frontiers | Research Topics

10 Years of frontiers in genetics: Past discoveries, current challenges and future perspectives

Edited by
William C. Cho, Jordi Pérez-Tur, Rosalba Giugno, Mehdi Pirooznia, Kathleen Boris-Lawrie, Dov Greenbaum, Blanka Rogina, Mojgan Rastegar, Rui Henrique, Peng Xu and Joao Batista Teixeira da Rocha

Published in
Frontiers in Genetics
Frontiers in Toxicology
Frontiers in Pediatrics
Frontiers in Pharmacology
Frontiers in Neuroscience
Frontiers in Cell and Developmental Biology
Frontiers in Ecology and Evolution

FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-2477-0
DOI 10.3389/978-2-8325-2477-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

10 Years of frontiers in genetics: Past discoveries, current challenges and future perspectives

Topic editors

William C. Cho — QEH, Hong Kong, SAR China
Jordi Pérez-Tur — Institute of Biomedicine of Valencia, Spanish National Research Council (CSIC), Spain
Rosalba Giugno — University of Verona, Italy
Mehdi Pirooznia — Johnson & Johnson, United States
Kathleen Boris-Lawrie — University of Minnesota Twin Cities, United States
Dov Greenbaum — Yale University, United States
Blanka Rogina — University of Connecticut Health Center, United States
Mojgan Rastegar — University of Manitoba, Canada
Rui Henrique — Portuguese Oncology Institute, Portugal
Peng Xu — Xiamen University, China
Joao Batista Teixeira da Rocha — Federal University of Santa Maria, Brazil

Citation

Cho, W. C., Pérez-Tur, J., Giugno, R., Pirooznia, M., Boris-Lawrie, K., Greenbaum, D., Rogina, B., Rastegar, M., Henrique, R., Xu, P., da Rocha, J. B. T., eds. (2023). *10 Years of frontiers in genetics: Past discoveries, current challenges and future perspectives*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2477-0

Author MP was employed by the company Johnson & Johnson.
The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Table of contents

06 Editorial: 10 years of Frontiers in genetics: past discoveries, current challenges and future perspectives
William C. Cho, Jordi Pérez-Tur, Rosalba Giugno, Mehdi Pirooznia, Kathleen Boris-Lawrie, Dov Greenbaum, Mojgan Rastegar, Rui Henrique, Peng Xu, Joao Batista Teixeira da Rocha and Blanka Rogina

09 Analysis of Longevity Traits in Holstein Cattle: A Review
Honghong Hu, Tong Mu, Yanfen Ma, XingPing Wang and Yun Ma

24 Evidence That the Etiology of Congenital Hypopituitarism Has a Major Genetic Component but Is Infrequently Monogenic
Youn Hee Jee, Mariam Gangat, Olga Yeliosof, Adrian G. Temnycky, Selena Vanapruks, Philip Whalen, Evgenia Gourgari, Cortney Bleach, Christine H. Yu, Ian Marshall, Jack A. Yanovski, Kathleen Link, Svetlana Ten, Jeffrey Baron and Sally Radovick

36 Molecular Characterization of α- and β-Thalassaemia Among Children From 1 to 10 Years of Age in Guangxi, A Multi-Ethnic Region in Southern China
Sheng He, Dongming Li, Shang Yi, Xiuning Huang, Chaofan Zhou, Biyan Chen, Yangjin Zuo, Li Lin, Faqin Chen and Hongwei Wei

42 From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries
Binglan Li and Marylyn D. Ritchie

57 Genetically Predicted Longer Telomere Length May Reduce Risk of Hip Osteoarthritis
Jing Yang, Huiqing Xu, Bingyue Cai, Jiahe Wei, Lingling Sun, Yasong Li, Tianle Wang and Yingjun Li

64 Molecular Cytogenetics in the Era of Chromosomics and Cytogenomic Approaches
Thomas Liehr

73 Genetic Analysis for Coronary Artery Disease Toward Diverse Populations
Kazuo Miyazawa and Kaoru Ito

79 Identification and Validation of Prognostic Factors of Lipid Metabolism in Obstructive Sleep Apnea
Lu Peng, Xiaodi Wang and Dan Bing

96 Benefits of Cohort Studies in a Consortia-Dominated Landscape
Selam Zenebe-Gete, Rebecca Salowe and Joan M. O'Brien

99 Identification and Functional Verification Reveals that miR-195 Inhibiting *THRSP* to Affect Fat Deposition in Xinyang Buffalo
Shuzhe Wang, Cuili Pan, Xiaojie Ma, Chaoyun Yang, Lin Tang, Jieping Huang, Xuefeng Wei, Hui Li and Yun Ma

112 Episodes of Rapid Recovery of the Functional Activity of the *ras85D* Gene in the Evolutionary History of Phylogenetically Distant *Drosophila* Species
A. I. Chekunova, S. Yu. Sorokina, E. A. Sivoplyas, G. N. Bakhtoyarov, P. A. Proshakov, A. V. Fokin, A. I. Melnikov and A. M. Kulikov

134 Non-Invasive microRNA Profiling in Saliva can Serve as a Biomarker of Alcohol Exposure and Its Effects in Humans
Edward A. Mead, Nadia Boulghassoul-Pietrzykowska, Yongping Wang, Onaiza Anees, Noah S. Kinstlinger, Maximillian Lee, Shireen Hamza, Yaping Feng and Andrzej Z. Pietrzykowski

156 An Update Evolving View of Copy Number Variations in Autoimmune Diseases
Rong-hua Song, Chao-qun Gao, Jing Zhao and Jin-an Zhang

172 Association Between TNF-α-308, +489, –238 Polymorphism, and COPD Susceptibility: An Updated Meta-Analysis and Trial Sequential Analysis
Zhiyu Xia, Yufei Wang, Fu Liu, Hongxin Shu and Peng Huang

182 Case Report: Mutation in AIMP2/P38, the Scaffold for the Multi-Trna Synthetase Complex, and Association With Progressive Neurodevelopmental Disorders
Mahta Mazaheri, Mahdie Yavari, Hadi Zare Marzouni, Angela Stufano, Piero Lovreglio, Simona D'Amore and Hamid Reza Jahantigh

188 Emerging Roles for LGR4 in Organ Development, Energy Metabolism and Carcinogenesis
Linlin Yang, Jing Wang, Xiaodi Gong, Qiong Fan, Xiaoming Yang, Yunxia Cui, Xiaoyan Gao, Lijuan Li, Xiao Sun, Yuhong Li and Yudong Wang

203 Estimates of Variance Components and Heritability Using Random Regression Models for Semen Traits in Boars
Yifeng Hong, Limin Yan, Xiaoyan He, Dan Wu, Jian Ye, Gengyuan Cai, Dewu Liu, Zhenfang Wu and Cheng Tan

214 Strategies for High-Efficiency Mutation Using the CRISPR/Cas System
Shuying Feng, Zilong Wang, Aifang Li, Xin Xie, Junjie Liu, Shuxuan Li, Yalan Li, Baiyan Wang, Lina Hu, Lianhe Yang and Tao Guo

234 A Final Frontier in Environment-Genome Interactions? Integrated, Multi-Omic Approaches to Predictions of Non-Communicable Disease Risk
Alexandra J. Noble, Rachel V. Purcell, Alex T. Adams, Ying K. Lam, Paulina M. Ring, Jessica R. Anderson and Amy J. Osborne

248 Immune ULBP1 is Elevated in Colon Adenocarcinoma and Predicts Prognosis
Guo-Tian Ruan, Hai-Lun Xie, Li-Chen Zhu, Yi-Zhong Ge, Lin Yan, Cun Liao, Yi-Zhen Gong and Han-Ping Shi

262 Regulation of DNA Methylation During Plant Endosperm Development
Dongdong Lu, Jixian Zhai and Mengli Xi

269 Integrated Analysis of Competitive Endogenous RNA Networks in Acute Ischemic Stroke
Zongkai Wu, Wanyi Wei, Hongzhen Fan, Yongsheng Gu, Litao Li and Hebo Wang

291 Exploration of Reduced Mitochondrial Content–Associated Gene Signature and Immunocyte Infiltration in Colon Adenocarcinoma by an Integrated Bioinformatic Analysis
Jinlin Kang, Na li, Fen Wang, Yan Wei, Yangyang Zeng, Qifan Luo, Xuehua Sun, Hui Xu, Jin Peng and Fuxiang Zhou

306 Genetic Regulation of Vertebrate Forebrain Development by Homeobox Genes
Ryan F. Leung, Ankita M. George, Enola M. Roussel, Maree C. Faux, Jeffrey T. Wigle and David D. Eisenstat

341 Identification of Enhancer RNA CDK6-AS1 as a Potential Novel Prognostic Biomarker in Gastric Cancer
Shifeng Yang, Xiaoming Zou, Hao Yang, Jiacheng Li, Ange Zhang, Lisha Zhang, Changjian Li, Lei Zhu and Zhen Ma

352 Comparative Toxicotranscriptomics of Single Cell RNA-Seq and Conventional RNA-Seq in TCDD-Exposed Testicular Tissue
Alex Haimbaugh, Danielle Meyer, Camille Akemann, Katherine Gurdziel and Tracie R. Baker

367 Breaking the Immune Complexity of the Tumor Microenvironment Using Single-Cell Technologies
Simone Caligola, Francesco De Sanctis, Stefania Canè and Stefano Ugel

386 Genetic Diversity, Habitat Relevance and Conservation Strategies of the Silver Carp in the Yangtze River by Simple Sequence Repeat
Yu-Ting Luo, Di-An Fang, Yan-Feng Zhou, Dong-Po Xu, Yun-Xin Peng, Min-Ying Zhang, Cheng-Cheng Mao, Xue-Mei Tang, Jun Xu and Yang You

397 CLP1 is a Prognosis-Related Biomarker and Correlates With Immune Infiltrates in Rheumatoid Arthritis
Zhenyu Zhao, Shaojie He, Sheng Tang, Xiaofeng Lai, Jie Ren, XinCheng Yu, Jinhua Lin, Mohan Wang, Mariya M. El Akkawi, Shan Zeng and Dingsheng Zha

415 Regulation of 3D Organization and Its Role in Cancer Biology
Anghui Peng, Wang Peng, Ruiqi Wang, Hao Zhao, Xinyang Yu and Yihao Sun

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 05 May 2023
DOI 10.3389/fgene.2023.1192071

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Fangqing Zhao,
Beijing Institutes of Life Science (CAS), China

*CORRESPONDENCE
William C. Cho,
chocs@ha.org.hk,
williamcscho@gmail.com
Blanka Rogina,
Rogina@uchc.edu

RECEIVED 22 March 2023
ACCEPTED 17 April 2023
PUBLISHED 05 May 2023

CITATION
Cho WC, Pérez-Tur J, Giugno R, Pirooznia M, Boris-Lawrie K, Greenbaum D, Rastegar M, Henrique R, Xu P, Rocha JBTd and Rogina B (2023), Editorial: 10 years of Frontiers in genetics: past discoveries, current challenges and future perspectives.
Front. Genet. 14:1192071.
doi: 10.3389/fgene.2023.1192071

COPYRIGHT
© 2023 Cho, Pérez-Tur, Giugno, Pirooznia, Boris-Lawrie, Greenbaum, Rastegar, Henrique, Xu, Rocha and Rogina. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: 10 years of Frontiers in genetics: past discoveries, current challenges and future perspectives

William C. Cho[1]*, Jordi Pérez-Tur[2], Rosalba Giugno[3], Mehdi Pirooznia[4], Kathleen Boris-Lawrie[5], Dov Greenbaum[6,7], Mojgan Rastegar[8], Rui Henrique[9,10], Peng Xu[11], Joao Batista Teixeira da Rocha[12] and Blanka Rogina[13]*

[1]Queen Elizabeth Hospital, Kowloon, Hong Kong SAR, China, [2]Institut de Biomedicina de València-CSIC, CIBER-CIBERNED, ISCIII, Valencia, Spain, [3]Department of Computer Science, University of Verona, Verona, Italy, [4]Johnson & Johnson, Washington, DC, United States, [5]Department of Veterinary and Biomedical Sciences, University of Minnesota Twin Cities, St. Paul, MN, United States, [6]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, United States, [7]Zvi Meitar Institute for Legal Implications of Emerging Technologies, Reichman University, Herzliya, Israel, [8]Department of Biochemistry and Medical Genetics, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada, [9]Portuguese Oncology Institute of Porto (IPO Porto)/Porto Comprehensive Cancer Center (Por-to.CCC), R. Dr. António Bernardino de Almeida, Porto, Portugal, [10]School of Medicine and Biomedical Sciences (ICBAS), University of Porto (ICBAS-UP), Porto, Portugal, [11]Xiamen University, Xiamen, China, [12]Departamento de Bioquímica e Biologia Molecular (UFSM), Federal University de Santa Maria, Santa Maria, Brazil, [13]Department of Genetics and Genome Sciences, School of Medicine, University of Connecticut Health Center, Farmington, CT, United States

KEYWORDS

genetics, genomics, Frontiers in genetics, human physiology, genomics data analysis

Editorial on the Research Topic
10 years of Frontiers in genetics: past discoveries, current challenges and future perspectives

This Editorial is for the Research Topic dedicated to 10 years of Frontiers in Genetics. Frontiers in Genetics was launched in 2011 with the grand challenge proposed by Professor David Allison, founding Field Editor, to use new and exciting technologies—such as the mathematical and biological tools of modern genomics—to answer some of the key questions about evolutionary biological science. The second challenge was designed to address genomic structure, tissue-specific gene expression during development, aging, and disease. The final challenge was aimed at understanding how modification of gene expression could affect health and longevity using modern genomic tools such as knockout mice, transgenic plants or gene therapies.

The answers to these three challenges have resulted in the publication of leading edge genetics and genomic studies, ranging from basic to translational research, perspectives, and commentaries. Ten years later, Frontiers in Genetics has arguably accomplished its initial stated goals, and has even expanded remarkably, maturing into a platform to publish new discoveries, views, perspectives, advances and challenges in the field, ultimately becoming the largest open access journal in genetics.

In 2011, Frontiers in Genetics had twelve mother-field Specialty Sections. Today it has 23 Specialty sections, each directed by dedicated Chief Editors, as well as an outstanding

editorial board of over 7,500 dedicated and exceptional Associated Editors and Reviewers. During our first 10 years, Frontiers in Genetics has collectively published over 6,725 articles, all of which are freely available online to scientists, doctors, patients, policymakers and the general public. Some exceptional statistics: The current impact factor is 4.772, which keeps on rising, and the most cited manuscript, with over 502 citations, is by Sawana et al. (2014), which describes the molecular signature and phylogenomic analysis of the genus Burkholderia. There were overall of 1,12,857 articles citations between 2011 and 2021 and 182 articles were cited more than 100 times. There was an impressive number of total downloads of 1,15,36,824, with a perspective on the origins of the Ashkenaz, Ashkenazic Jews and Yiddish leading this impressive record with 4,34,504 views (Das et al., 2017). Lastly, there were 676 Research Topics posted between 2011 and 2021.

This Research Topic is a celebration of the 10 year Anniversary of this remarkable journal, gathering together editors and key contributors to highlight significant contributions of the past, present and future of genetics. This Topic, which comprises thirty rigorously reviewed manuscripts covering a wide spectrum of studies in humans, animals and plants, aims to give an overview of the most important areas and advances in genetics over the last 10 years, and to provide a platform to raise current challenges for exciting, new research.

The primary area of research within the Research Topic question how genes and genomes are related to phenotypes and human physiology. It includes research manuscripts and reviews that explore new etiology in human diseases, including, congenital hypopituitarism (https://doi.org/10.3389/fgene.2021.697549), α and β-Thalassemia in young children from regions in Southern China (https://doi.org/10.3389/fped.2021.724196), and the association between TNF- α-Polymorphisms and COPD susceptibility (https://doi.org/10.3389/fgene.2021.772032). Several manuscripts describe the value of different predictors in human disease, such as the value of longer telomere lengths in reducing risk of hip osteoarthritis (https://doi.org/10.3389/fgene.2021.718890), prognostic factors of lipid metabolism in obstructive sleep apnea (https://doi.org/10.3389/fgene.2021.747576), genetic analysis of coronary artery disease (https://doi.org/10.3389/fgene.2021.766485), use of integrated analysis of the RNA network in acute ischemic stroke (https://doi.org/10.3389/fgene.2022.833545), as well as the use of CLP1 as a prognostic biomarker of immune infiltrates in rheumatoid arthritis (https://doi.org/10.3389/fphar.2022.827215). The topic also includes a review on the role of copy number variation in autoimmune diseases (https://doi.org/10.3389/fgene.2021.794348).

Several original research articles describe the novel use of specific genes in diagnosis and prognosis of cancer, such as, ULBP1 in the case of adenocarcinoma (https://doi.org/10.3389/fgene.2022.762514), or use of enhancers as biomarkers of gastric cancer (https://doi.org/10.3389/fgene.2022.854211). A manuscript that describes the role of reduced mitochondrial content and immunocyte infiltration is also included in this topic (https://doi.org/10.3389/fgene.2022.832331). A review describes use of single-cell technologies and computational techniques for immune-profiling of the tumor environment (https://doi.org/10.3389/fgene.2022.867880). A case report describes identification of a novel mutation in AIMP2/P38, and its role in leukodystrophy, a progressive neurodevelopmental disorder (https://doi.org/10.3389/fgene.2022.816987). A second review article describes the role of homeobox genes in the genetic regulation of vertebrate forebrain development (https://doi.org/10.3389/fnins.2022.843794).

As another goal of Frontiers in Genetics is to support development of cutting edge technological and analytical tools to study genomic data, this topic also includes a review article that describes the role of genome-wide association studies (GWAS) in identification of thousands of single nucleotide polymorphism that have been associated with different human diseases and traits (https://doi.org/10.3389/fgene.2021.713230). In addition, an opinion article discusses the advantages and power of forming GWAS consortia to avoid some challenges associated with these studies and the cost of recruiting a large cohort (https://doi.org/10.3389/fgene.2021.801653). And another review discusses the role of molecular cytogenetics during highly used novel techniques such as chromosomics and cytogenomics analysis (https://doi.org/10.3389/fgene.2021.720507).

One of the new frontiers in the post-genomic era is three-dimensional (3D) genomics, which explores the relationship between chromatin spatial conformation and its effects on gene transcription. A review article of this topic describes the contribution of 3D organization in cancer biology (https://doi.org/10.3389/fcell.2022.879465). An additional review article discusses the role of the leucine-rich repeats containing G-protein coupled receptor 4 (LGR4) in energy metabolism, development and carcinogenesis https://doi.org/10.3389/fgene.2021.720507).

Another area that is covered in Frontiers in Genetics is gene flow among species and populations, including genomic research of farm animals. Here, we have included several manuscripts describing new findings on the role of genomics in animals, such as a review of longevity traits in Holstein Cattle (https://doi.org/10.3389/fgene.2021.695543), the molecular mechanisms of fat deposition in Xianyang Buffalo (https://doi.org/10.3389/fgene.2021.736441), and the study of variance components and heritability of semen traits over the reproductive life in boars (https://doi.org/10.3389/fgene.2022.805651). Adaptation to habitat and genetic diversity of the Silver Carp was also described (https://doi.org/10.3389/fevo.2022.850183), as well as a study of the role of evolutionary history in the formation of the regulatory regions of specific genes in different *Drosophila* species (https://doi.org/10.3389/fgene.2021.8072340.).

Interactions between organisms and their environments, as well as molecular and cellular genetics, are also covered in Frontiers in Genetics and in this Topic. An article describes an analysis of the regulation of DNA methylation during plant endosperm development (https://doi.org/10.3389/fgene.2022.760690). An additional manuscript describes the use of single-cell RNA-Seq and RNA-Seq to study the effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) on zebrafish testes during sexual differentiation and its effects on sterility https://doi.org/10.3389/ftox.2022.821116.

There are number of cutting-edge technologies that will play an important part of future frontiers in genetics such as single-cell multiomics, new gene editing tool, e.g., Retron Library Recombineering (RLR) and others. This topic has several articles that focuses of these new frontiers, such as a review that discusses strategies to design high-efficiency mutations using the CRISPR/Cas system (https://doi.org/10.3389/fcell.2021.803252). Another review

article describes the use of integrated and multi-omic approaches to study environmental and genomic interactions, and the impact of such interactions on human and animal health (https://doi.org/10.3389/fgene.2022.831866).

An interesting original study describes the use of micro RNA profiling in saliva to produce biomarkers of alcohol exposure in humans (https://doi.org/10.3389/fgene.2021.804222).

Frontiers would also like to take this opportunity to thank and congratulate the founding Field Chief Editor of Frontiers in Genetics, Professor, Dr. David B. Allison from Indiana University Bloomington, USA, who was leading the Field from 2010 until 2017. Dr. Allison's vision and recognition of the value of Open Science directly resulted in establishing the journal's worldwide reputation (Allison, 2011). This strong vision continued with Professor Emmanouil Dermitzakis, Ph.D., from University of Geneva Switzerland, the second Field Editor of Frontiers in Genetics, who served in this visionary role from 2017 until 2021. During his leadership, Frontiers in Genetics continued to expand the number of Specialties. In 2021, Professor Enrico Domenici, Ph.D., from University of Trento, Italy, became the third Field Chief Editor. Professor Domenici continues to successfully provide leadership and guidance in the further expansion of the field, https://www.frontiersin.org/journals/genetics/about.

Standing on the shoulders of giants, Frontiers in Genetics will continue to follow its mission and original vision to bring together the world's leading experts in genetics to freely disseminate cutting-edge studies and accelerate the progress of research in all areas of genes and genomes in humans, plants, livestock and other model organisms. This also includes the increasingly pertinent ethical, legal and social implications of genomic studies that are also included in a section within Frontiers in Genetics.

The record of the first 10 years, and the continuation of truly wonderful experience, will move us further into the next decade and many more to come.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This research was supported by the National Institute on Health, grant number R01AG059586, R01AG059586-03S1, and by the University of Connecticut (UConn) Claude D. Pepper Older Americans Independence Center (P30-AG067988). BR is a recipient of a Glenn Award for Research in Biological Mechanisms of Aging. Research in the MR lab is supported by the Canadian Institutes of Health Research (CIHR) Project Grant 202203PJG-482520-CIA-CDAA-111824, CIHR Project Grant 202209PJT-486512-CIA-CDAA-111824, University of Manitoba Collaborative Research Program, Ontario Rett Syndrome Association Hope Fund, Children Hospital Research Institute of Manitoba (CHRIM), and Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2016-06035. Work at JP-T's lab was supported by CIBER-Consorcio Centro de Investigación Biomédica en Red-(209), Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación.

Acknowledgments

The authors are grateful to Trinity King and Frontiers in Genetics Editorial Office for their kind help with Frontiers in Genetics statistics and this Research Topic.

Conflict of interest

Author MP was employed by the company Johnson & Johnson.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Allison, D. B. (2011). The grand challenge for frontiers in genetics: To understand past, present, and future. *Front. Genet.* 2, 2. doi:10.3389/fgene.2011.00002

Das, R., Wexler, P., Pirooznia, M., and Elhaik, E. (2017). The origins of Ashkenaz, ashkenazic Jews, and Yiddish. *Front. Genet.* 8, 87, doi:10.3389/fgene.2017.00087

Sawana, A., Adeolu, M., and Gupta, R. S. (2014). Molecular signatures and phylogenomic analysis of the genus Burkholderia: Proposal for division of this genus into the emended genus Burkholderia containing pathogenic organisms and a new genus paraburkholderia gen. Nov. Harboring environmental species. *Front. Genet.* 5, 429. doi:10.3389/fgene.2014.00429

frontiers in Genetics

REVIEW
published: 03 August 2021
doi: 10.3389/fgene.2021.695543

Analysis of Longevity Traits in Holstein Cattle: A Review

*Honghong Hu†, Tong Mu†, Yanfen Ma, XingPing Wang and Yun Ma**

Ningxia Key Laboratory of Ruminant Molecular and Cellular Breeding, School of Agriculture, Ningxia University, Yinchuan, China

Dairy cow longevity is an essential economic trait that can supplement the breeding value of production traits, which is related to the herd time and lifetime milk yield of dairy cows. However, longevity is a relatively difficult trait to select for dairy cow breeding due to low heritability and numerous influence factors of the longevity in dairy cows. Longevity trait has been used as an important breeding target of a comprehensive selection index in many dairy developed countries; however, it has not been included in performance index in many developing countries. At present, cows in these countries are still in the primary stage of "large quantity, low quality, high cost, and low yield." The average parity of dairy cows is less than 2.7, which is difficult to maintain the production efficiency to meet the demands of the dairy industry. Therefore, there is an urgent need to select and breed for the longevity of dairy cows. The various definitions and models (including linear, threshold, random regression, sire, and survival analysis) of longevity were reviewed and standardized. Survival analysis is the optimal model to evaluate longevity, and the longevity heritability is 0.01–0.30 by using different definitions and models. Additionally, the relationship between longevity and other traits was summarized, and found that longevity was regulated by multiple factors, and there were low or medium genetic correlations between them. Conformation traits, milk production traits, reproductive traits, and health traits may be used as indicators to select and breed the longevity of dairy cows. The genetic assessment methods, heritability, influencing factors, importance, breeding, and genetics of longevity were reviewed in the manuscript, which could provide a valuable reference for the selective breeding to extend the productive life of Holstein cattle.

Keywords: Holstein cattle, longevity, heritability, breeding method, genetic correlation

OPEN ACCESS

Edited by:
Peng Xu,
Xiamen University, China

Reviewed by:
Guillermo Giovambattista,
CONICET Institute of Veterinary Genetics (IGEVET), Argentina
Juan Pablo Gutiérrez,
Complutense University of Madrid, Spain

***Correspondence:**
Yun Ma
mayun_666@126.com

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *15 April 2021*
Accepted: *25 June 2021*
Published: *03 August 2021*

Citation:
Hu H, Mu T, Ma Y, Wang X and Ma Y (2021) Analysis of Longevity Traits in Holstein Cattle: A Review. Front. Genet. 12:695543. doi: 10.3389/fgene.2021.695543

INTRODUCTION

The longevity of dairy cows refers to the time from the first calving to exit the herd when cows do not have sufficient productivity. The production life of dairy cows is less than 3–4.5 years (Kerslake et al., 2018), but in fact the maximum annual production occur in the fifth lactation period, and the highest annual profit typically achieved in the sixth lactation period (Horn et al., 2012). The natural lifespan of cows is approximately 20 years, but the average culled time is much earlier than the natural life. Moreover, cows will be eliminated if they cannot reach the peak production to obtain the highest profit (Najafabadi et al., 2016).

The longevity of dairy cows is a complex trait with low heritability and a lack of supporting data, and longevity is affect by many factors, such as the inherent factors (lactation, health, conformation traits, and reproductive performance; Ferris et al., 2014) and the external factors (milk price, nutrition, management, policy, feeding cost, and replacement heifers; Grandl et al., 2016; Vries and Marcondes, 2020). Therefore, it is a difficult task for breeding longevity traits in dairy cows, and it is necessary to select the traits of longevity, which determines the utilization value of dairy cows, improves the economic benefits of dairy farms. It is especially important for the development of dairy industry.

Longevity of dairy cows was studied at home and abroad, and various definitions and methods of longevity were proposed (**Table 1**). However, because these terms are often used interchangeably and confusedly, which can confound the study of longevity traits. Therefore, it is necessary to standardize the terms of longevity traits. Herd life refers to the days from birth to culling or death (Zhang et al., 2021b), productive life refers to the days from the first calving to culling or death (Raguz et al., 2011), milking life refers to the days from the first calving to culling or death but excludes all dry periods (Zhang et al., 2021b), and stayability refers to the probability that a cow remains in the herd enough time to raise a certain number of calves that pay for her development and maintenance costs (Costa et al., 2020). Nevertheless, longevity is included as an important indicator in the comprehensive selection index of dairy cows in various countries. Prolonging the productive life of dairy cows can reduce involuntary culling (Zavadilová and Stipkova, 2012) and improve the voluntary culling of dairy cows (Sewalem et al., 2008), but can also help to increase the market profit of dairy enterprises (Albert, 2020), meet consumer demand, enhance animal welfare, respond to climate change, and promote environmentally sustainable development (Grandl et al., 2019). We mainly focuses on genetic evaluation, influencing factors, breeding, genetics, and breeding methods of longevity were reviewed to provide a reference for the selection and breeding of longevity traits in Holstein cattle.

GENETIC EVALUATION AND HERITABILITY

Longevity is a very important economic trait. Several countries have explored different models to evaluate longevity based on different definitions, data properties, and quality to improve the selection of longevity traits by genetic evaluation. Typically, these models include linear models (Allaire, 1976), threshold models (Boettcher et al., 1999), random regression models (Pelt et al., 2015), sire models (Brotherstone et al., 1997),

TABLE 1 | Terms and definitions of longevity in dairy cows in the literature.

Terms	Definitions	Source
Productive life	The day from first calving to culling. The actual productive life and mainly depends on productivity. The total number of days in milk up to 84 mo of age with a restriction of 305, 500, or 999 d per lactation (PL305, PL500, or PL999, respectively).	Raguz et al., 2011 Vukasinovic et al., 1997 Tsuruta et al., 2005
Functional productive life	The ability of the cow to avoid culling for involuntary reasons such as sterility or disease.	Vukasinovic et al., 1997
Length of productive life	The number of days from first calving until culling or censoring. The days between first calving and disposal. The number of completed lactations.	Caraviello et al., 2004 Martinez et al., 2004 Yazdi et al., 1999
Herd life	The day from first calving to culling. The total number of days from the first calving date to the last (culling) date. The length of time that individual cows remain in herds after their first calving. Survival to the next lactation.	Brickell and Wathes, 2011 Tsuruta et al., 2005 Hare et al., 2006 Jairath et al., 1998
Longevity	The total months in milk by 84 mo of age. The number of days from first calving until culling or censoring The length of time during which an animal is able to stay producing in the herd, and survival. The individual farmer to eliminate cows for low milk production (voluntary culling)	Vanraden and Klaaskate, 1993 Sasaki et al., 2012 Jamrozik et al., 2013 Strapáková et al., 2019
True longevity	The ability to delay any culling.	Jenko et al., 2013
Stayability	A measure of cow survival that does not require recording of cull data. The probability of a cow remain in the herd enough time to raise a certain number of calves that pay for her development and maintenance costs. The ability of the cow to calve at least three times until 76 months of age. The measure of whether or not an animal remains and produces in the herd until a specified point in time.	Handcock et al., 2020 Costa et al., 2020 Ramos et al., 2020 Jamrozik et al., 2013
Functional longevity	The number of days between the first calving and culling. The different lengths for the time interval for survival. The cow's ability to avoid involuntary culling or culling not correlated with its own production. The capability of cows to delay involuntary culling for infertility or diseases.	Zavadilova et al., 2011 Pelt et al., 2015 Stanojević et al., 2018 Strapáková et al., 2019
Lifespan	The number of lactations an animal completes or is expected to complete prior to culling.	Brotherstone et al., 1997
Milking life	The days from the first calving to culling or death but excludes all dry periods.	Zhang et al., 2021b

and survival analysis (Buenger et al., 2001). Linear models, threshold models, and random regression models can process multiple traits simultaneously; thereby directly estimate the genetic correlation between longevity and other traits with a relatively fast calculation speed (Imbayarwo-Chikosi et al., 2015). Survival analysis can appropriately accommodate censored data, consider time-dependent environmental impact, and manage the skewed distribution of longevity characteristics (Imbayarwo-Chikosi et al., 2016). The estimated value of the trait is remarkably close to the measured value, which can easily be adapted to longevity data and provide accurate results, but the calculation speed is relatively slow. In addition, linear, threshold models, and random regression models generally produce lower estimation of longevity heritability than survival analysis models on the original scale (Ducrocq, 1997; Setati et al., 2004; Jamrozik et al., 2008; Kern et al., 2014).

Survival analyses included parametric, semi-parametric, and non-parametric methods (Smith and Westgarth, 1957). Cox proportional hazard model is a semi-parametric method, and the Weibull distribution model is a parametric method, which have been used to estimate the longevity traits of dairy cows (Zhao, 2013). The Cox proportional hazard model is used to analyze the factors that affect the survival time without a clear benchmark risk rate function, which has a wide range of applications and shows high statistical efficiency (Stokes, 2019). The Weibull regression model is a multi-factor analysis model and is based on the Weibull distribution. The weight of each factor in the production life can be obtained with the change of time. The Weibull model can more adapt to the censoring, covariates changes with time, and the screening process more intuitively. Therefore, Weibull regression is more accurate than the Cox proportional hazard model, but it is also more complex.

Researchers use different models to genetically assess different longevity definitions, and find that the longevity heritability is low with rates 0.01–0.30 (**Table 2**). Although the selection process might be slow, there is sufficient genetic variation. The breeding of longevity traits could be improved indirectly by selecting traits that should be a strong genetic correlation with longevity. Indirect selection is useful if the square of the genetic correlation between the indirect trait and longevity is greater than the heritability of the longevity trait. In addition, the breeding of longevity traits could be improved by combining conventional breeding strategies with modern genome selection technology.

LONGEVITY WITH OTHER TRAITS

Longevity and Conformation

Conformation traits with genetic traits can be monitored in early life (usually the first lactation), thus it is an attractive indirect traits of longevity (Miglior et al., 2017). Conformation traits include lactation system traits (breast texture, udder attachment, udder depth, teat placement, median suspensory, rear breast width, and rear breast height), foot and leg traits (bone quality, rear view of the rear leg, and foot angle), and body traits (weight, body height, and body depth). The higher total conformation score, the longer the longevity. Dairy workers worldwide have been committed to improve the longevity of dairy cows by looking for the ideal score for each conformation trait and exploring the correlation between these traits and longevity (**Figure 1**; Zavadilova et al., 2009, 2011; Kern et al., 2015; Mao, 2015; Imbayarwo-Chikosi et al., 2016). The nine-point scale is often used to score the conformation traits of dairy cows, and the optimal score value is selected only when the productive life is longest and the milk yield is highest.

Longevity has a weak genetic correlation with the lactation system and body condition (Zavadilová and Stipkova, 2012), and it is unfavorable for selecting the longevity trait. While body height, chest width, loin strength, rump angle, rump width, foot angle, bone quality, rear view of the rear leg, breast texture, fore udder attachment, rear udder attachment, udder depth, teat placement, median suspensory, and top-line were all positive genetic correlated with the longevity (Sewalem et al., 2004; Wu, 2007), and it is favorable for longevity trait selection, among these traits, foot angle and rump width are the highest correlations with the productive life of cows. On the contrary, there is a significantly negative genetic correlation among teat length, fore teat placement, and longevity (Vacek et al., 2006; Čanji et al., 2008). Moreover, several studies have found that other traits are negatively genetic correlated with longevity (Vollema and Groen, 1997; Perez-Cabal et al., 2006; Muntean et al., 2018). Above all, longevity has a high genetic correlation with lactation system traits, a low genetic correlation between foot and leg, and a low or medium negative genetic correction with body traits (**Figure 2**). These differences among studies may be due to the different breeding needs, herd size, genetic background, the term of longevity, and the analytical models used in the study of cow longevity in different countries. Additionally, the oral score also plays an important role in productive life. Muntean et al. (2018) found that Bolan cows with a better oral score (score of 1) have longer life expectancy. If degeneration of the incisors affects the feeding ability of the cows, which in turn affects the nutritional level needed to maintain physical condition. Cows can live longer and produce more milk when they achieve ideal scores conformation. Therefore, these conformation traits are the ideal indirect trait for selecting and breeding longevity of dairy cows.

Longevity and Milk Production Traits

To reduce the cost of reserve cows during the breeding process, it is important to select cows with excellent milk production and longevity (Wasana et al., 2015). A study of 36,663 Slovenian brown cattles found that longevity was positively genetic correlated with 305-day milk yield, with a correlation coefficient of 0.23 (Jenko et al., 2015). Dairy cows had a longer lifespan and higher final milk production when the first milk production of primiparous cows was less than 30 L/d (Haworth et al., 2008). Kaupe et al. (2007) studied 1,291 Holstein cows and found that longevity was significantly negatively phenotypic correlated with milk fat and somatic cell count, with a correlation

TABLE 2 | Heritability of longevity.

Sample	Country	Trait	Model	Heritability	References
Holstein	America	Herd life	Linear mixed model	0.25	Allaire, 1976
Holstein	United Kingdom	Length of productive life	Sire	0.10	Hoque and Hodges, 1980
Holstein	America	Herd life	Sire	0.035	Dentine et al., 1987
Holstein	America	Herd life/Functional Herd life	Sire	0.03	Boldman et al., 1992
Holsteins	America	Herd life	Multiple – trait mixed model	0.01–0.10	Short and Lawlor, 1992
Holsteins	America	longevity	Linear model	0.085	Vanraden and Klaaskate, 1993
Swiss Brown	Switzerland	Herd life	Sire	0.03–0.14	Vukašinović et al., 1995
Black and White	Netherlands	Length of productive life/ Herd life	Sire	0.035–0.136	Vollema and Groen, 1996
Holstein-Friesian dairy cattle	UK	Lifespan	Sire	0.063	Brotherstone et al., 1997
Cow	Germany	Functional longevity	Survival analysis	0.18	Buenger et al., 2001
Holsteins	America	Herd life/productive life	Multiple trait sire model	0.08–0.10	Tsuruta et al., 2005
South African Angus	South Africa	Stayability	Threshold	0.24–0.30	Maiwashe et al., 2009
Holstein	Tunisie	Longevity	Proportional hazard model	0.19	M'hamdi et al., 2010
Holstein	Japan	Longevity	Weibull proportional hazard model	0.119–0.123	Sasaki et al., 2012
Brown cattle	Slovenia	True longevity Functional longevity	Piecewise Weibull baseline Survival models	0.094 0.099	Jenko et al., 2013
Black and white	Netherlands	Longevity	Random regression sire-maternal grandsire	0.115–0.149	
Dairy cattle	Dutch	Functional longevity	Random regression	0.146	Pelt et al., 2015
Holstein dairy	Iran	Length of productive life	Weibull proportional risk model	0.018	Najafabadi et al., 2016
Danish dairy cattle	Danish	Longevity	NAV longevity model	0.022–0.090	Clasen et al., 2017
South African Holstein Cattle	South Africa	Functional longevity	Weibull proportional hazards model	0.11	Imbayarwo-Chikosi et al., 2017
Holstein	Iran	Length of productive life/ Functional productive life	Regression	0.032/0.012	Mirhabibi et al., 2018
Holstein Slovak Simmental	Slovak	Length of productive life	Weibull sire model	0.13 0.05	Strapáková et al., 2019

coefficient of −0.08 and −0.52, respectively, and longevity was positively phenotypic correlated with protein content, with a correlation coefficient of 0.01. Longevity had a significant positively genetic correlation with milk fat and milk yeild, the correlation coefficients are 0.46 and 0.43, respectively (Weigel et al., 1998), and it was positively genetic correlated with fat and protein, the correlation coefficients are 0.56–0.61 (Haile-Mariam and Pryce, 2015). However, other studies have found a negative correlation between production life and milk production (Mirhabibi et al., 2018; Bieber et al., 2019), which indicate that high-yield cows have high energy metabolism during lactation, and their breasts may be more vulnerable to the milking equipment. There is a positive genetic correlation between somatic cells and blood immune-related cells. Sartori et al. (2020) found that the high positive genetic correlation between longevity and fighting ability (average ra = 0.669). Therefore, a higher resistance to mastitis is associated with their better welfare and longer the lifespan of dairy cows, which lead to a higher milk yield and production to farm benefit.

Furthermore, the urea nitrogen content, lactose content, and milking temperament also impact longevity of dairy cows. Miglior et al. (2006) found that there was a linear correlation between the lifespan and the urea nitrogen content, cows culling rate decreased with the increasing of urea nitrogen content, but the cows culling rate would increased when the content of urea nitrogen still increased to a certain extent. Because urea nitrogen content above 19 mg/dL had an important influence on the reproductive traits of dairy cows (Wang et al., 2010), which would increase their return to estrus and shorten their production life. On the contrary, the content of urea nitrogen less than 12 mg/dL might reduce milk production and milk protein content by affecting the rumen microflora quantity and reducing dietary protein index, thereby shortening the life of dairy cows (Zhang et al., 2011). Costa et al. (2020) found that the lactose content in milk decreased and dairy cows were more likely to be eliminated with the increase of breast infection degree and parity. Miglior et al. (2006) also found that cows with low lactose level had higher risk of elimination, whereas dairy cows with high lactose levels had a low risk of elimination. Because high lactose level may affect the pregnancy rate, early luteal reaction, and subclinical and clinical ketosis of dairy cows, which can shorten lifespan. Milking temperament can be defined according to milking behavior (Hedlund and Løvlie, 2015) and aggressiveness occurs during feeding (Uetake et al., 2004). Cows with a fast milking speed have a higher somatic cell count and susceptible mastitis (Sewalem et al., 2010). A docile temperament is beneficial for increasing the calving rate, conception rate, milk production, and production life (Chang et al., 2019).

Longevity and Reproductive Performance

Improvement in reproductive performance and the using of sex-controlled semen increased the number of replacement heifers in most dairy economies, which also cull more reserve cows (Overton and Dhuyvetter, 2020). Sewalem et al. (2005)

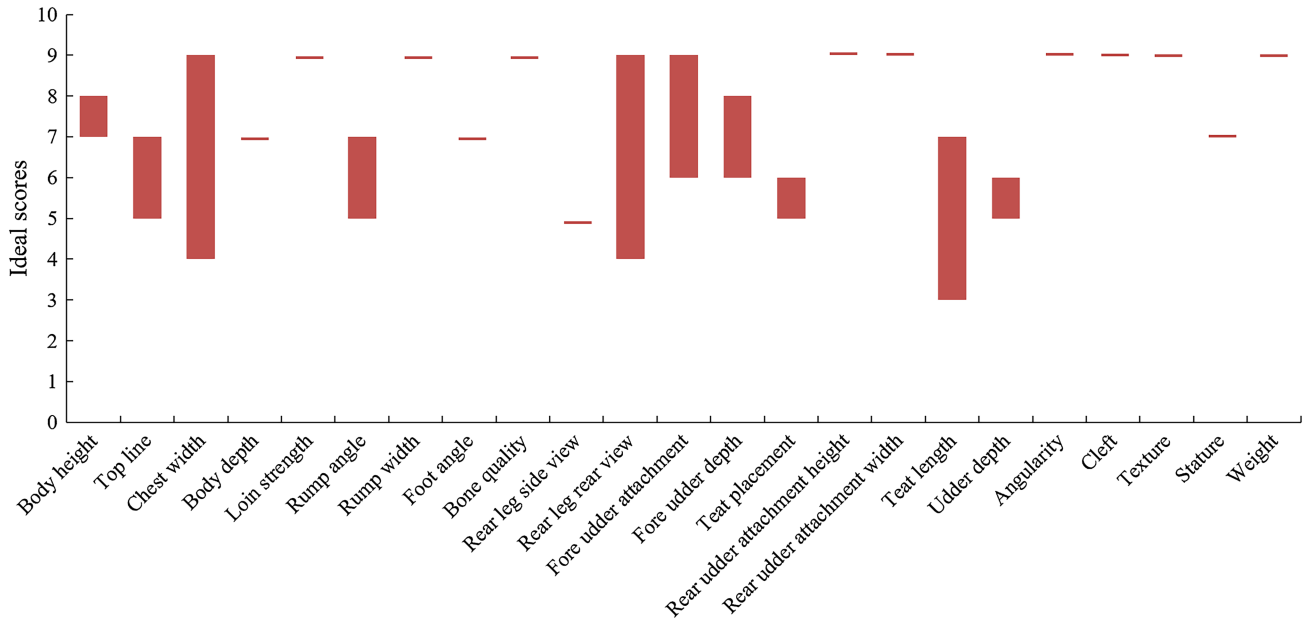


FIGURE 1 | Ideal score for various body traits of dairy cows.

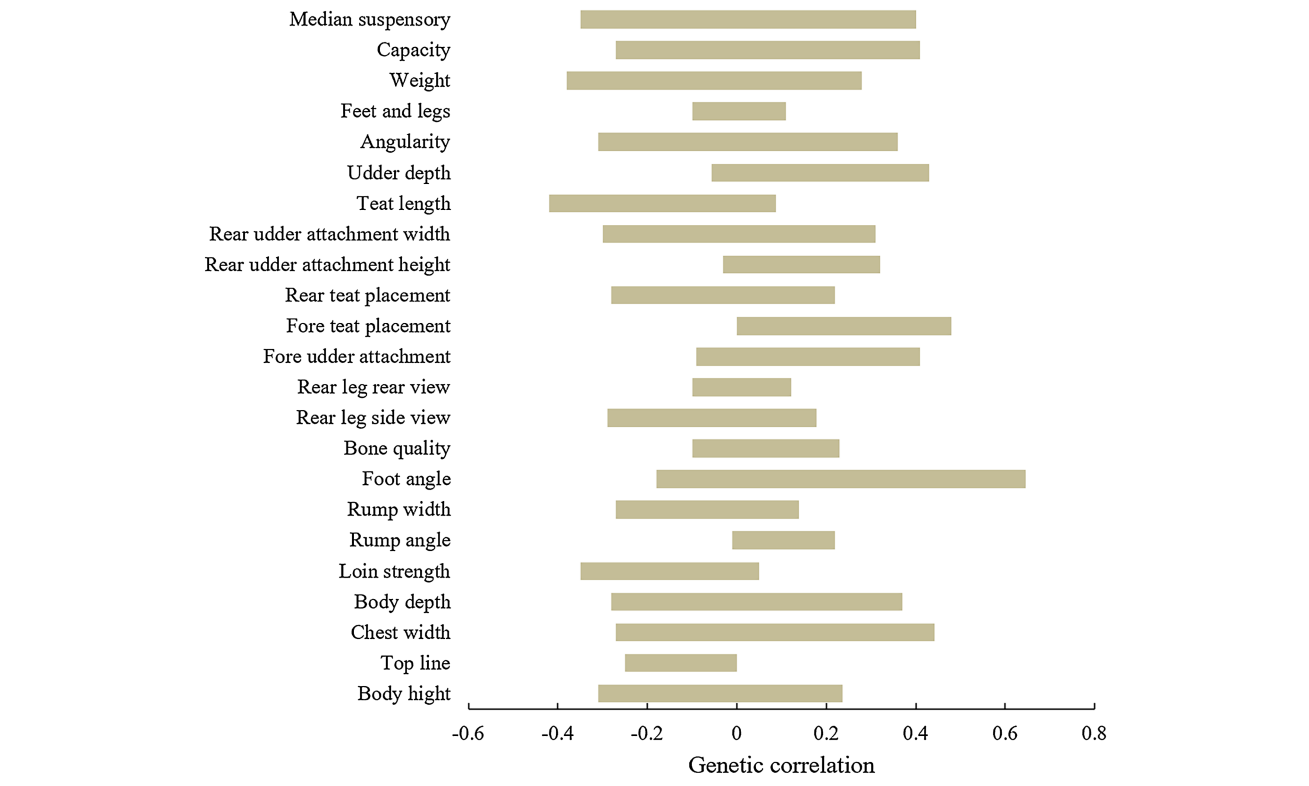


FIGURE 2 | Genetic correlations between longevity and type traits.

analyzed the factors affecting the herd life of Holstein, Jersey, and Ayrshire cows, and found that the risk of being culled was higher for older heifers than for heifers calving at an age between 24 and 28 months in all breeds. Late calvings are presumably caused by some problems associated with herd management, fertility, other health problems, and higher

rearing costs. In addition, cows first calving at <21 month of age have a high risk of culling due to dystocia and the high elimination is related to the quota system that exists in Canada (Nilforooshan and Edriss, 2004). Anim et al. (2020) analyzed different lactations recordings and found that productive life of dairy cows gradually decrease with the increasing of the age of the first calving, especially for cows calving at an age over 29 months. Mossa and Ireland (2019) found that cows had low fertility, low milk yield, and a high elimination risk with low and high follicle numbers. Therefore, to reduce costs, cows should be bred for the first time at approximately 14 months so that they can achieve the first calving at approximately 23–24 months. However, it is important to choose the first calving time of the cows according to the climate and breeding needs of their country.

The calving interval is a reproductive trait with low heritability (Pryce et al., 2000). The ideal state of dairy cows is lactation for 10 months and dry milk for 2 months. When the time of artificial insemination is less than 60 days, the conception rate is found to be significantly lower. Do et al. (2013) found that there was a negative correlation between the first two calving intervals and the production life of dairy cows (the correlation coefficient is −0.265), and cows had a longer calving interval and higher milk production with a short production life (Wu, 2007). Calving ease significantly reduces the service life and reproductive performance of dairy cows (Maturana et al., 2007). Although the criteria for calving difficulty vary from country to country, the result is that unassisted cows live longer than cows that require assistance during delivery (Hossein-Zadeh, 2016). Dystocia reduces herd profitability, impairs the reproductive performance of cows in the next breeding cycle, and reduces milk production (Hossein-Zadeh, 2016). Rajala-Schultz and Gröhn (1999) found that the risk of culling for dystocic cows was significantly increased during the first 30 days after calving and at the end of the lactation. In addition, the number of pregnancies (Molina-Coto et al., 2020), open days (Pinedo and Vries, 2010), and calf survival rates significantly increased the risk of cows being eliminated (Yalew et al., 2011). Therefore, there was a low or moderate genetic association between reproductive traits and longevity traits, it is necessary to continue to further examine the correlation between these traits and longevity in follow-up studies of the breeding traits of dairy cows. In particular, the first calving, calving interval, calving ease, and calf survival rates may be the indirect traits that can be used for the early selection for longevity.

Longevity and Disease Traits

There is a unfavorable genetic correlation between longevity and health traits, especially metabolic diseases, which have a strong genetic impact on productive life, and the genetic correlation is −0.98 (Shabalina et al., 2020). Therefore, health traits can be considered as an index to measure productive life. Hadley et al. (2006) found that up to 80% culled dairy cows had the health problems. Therefore, appropriate treatment methods should be adopted to reduce the prevalence of disease in dairy cows to the greatest extent. The main diseases related to the longevity of dairy cows include mastitis, foot disease, metabolic disease, reproductive disease, digestive tract disease, tuberculosis, and brucellosis. Dairy cows suffering from clinical mastitis and reproductive diseases are the most likely to be culled, which have low or moderate negative genetic correlations with longevity (Holtsmark et al., 2008). The culled risk during the remaining lactation period increases after cow suffering from mastitis, and mastitis resistance is positively correlated with productive life (Neerhof et al., 2000). Therefore, longevity can be extended by selecting cows with high resistance to mastitis. Reproductive diseases, such as abnormal estrus, retained placenta, ovarian quiescence, ovarian cysts, persistent corpus luteum, and endometritis abortion can shorten the lifespan, which can prolong the calving interval and decrease milk production (Pascottini and Leblanc, 2020).

Foot disease affects the longevity of dairy cows by affecting their activities, feeding behavior, and production performance. Charfeddine and Perez-Cabal (2016) found that the existence of ulcers or white lines in foot disease was associated with low milk yield of Spanish Holstein cows, the existence of ulcers, or white line disease in early lactation will extend day open and long calving intervals. Metabolic disease is an important cause for the involuntary culling of dairy cows. Milk fever is the most important disease affecting the survival rate, followed by ketosis, fatty liver, and abomasal translocation (Probo et al., 2018). These diseases are related to both longevity and milk production, the risk of dairy cow culling increases when one metabolic disease coexists with another. More importantly, the metabolic stress caused by metabolic diseases can damage the mitochondria and further affect the longevity of dairy cows. Ketosis is the most important metabolic disease in all countries (Huber et al., 2016); ketosis will secrete less milk and milk fat content. The sick cow will no longer continue to secrete milk with the aggravation of the disease, consequently, they will be involuntary culled and the production life will be shortened.

Bovine infectious tuberculosis and brucellosis are two major infectious diseases in dairy cattle breeding, which not only endanger human health but also shorten the longevity of dairy cows (Ma and Xie, 2020). Breast tuberculosis is an infection of the mammary gland that results in a lump on the outside of the breast, causing in low milk production and short lifespan of cows (Yang and Zhou, 2010). Brucellosis is a common epidemic disease in pastures; cows suffering from brucellosis are prone to abortion, retained placenta, and long-term infertility (Wang, 2020). Additionally, the longevity traits of dairy cows in various countries with the increasing understand, the longevity of dairy cows should not only be studied after calving, but also before calving. The calf survival rate is affected by respiratory disease and diarrhea, which occur annually in most intensive dairy cows (Timsit et al., 2017). According to the National Animal Health Surveillance System of the United States, half of the deaths of calves in the United States dairy industry are caused by diarrhea (Yong-II and Kyoung-Jin, 2014). Therefore, a cow herd with high disease resistance will be beneficial to extend the longevity of the cows.

Longevity and Nutrition

Nutritional factors are directly or indirectly related to the growth and development of animals, physiological and biochemical indices, and immune indices (Michael et al., 2019). High protein levels can have various toxic effects on the ovum, oosperm, and embryo, and it will reduce prostaglandin synthesis and progesterone, delay estrus, and ovulation. Moreover, low protein content in the diet will cause reproductive diseases, such as delayed follicular development, abnormal estrus, low conception rate, and retained placenta (Liu, 2017). Excessive carbohydrate content in the diet will lead to rumen acidosis in dairy cows (Agovino, 2018). An imbalance in the calcium and phosphorus ratio can readily cause the development of metabolic disorders and osteoporosis. The feeding, ruminating, and digestion characteristics of cows are significantly different at different ages. Therefore, total mixed ration (TMR) technology should be promoted according to the nutritional needs of dairy cows at different ages, growth stages, pregnancies, and lactation (Schingoethe, 2017). The roughage and concentrate should be cut, stirred, mixed, and fed in proportion to ensure a balanced nutritional intake for dairy cows. Additionally, regular sampling and analysis of the TMR mixed diet should be carried out to reduce the inconsistency between the trough feed and the allocated feed in the mixed diet, because imbalance nutrient can result in diseases and thereby shorten the longevity of dairy cows.

Longevity and Management

The improvement of cow comfort and welfare is conducive to improve the profitability of dairy farms and longevity. Bouffard et al. (2017) studied Holstein cows in 100 tiestall dairy farms and found that most cows tied in the cattle pen had an increase in the incidence of foot disease and shortened the lifespan by comparison free-ranging of cows. Fuerst-Waltl et al. (2018) found that the culling rate of free-ranging Simmental cattle was 15% lower than stall-fed Simmental cattle. Exercise may be an important factor for improving the health of dairy cows. Therefore, it is important to adhere to the concept of animal welfare and provide more consultation services for the breeding of dairy cows (Karin et al., 2018). It is important to control the herd size (Sawa et al., 2016), keep warm in winter, and provide sufficient ventilation to prevent heatstroke in summer. In addition, flooring type and slipperiness, barn cleanliness, bedding type and quantity, and stall design are all associated with increased odds of lameness (Wu et al., 2018). Additionally, managers should regularly assess performance and health status data in accordance with the range management system to detect adverse trends for timely treatment. In daily life, cows should have the opportunity to bask in the sun every day, sufficiently brush the cow's body, which ensures they are clean and dry at all times to promote blood circulation and metabolism and grasp the suitable age for breeding. The breasts should be massaged every day, which maintains normal function of breast and promotes lactation. Ensuring cows should have appropriate daily activity to improve their physique, feed conversion time, and efficiency, which in turn ultimately improve the milk yield. In particular, the dirt between cows toes should be cleaned regularly, and the cows' feet should be pruned. Protecting the cows' feet in strict accordance with the operating procedures of foot repair technology will reduce the occurrence of limb foot disease (He, 2011).

LONGEVITY IN BREEDING SYSTEMS IN VARIOUS COUNTRIES

For many years, most selection indices worldwide focused on increasing milk production (Miglior et al., 2005). With relative emphasis on production in various countries, the selection indices have gradually shifted toward a more balanced breeding goal of improving production, especially protein yield and percentage, longevity, udder health, conformation, and reproduction (Vanraden, 2002). Longevity of dairy cows has been studied in many countries since the 1950s (Zhang et al., 2020). Longevity traits are heritable and can be improved by selecting (Miglior et al., 2017). The definition and model of longevity traits are also developing with further research on longevity traits in developed countries (Miglior et al., 2005). However, there is currently no consensus on definition of longevity trait or on the methodology for the evaluation across countries. Therefore, the definition and model for longevity trait selection, and the longevity trait weight in the comprehensive selection index of each country is inconsistent.

The United States considers productive life, which combines direct longevity defined as total months in milk through 84 mo of age, along with somatic cell score, milk yield, milk fat yield, milk protein yield, and some conformation traits (Cruickshank et al., 2002), and usually analyzed by single trait-best linear unbiased prediction (BLUP)-animal model. Canada considers longevity is a total five traits, survival status from the first calving to day 120, day 120 to 240, day 240 to second calving, second to third calving, and third to fourth calving, which combines direct longevity somatic cell score, milking speed, non-return rate, calving to first service interval, and some conformation traits, and usually analyzed by five traits-BLUP-animal model (Sewalem et al., 2007). Nodic considers longevity is a total of five traits, the partial productive life from the first to the second, third, fourth, fifth, and sixth calvings, which usually analyzed by muti-traits-BLUP-animal model. Germany considers longevity is a total of nine traits, survival status from day 0 to 49, day 50 to 249, and day 250 to the next calving after the current calving among the first three lactations, which usually analyzed by muti-traits -BLUP -animal model (Zhang et al., 2020). France considers productive life, which combines direct longevity along with somatic cell count, clinical mastitis, and some reproduction and conformation traits, which usually analyzed by Single trait-survival

analysis–Sire-maternal grand sire model (Zhang et al., 2020). Australia considers longevity is survival status from the first to the second, third, fourth, fifth, sixth, seventh, and eighth calvings, which usually analyzed by muti-traits-BLUP-animal model (Zhang et al., 2020).

Longevity was included in the total performance index (TPI) by the American Holstein Dairy Association in 2001, which accounted for 8% of the latest TPI established in 2021. Canada added longevity to the lifetime performance index (LPI) in August 2001, longevity accounted for 5% of the latest LPI index in 2021. In 1990, longevity is added to the nordic total merit (NTM), and longevity accounted for 7% of this index. In Germany, milk yield, conformation, and functional traits were first included in the selection index relative zuchtwert gesamt (RZG) in 1997, function longevity accounted for 20% of the latest RZG index in 2020. Longevity accounted for 5% of the French Index de Synthèse UPRAISU in 2021. The weight of longevity traits in the TPI of Australian and Dutch dairy cows was 8 and 12%, respectively. Therefore, longevity plays an increasingly important role in dairy cattle breeding in various countries.

BREEDING AND GENETIC SELECTION

Improvements and genetic selection for milk yield have led to substantially steady increase in milk production over recent decades in many countries, including Canada, the United States, and throughout Europe and Australasia (Schuster et al., 2020). However, due to only focus on the improvement of milk yield, other traits, and the disease resistance of cows decrease (Ingvartsen et al., 2003). Therefore, the aim of breeding in various countries gradually shift to more balanced breeding, functional traits, and especially longevity traits. All functional traits have also been included in the selection indices of respective countries (Miglior et al., 2005). There are two approaches to study longevity, the first is to select longevity directly, and the second is to select the underlying functional traits as the breeding goal. The indirectly selects for traits that are difficult to measure or do not have complete data records (Vollema, 1998), and it is important to determine the heritabilities and correlations of these traits.

Since longevity is a low-heritability trait globally (Schuster et al., 2020), and true longevity of cows cannot be known until cows were culled, which makes a longest generation interval (Ducrocq et al., 1988). The emergence of genomic selection technology has shortened the generation interval, improved the breeding value, and accelerated the genetic progress of longevity (Liu et al., 2011). Although genomic selection accelerates genetic improvement, cows can be replaced slightly faster, and raise all born heifer calves to replace cows is not the most profitable strategy in many cases (Vollema, 1998). Thus, the optimal asset replacement theory is only the necessary number of heifers is selected and the lowest genetic merit is sold, which can help to expand the dominant population (Weigel et al., 2012).

In addition, the underlying molecular mechanisms of longevity remain not to be incompletely understand, which slow research progress on the longevity of dairy cows. Molecular breeding methods can more accurately determine the genetic potential for specific traits. At present, many molecular markers, genotypes and metabolites are being assessed in terms of their correlations with the longevity of dairy cows, which can assist in the selection of longevity traits in dairy cows.

Seeker et al. (2018) found that bovine relative leukocyte telomere length (RLTL) was a heritable trait, and the association with longevity traits can be used in breeding programs to increase the lifespan of dairy cows. Bovine telomeres shorten with age, and the relationship between RLTL at different life stages and the productive life of dairy cows has not been explored. RLTL is generally positively correlated with longevity in humans and vertebrates, but telomere length is negatively correlated with longevity, which implies that short telomeres are associated with an increased mortality risk (Wilbourn et al., 2018). The only other study on the relationship between RLTL and productive life displays a significant but weak correlation in dairy cows (Brown et al., 2012). Therefore, a follow-up study is needed to assess the longevity benefits of a strategy of breeding according to RLTL through genome-wide association analysis. Huber et al. (2016) sequenced 19 dairy cows plasma by using a metabolomics approach to identify the metabolites of long-chain acylcarnitine, spermidine, and biogenic amines associated with prolonged production life, but these metabolites have not yet been confirmed. If these potential new biomarkers are confirmed, they can be used for the genetic selection of bulls and dairy cattle breeding, and increase the number of dairy cows with "extended life" metabolic genetic traits. Ioannidis et al. (2018)found that plasma microRNAs were associated with telomere length, milk yield, milk composition, somatic cell count, reproduction, and blood metabolites related to body energy balance and metabolic stress, suggesting that these microRNAs may be significantly associated with a productive life. Additionally, researchers from various countries have used genome-wide association studies (GWAS) and single-nucleotide polymorphism (SNP) to identify many genes, quantitative trait loci (QTLs), and SNPs are significantly related to longevity (**Table 3**), indeed, a large number of genes (**Figure 3**) and QTL (**Figure 4**) have been discovered from cattle QTL projects.[1] At present, these genes and polymorphisms are also being studied and tested to determine their relationship with longevity and extend longevity of dairy cows.

CONCLUSION AND FUTURE PERSPECTIVES

The dairy industry is gradually moving toward intensive, large-scale, standardized, and mechanized breeding worldwide.

[1]https://www.animalgenome.org/cgi-bin/QTLdb/index

TABLE 3 | Genes, single-nucleotide polymorphisms (SNPs), and quantitative trait loci (QTL) related longevity of cattle.

Author	Country	Breed	Definition	Analytical method	chromosome	Gene, SNP, and QTL
Pryce et al., 2010	Australia	Holstein, Jersey	Longevity	SNP-by-SNP	6, 26	13.5–23.7 Mb (6 SNP), 70.4–75.6 Mb(5 SNP), and 0.33–1.46 Mb(6 SNP)
Ashwell and Tassel, 1999	United States	Holstein	Productive life	Microsatellite Marker	2, 21, 23	BMS2519, BM103, and 513
Saowaphak et al., 2017	Thailand	Holstein	Productive life	GWAS	5, BTAX	SYT1, DOCK11, IL13RA1, and KLHL13
Nayeri et al., 2017	Canada	Holstein	Longevity	GWAS	18, 6, 14	CTU1, NPFFR2, CYHR1, CPSF1, DGAT1, and GRINA
Zhao, 2013	China	Holstein	Productive life	SSCP	12	FoxO1(TT)
Wang et al., 2018	China	Holstein	Productive life	SNP mutation analysis	2, 3	CXCR1-816 A>C, LF-28A>C, and SLC35A3-66G>T
Steri et al., 2019	Germany	Holstein	Longevity	GWAS	16, 30	43.25–43.69 Mb, 17.56–18.06 Mb
Szyda et al., 2011	Poland	Holstein	Functional longevity	SNP mutation analysis	3, 4, 6, 14, 23	LEPR, LEP, ABCG2, DGAT1, and BTN1A1
Bjelka and Novák, 2019	Czech Republic	Czech Red Pied	Longevity	HiSeq X-Ten	6	TLR1 798C>T, TLR1 1762G>A
Khatib et al., 2005	United States	Holstein	Productive life	SNP mutation analysis	21	PI
Tassell et al., 2000	United States	Holstein	Productive life	Microsatellite Markers	2, 21, 23	BMS2519, BM103, and MB026
Mészáros et al., 2014	Sweden	Fleckvieh	Longevity	GWAS	5, 6, 8, 14	SYT10, ADAMTS3, NTRK2, and SNTG1
Cole et al., 2011	United States	Holstein	Productive life	GWAS	7, BTAX	INSR, LOC520057, and GRIA3
Steri et al., 2019	Italy	Holstein	Longevity	GWAS	16, 30,	UBIAD1, MTOR, ANGPTL7, EXOSC10, SRM, MASP2, TARDBP, CASZ1, GPC3, and PHF6
Zhang et al., 2016	Netherland	Holstein, Red Dairy Cattle, Jersey	Longevity	GWAS	23, 9, 10	KCNK16, PPP1R14C, and GCH1
Kaupe et al., 2007	Germany	German Holstein	productive lifespan	SNP mutation analysis	21	UTMP SNP1296
Zhang et al., 2021b	China	Chinese Holsteins	Productive life; Herd life; milking life; Partial Lifespan	GWAS	2, X, 9, 18, 4, 20, 25	RPRM, GRIA3, GTF2H5, CA5A, CACNA2D1, FGF10, and DNAJA3
Shabalina et al., 2020	Germany	Holstein	Length of productive life	GWAS	4, 10, 13, 28	ETV1, ONECUT1, MACROD2, and SIRT1

With the increasing attention being paid to the breeding value, lifetime benefits, climate change, and environmental sustainability of dairy cattle globally, the focus of dairy cattle breeding has begun to shift to more balanced breeding goals, including longevity, health, welfare, milk yield, milk quality, and environmental sustainability. Longevity have been incorporated into breeding programs in many developed countries according to the dairy production importance (Miglior et al., 2005), and have been selected for longevity traits. Extending of dairy cow longevity has become an urgent need for the development of the dairy industry. Nevertheless, longevity does not yet include into national selection indices in many developing countries because of the complexity. For example, the newly revised Chinese Dairy Performance Index (CPI) in 2020 only includes milk protein content, milk fat content, somatic cells, conformation, lactation system, and feet, and longevity has not been included in the selection index. To increase the breeding of longevity traits, it should be included in selection indices in each country, because longevity trait has economic value as its improvement that can reduce production costs (Allaire, 1981). It is worth noting that the correlation between longevity and milk yield remains unclear. In the future, if longevity is included in the selection index, it will be necessary to comprehensively consider whether it will affect the milk yield of dairy cows.

The short longevity of dairy cows not only seriously affects productivity, but also hinders the scope for selection for other traits. In traditional breeding, researchers in various countries have used different models and trait definitions to directly or indirectly select for longevity traits (Gill and Allaire, 1976; Hoque and Hodges, 1980; Buenger et al., 2001). Multiple-trait evaluations combining indirect measures of longevity with direct measures are helpful to improve the accuracy of longevity evaluations (Miglior et al., 2017). In addition, the terms used to describe the longevity trait are inconsistent across countries, and if all terms are used interchangeably, there will be inconsistencies and ambiguities

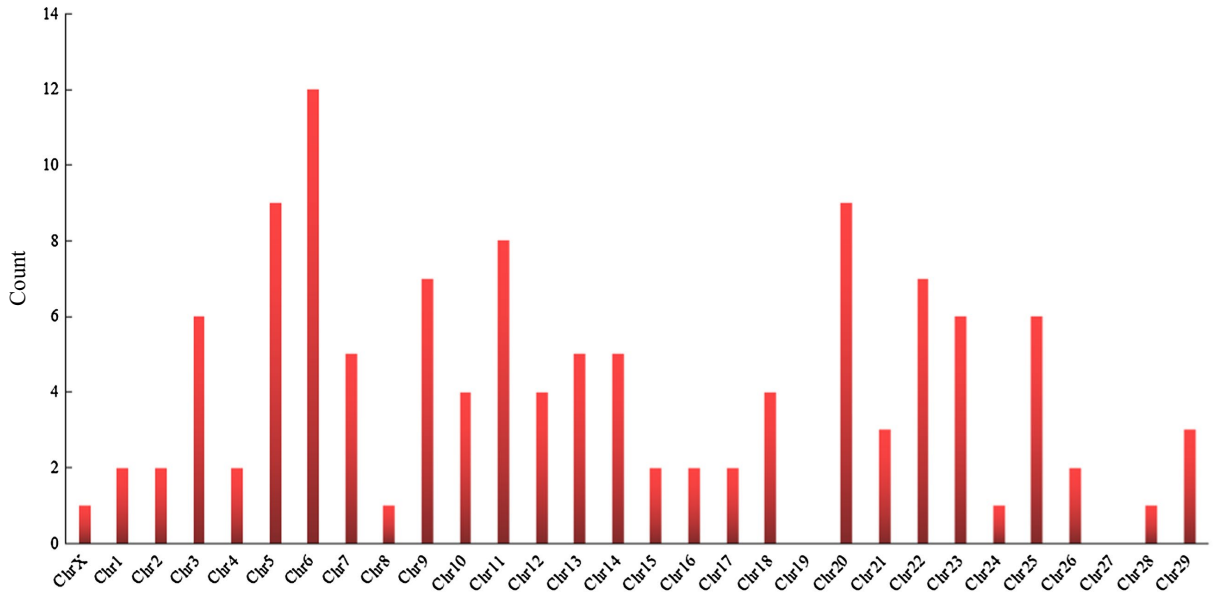


FIGURE 3 | Chromosomal distribution of genes associated with productive life.

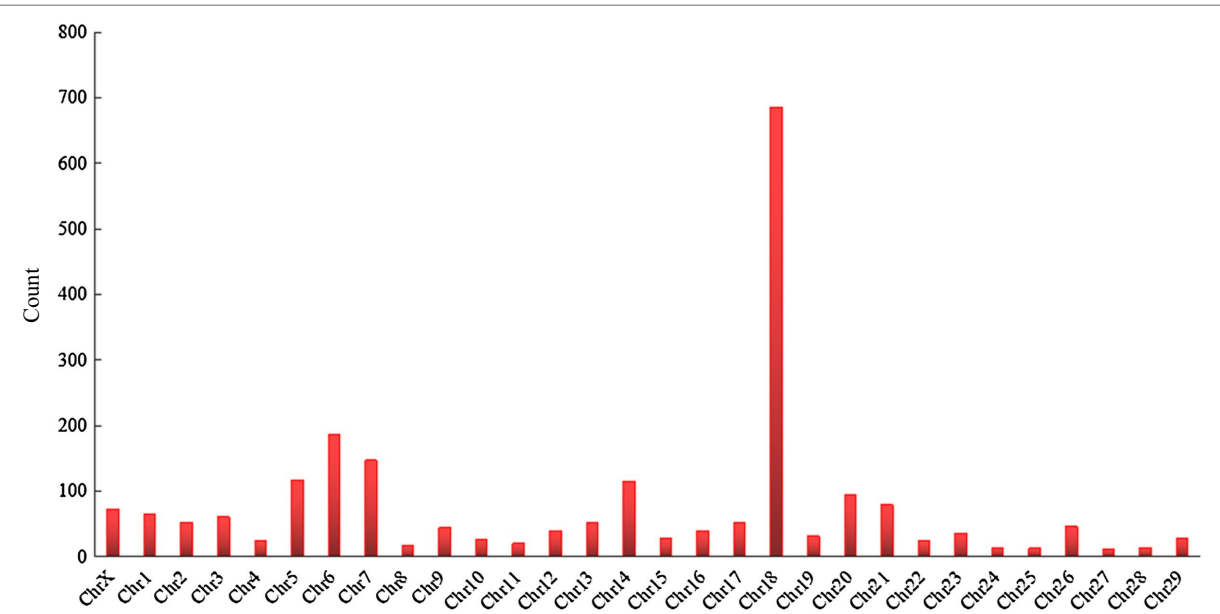


FIGURE 4 | Chromosomal distribution of QTL associated with productive life.

in the definitions of longevity (Caraviello et al., 2004; Brickell and Wathes, 2011; Raguz et al., 2011). Therefore, it is necessary to standardize the different terms related to longevity traits, and researchers can choose different definitions of longevity depending on the purpose of their study.

With the development of molecular technology, genome selection can significantly improve the genetic improvement speed of dairy cows and shorten the generation interval. However, genomic selection accelerates herd replacement, the accuracy gradually decreases due to recombination and the linkage between SNPs and causal genes disappears with time (Zhang et al., 2021a). Longevity traits have low heritability and are highly influenced by the environment, since dairy cattle are farmed globally, and the environment varies greatly from country to country, so it is important to study the interaction between longevity traits and the environment. In the future, it will be particularly important to correctly identify the early indicator traits and genetic markers of longevity, improve the accuracy of longevity

assessment, collect complete records of dairy cows' conformation traits, lactation system, reproductive traits, health traits, limbs, and hooves, and perform dairy herd improvement (DHI) determinations. However, it is important to find a balance between collecting record data and resulting benefits to farmers. Good performance recording in combination with an appropriate model of genetic evaluation and a well-organized selection process have been shown to be useful for breeding of longevity in dairy cows, which will make better progress in this field. In additional, national policies and animal welfare can be a challenge in choosing a longevity herd or the optimal herd in breeding longevity traits in the future when production efficiency and profit are the primary goals. Therefore, the future selection of dairy cattle for longevity breeding will require a fully integrated and balanced breeding model.

AUTHOR CONTRIBUTIONS

HH and TM contributed to the writing, preparing the original draft, investigation and validation. YaM and XW contributed to revision and supervision. YuM contributed to the conceptualization. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by grants from the Modern Agro-industry Technology Research System (CARS-36). The development and application of health control technology for high yield dairy cows (No. 2021BEF01001).

REFERENCES

Agovino, M. (2018). Optimising cow longevity, reproductive performance and milk yield. *Int. Dairy Topics* 17, 7–8.

Albert, D. V. (2020). Symposium review: why revisit dairy cattle productive lifespan. *J. Dairy Sci.* 103, 3838–3845. doi: 10.3168/jds.2019-17361

Allaire, G. S. G. R. (1976). Genetic and phenotypic parameters for a profit function and selection method for optimizing profit in dairy cattle. *J. Dairy Sci.* 7, 1325–1333. doi: 10.3168/jds.S0022-0302(76)84363-9

Allaire, F. R. (1981). Economic consequences of replacing cows with genetically improved heifers. *J. Dairy Sci.* 64, 1985–1995. doi: 10.3168/jds.S0022-0302(81)82801-9

Anim, I., Res, J., Kuevi, D., Dragin, S., and Mirkov, M. (2020). Effect of age at first calving and other non-genetic factors on longevity and production traits in holstein cattle under vojvodina province condition, serbia. *Indian J Anim.* 54, 499–505. doi: 10.18805/ijar.B-1063

Ashwell, M. S., and Tassel, C. P. V. (1999). Detection of putative loci affecting milk, health, and type traits in a US holstein population using 70 microsatellite markers in a genome scan. *J. Dairy Sci.* 82, 2497–2502. doi: 10.3168/jds.S0022-0302(99)75501-3

Bieber, A., Wallenbeck, A., Leiber, F., Fuerst-Waltl, B., Winckler, C., and Gullstrand, P. (2019). Production level, fertility, health traits, and longevity in local and commercial dairy breeds under organic production conditions in Austria, Switzerland, Poland, and Sweden. *J. Dairy Sci.* 102, 5330–5341. doi: 10.3168/jds.2018-16147

Bjelka, M., and Novák, K. (2019). Association of TLR gene variants in a czech red pied cattle population with reproductive traits. *Vet. Immunol. Immunopathol.* 220:109997. doi: 10.1016/j.vetimm.2019.109997

Boettcher, P. J., Jairath, L. K., and Dekkers, J. (1999). Comparison of methods for genetic evaluation of sires for survival of their daughters in the first three lactations. *J. Dairy Sci.* 82:1034. doi: 10.3168/jds.S0022-0302(99)75324-5

Boldman, K. G., Freeman, A. E., Harris, B. L., and Kuck, A. L. (1992). Prediction of sire transmitting abilities for herd life from transmitting abilities for linear type traits1. *J. Dairy Sci.* 75, 552–563. doi: 10.3168/jds.S0022-0302(92)77792-3

Bouffard, V., Passillé, A. M. D., Rushen, J., Vasseur, E., Nash, C. G. R., and Haley, D. B. (2017). Effect of following recommendations for tiestall configuration on neck and leg lesions, lameness, cleanliness, and lying time in dairy cows. *J. Dairy Sci.* 100, 2935–2943. doi: 10.3168/jds.2016-11842

Brickell, J. S., and Wathes, D. C. (2011). A descriptive study of the survival of Holstein-friesian heifers through to third calving on english dairy farms. *J. Dairy Sci.* 94, 1831–1838. doi: 10.3168/jds.2010-3710

Brotherstone, S., Veerkamp, R. F., and Hill, W. G. (1997). Genetic parameters for a simple predictor of the lifespan of Holstein-friesian dairy cattle and its relationship to production. *Anim. Sci.* 65, 31–37. doi: 10.1017/S135772980001626X

Brown, D. E., Dechow, C. D., Liu, W. S., Harvatine, K. J., and Ott, T. L. (2012). Hot topic: association of telomere length with age, herd, and culling in lactating holsteins. *J. Dairy Sci.* 95, 6384–6387. doi: 10.3168/jds.2012-5593

Buenger, A., Ducrocq, V., and Swalve, H. H. (2001). Analysis of survival in dairy cows with supplementary data on type scores and housing systems from a region of Northwest Germany. *J. Dairy Sci.* 84, 1531–1541. doi: 10.3168/jds.S0022-0302(01)70187-7

Čanji, V., Strapak, P., Strapakova, E., and Juhas, P. (2008). Effect of conformation traits on longevity of cows of slovak simmental breed. *Slovak J. Anim. Sci.* 41, 83–90.

Caraviello, D. Z., Weigel, K. A., and Gianola, D. (2004). Analysis of the relationship between type traits and functional survival in US holstein cattle using a weibull proportional hazards model. *J. Dairy Sci.* 87, 2677–2686. doi: 10.3168/jds.S0022-0302(04)73394-9

Chang, Y., Li, X., Zhang, H. L., Yuan, J. G., Guo, G., and Liu, L. (2019). Genetic analysis for temperament of chinese Holstein in Beijing area. *Acta Veterinaria et Zootechnica Sinica.* 50, 44–52. doi: 10.11843/j.issn.0366-6964.2019.04.004

Charfeddine, N., and Perez-Cabal, M. A. (2016). Effect of claw disorders on milk production, fertility, and longevity, and their economic impact in spanish Holstein cows. *J. Dairy Sci.* 100, 653–665. doi: 10.3168/jds.2016-11434

Clasen, J. B., Norberg, E., Madsen, P., Pedersen, J., and Kargo, M. (2017). Estimation of genetic parameters and heterosis for longevity in crossbred danish dairy cattle. *J. Dairy Sci.* 100, 6337–6342. doi: 10.3168/jds.2017-12627

Cole, J. B., Wiggans, G. R., Ma, L., Sonstegard, T. S., Lawlor, T. J., and Crooker, B. A. (2011). Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. *BMC Genomics* 12:408. doi: 10.1186/1471-2164-12-408

Costa, E. V., Ventura, H. T., Veroneze, R., Silva, F. F., Pereira, M. A., and Lopes, P. S. (2020). Estimated genetic associations among reproductive traits in Nellore cattle using Bayesian analysis. *Anim. Reprod. Sci.* 214:106305. doi: 10.1016/j.anireprosci.2020.106305

Cruickshank, J., Weigel, K. A., Dentine, M. R., and Kirkpatrick, B. W. (2002). Indirect prediction of herd life in Guernsey dairy cattle. *J. Dairy Sci.* 85, 1307–1313. doi: 10.3168/jds.S0022-0302(02)74195-7

Dentine, M. R., Mcdaniel, B. T., and Norman, H. D. (1987). Evaluation of sires for traits associated with herdlife of grade and registered Holstein cattle 1. *J. Dairy Sci.* 70, 2623–2634. doi: 10.3168/jds.S0022-0302(87)80332-6

Do, C., Wasana, N., Cho, K., Choi, Y., Choi, T., Park, B., et al. (2013). The effect of age at first calving and calving interval on productive life and lifetime profit in korean holsteins. *Asin Austral. J. Anim.* 26, 1511–1517. doi: 10.5713/ajas.2013.13105

Ducrocq, V. (1997). Survival analysis, a statistical tool for longevity data. Proceedings of 48th Annual Meeting of the EAAP, 25–28.

Ducrocq, V., Quaas, R. L., Pollak, E. J., and Casella, G. (1988). Length of productive life of dairy cows. 2. Variance component estimation and sire evaluation. *J. Dairy Sci.* 71, 3071–3079. doi: 10.3168/jds.S0022-0302(88)79907-5

Ferris, C. P., Patterson, D. C., Gordon, F. J., Watson, S., and Kilpatrick, D. J. (2014). Calving traits, milk production, body condition, fertility, and survival of Holstein-Friesian and norwegian red dairy cattle on commercial dairy farms over 5 lactations—ScienceDirect. *J. Dairy Sci.* 97, 5206–5218. doi: 10.3168/jds.2013-7457

Fuerst-Waltl, B., Aichhorn, T., and Fuerst, C. (2018). Mountain pasturing of rearing stock reduces the culling risk as dairy cows. *Int. J. Anim. Biosci.* 13, 209–212. doi: 10.1017/S1751731118001465

Gill, G. S., and Allaire, F. R. (1976). Relationship of age at first calving, days open, days dry, and herdlife to a profit function for dairy cattle. *J. Dairy Sci.* 59, 1131–1139. doi: 10.3168/jds.S0022-0302(76)84333-0

Grandl, F., Furger, M., Kreuzer, M., and Zehetmeier, M. (2019). Impact of longevity on greenhouse gas emissions and profitability of individual dairy cows analysed with different system boundaries. *Animal* 13, 198–208. doi: 10.1017/S175173111800112X

Grandl, F., Luzi, S. P., Furger, M., Zeitz, J. O., Leiber, F., Ortmann, S., et al. (2016). Biological implications of longevity in dairy cows: 1. Changes in feed intake, feeding behavior, and digestion with age. *J. Dairy Sci.* 99, 3457–3471. doi: 10.3168/jds.2015-10261

Hadley, G. L., Awolf, C., and Harsh, S. B. (2006). Dairy cattle culling patterns, explanations, and implications. *J. Dairy Sci.* 89, 2286–2296. doi: 10.3168/jds.S0022-0302(06)72300-1

Haile-Mariam, M., and Pryce, J. E. (2015). Variances and correlations of milk production, fertility, longevity, and type traits over time in Australian Holstein cattle. *J. Dairy Sci.* 98, 7364–7379. doi: 10.3168/jds.2015-9537

Handcock, R. C., Lopez-Villalobos, N., Mcnaughton, L. R., Back, P. J., Edwards, G. R., and Hickson, R. E. (2020). Body weight of dairy heifers is positively associated with reproduction and stayability. *J. Dairy Sci.* 103, 4466–4474. doi: 10.3168/jds.2019-17545

Hare, E., Norman, H. D., and Wright, J. R. (2006). Survival rates and productive herd life of dairy cattle in the United States—science direct. *J. Dairy Sci.* 89, 3713–3720. doi: 10.3168/jds.S0022-0302(06)72412-2

Haworth, G. M., Tranter, W. P., Chuck, J. N., Cheng, Z., and Wathes, D. C. (2008). Relationships between age at first calving and first lactation milk yield, and lifetime productivity and longevity in dairy cows. *Vet. Rec.* 162:643. doi: 10.1136/vr.162.20.643

He, Y. (2011). How to extend the production and utilization lifespan of dairy cows to the maximum extent. *China Anim. Ind.* 24:90. doi: 10.3969/j.issn.2095-2473.2011.24.054

Hedlund, L., and Løvlie, H. (2015). Personality and production: nervous cows produce less milk. *J. Dairy Sci.* 98, 5819–5828. doi: 10.3168/jds.2014-8667

Holtsmark, M., Heringstad, B., Madsen, P., and Odegard, J. (2008). Genetic relationship between culling, milk production, fertility, and health traits in norwegian red cows. *J. Dairy Sci.* 91, 4006–4012. doi: 10.3168/jds.2007-0816

Hoque, M., and Hodges, J. (1980). Genetic and phenotypic parameters of lifetime production traits in Holstein cows. *J. Dairy Sci.* 63, 1900–1910. doi: 10.3168/jds.S0022-0302(80)83157-2

Horn, M., Knaus, W., Kirner, L., and Steinwidder, A. (2012). Economic evaluation of longevity in organic dairy cows. *Org. Agric.* 2, 127–143. doi: 10.1007/s13165-012-0027-6

Hossein-Zadeh, G. N. (2016). Effect of dystocia on subsequent reproductive performance and functional longevity in Holstein cows. *J. Anim. Physiol. Anim. Nutr.* 100, 860–867. doi: 10.1111/jpn.12460

Huber, K., Dänicke, S., Rehage, J., Sauerwein, H., Otto, W., and Rolle-Kampczyk, U. (2016). Metabotypes with properly functioning mitochondria and anti-inflammation predict extended productive life span in dairy cows. *Sci. Rep.* 6:24642. doi: 10.1038/srep24642

Imbayarwo-Chikosi, V. E., Ducrocq, V., Banga, C. B., Halimani, T. E., Van, W. J. B., and Maiwashe, A. (2016). Impact of conformation traits on functional longevity in south african Holstein cattle. *Anim. Prod. Sci.* 58:481. doi: 10.1071/AN16387

Imbayarwo-Chikosi, V. E., Ducrocq, V., Banga, C. B., Halimani, T. E., Vanwyk, J. B., and Maiwashe, A. (2017). Estimation of genetic parameters for functional longevity in the south african Holstein cattle using a piecewise weibull proportional hazards model. *J. Anim. Breed. Genet.* 134, 364–372. doi: 10.1111/jbg.12264

Imbayarwo-Chikosi, V., Dzama, K., Halimani, T., Wyk, J. V., Maiwashe, A., and Banga, C. (2015). Genetic prediction models and heritability estimates for functional longevity in dairy cattle. *South Afr. J. Anim. Sci.* 45:105. doi: 10.4314/sajas.v45i2.1

Ingvartsen, L. K., Dewhurst, R. J., and Friggens, N. C. (2003). On the relationship between lactational performance and health: is it yield or metabolic imbalance that cause production diseases in dairy cattle? A position paper—sciencedirect. *Livest. Prod. Sci.* 83, 277–308. doi: 10.1016/S0301-6226(03)00110-6

Ioannidis, J., Sánchez-Molano, E., Psifidi, A., Donadeu, F. X., and Banos, G. (2018). Association of plasma microRNA expression with age, genetic background and functional traits in dairy cattle. *Sci. Rep.* 8:12955. doi: 10.1038/s41598-018-31099-w

Jairath, L., Dekkers, J. C., Schaeffer, L. R., Liu, Z., Burnside, E. B., and Kolstad, B. (1998). Genetic evaluation for herd life in Canada. *J. Dairy Sci.* 81, 550–562. doi: 10.3168/jds.S0022-0302(98)75607-3

Jamrozik, J., Fatehi, J., and Schaffer, L. R. (2008). Comparison of models for genetic evaluation of survival traits in dairy cattle, a simulation study. *J. Anim. Breed. Genet.* 125, 75–83. doi: 10.1111/j.1439-0388.2007.00712.x

Jamrozik, J., Mcgrath, S., Kemp, R. A., and Miller, S. P. (2013). Estimates of genetic parameters for stayability to consecutive calvings of Canadian simmentals by random regression models. *J. Anim. Sci.* 91, 3634–3643. doi: 10.2527/jas.2012-6126

Jenko, J., Ducrocq, V., and Kovac, M. (2013). Comparison of piecewise weibull baseline survival models for estimation of true and functional longevity in brown cattle raised in small herds. *Animal* 7, 1583–1591. doi: 10.1017/S1751731113001055

Jenko, J., Perpar, T., and Kovač, M. (2015). Genetic relationship between the lifetime milk production, longevity and first lactation milk yield in slovenian brown cattle breed. *Mljekarstvo* 65, 111–120. doi: 10.15567/mljekarstvo.2015.0205

Karin, A. S., Ian, D., Anki, R., and Ulf, E. (2018). Farm characteristics and management routines related to cow longevity: a survey among swedish dairy farmers. *Acta Vet. Scand.* 60:38. doi: 10.1186/s13028-018-0390-8

Kaupe, B., Brandt, H., Prinzenberg, E. M., and Erhardt, G. (2007). Joint analysis of the influence of CYP11B1 and DGAT1 genetic variation on milk production, somatic cell score, conformation, reproduction, and productive lifespan in german Holstein cattle1. *J. Anim. Sci.* 85, 11–21. doi: 10.2527/jas.2005-753

Kern, E. L., Cobuci, J. A., Costa, C. N., McManus, C. M., Campos, G. S., Almeida, T. P., et al. (2014). Genetic association between herd survival and linear type traits in Holstein cows under tropical conditions. *Ital. J. Anim. Sci.* 13, 665–672. doi: 10.4081/ijas.2014.3419

Kern, E. L., Cobuci, J. A., Costa, C. N., Mcmanus, C. M., and Neto, J. B. (2015). Genetic association between longevity and linear type traits of Holstein cows. *Sci. Agric.* 72, 203–209. doi: 10.1590/0103-9016-2014-0007

Kerslake, J. I., Amer, P. R., O'Neil, P. L., Wong, S. L., Roche, J. R., and Phyn, C. V. C. (2018). Economic costs of recorded reasons for cow mortality and culling in a pasture-based dairy industry. *J. Dairy Sci.* 101, 1795–1803. doi: 10.3168/jds.2017-13124

Khatib, H., Heifetz, E., and Dekkers, J. C. (2005). Association of the protease inhibitor gene with production traits in Holstein dairy cattle. *J. Dairy Sci.* 88, 1208–1213. doi: 10.3168/jds.S0022-0302(05)72787-9

Liu, B. C. (2017). Effects of nutritional factors on reproductive performance of dairy cows. *Modern Anim. Husb. Sci. Technol.* 7:56.

Liu, Z., Seefried, F. R., Reinhardt, F., Rensing, S., and Reents, R. (2011). Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genet. Sel. Evol.* 43:19. doi: 10.1186/1297-9686-43-19

Ma, M., and Xie, D. L. (2020). Suggestions on monitoring, quarantine and prevention and control of two diseases in dairy farms. *Chin. Livest. Poult. Breed.* 016:122.

Maiwashe, A., Nephawe, K. A., and Theron, H. E. (2009). Analysis of stayability in south african angus cattle using a threshold model. *S. Afr. J. Anim. Sci.* 39, 55–60. doi: 10.4314/sajas.v39i1.43546

Mao, J. (2015). Genetic analysis betweet type traits, milk production traits, SCS and longevity traits of holtein cattle in shanghai. master degree thesis/dissertation. Jiangsu: Nanjing Agricultural University.

Martinez, G. E., Koch, R. M., Cundiff, L. V., Gregory, K. E., and Vleck, L. D. V. (2004). Genetic parameters for six measures of length of productive life and three measures of lifetime production by 6 yr after first calving for Hereford cows. *J. Anim. Sci.* 82, 1912–1918. doi: 10.2527/2004.8271912x

Maturana, E., Ugarte, E., and González-Recio, O. (2007). Impact of calving ease on functional longevity and herd amortization costs in basque holsteins using survival analysis. *J. Dairy Sci.* 90, 4451–4457. doi: 10.3168/jds.2006-734

Mészáros, G., Eaglen, S., Waldmann, P., and Slkner, J. (2014). A genome wide association study for longevity in cattle. *Open J. Genet.* 4, 46–55. doi: 10.4236/ojgen.2014.41007

M'hamdi, N., Aloulou, R., Bouallegue, M., Brarc, S. K., and BenHamoudad, M. (2010). Study on functional longevity of tunisian Holstein dairy cattle using

a weibull proportional hazard model. *Livest. Sci.* 132, 173–176. doi: 10.1016/j.livsci.2010.05.011

Michael, A., Occhio, D. X., Pietro, S., Baruselli, B., and Campanile, C. G. (2019). Influence of nutrition, body condition, and metabolic status on reproduction in female beef cattle: A review. *Theriogenology* 125, 277–284. doi: 10.1016/j.theriogenology.2018.11.010

Miglior, F., Fleming, A., Malchiodi, F., Brito, L. F., Martin, P., and Baes, C. F. (2017). A 100-year review: identification and genetic selection of economically important traits in dairy cattle. *J. Dairy Sci.* 100:10251. doi: 10.3168/jds.2017-12968

Miglior, F., Muir, B. L., and Doormaal, B. V. (2005). Selection indices in Holstein cattle of various countries. *J. Dairy Sci.* 88, 1255–1263. doi: 10.3168/jds.S0022-0302(05)72792-2

Miglior, F., Sewalem, A., Jamrozik, J., Lefebvre, D. M., and Moore, R. K. (2006). Analysis of milk urea nitrogen and lactose and their effect on longevity in Canadian dairy cattle. *J. Dairy Sci.* 89, 4886–4894. doi: 10.3168/jds.S0022-0302(06)72537-1

Mirhabibi, S., Kashan, N., and Gharahveysi, S. (2018). Genetic evaluation of survival traits in the Holstein dairy cows of Iran. *Egypt. J. Vet. Sci.* 49, 71–74. doi: 10.21608/ejvs.2018.3070.1031

Molina-Coto, R., Moore, S. G., Mayo, L. M., Lamberson, W. R., Poock, S. E., Lucy, M. C., et al. (2020). Ovarian function and the establishment and maintenance of pregnancy in dairy cows with and without evidence of postpartum uterine disease. *J. Dairy Sci.* 103, 10715–10727. doi: 10.3168/jds.2020-18694

Mossa, F., and Ireland, J. J. (2019). Anti-müllerian hormone (AMH): a biomarker for the ovarian reserve, ovarian function and fertility in dairy cows. *J. Anim. Sci.* 97, 1446–1455. doi: 10.1093/jas/skz022

Muntean, C. T., Herring, A. D., Riley, D. G., Gill, C. A., Sawyer, J. E., and Sanders, J. O. (2018). Evaluation of F1 cows sired by Brahman, boran, and Tuli bulls for reproductive, maternal, and cow longevity traits. *J. Anim. Sci.* 96, 2545–2552. doi: 10.1093/jas/sky169

Najafabadi, H. A., Mahayari, S. A., Edriss, M. A., and Strapakova, E. (2016). Genetic analysis of productive life length in Holstein dairy cows using weibull proportional risk model. *Arch. Anim. Breed.* 59, 387–393. doi: 10.5194/aab-59-387-2016

Nayeri, S., Sargolzaei, M., Abo-Ismail, M. K., Miller, S., Schenkel, F., and Moore, S. S. (2017). Genome-wide association study for lactation persistency, female fertility, longevity, and lifetime profit index traits in Holstein dairy cattle. *J. Dairy Sci.* 100, 1246–1258. doi: 10.3168/jds.2016-11770

Neerhof, H. J., Madsen, P., Ducrocq, V. P., Vollema, A. R., Jensen, J., and Korsgaard, I. R. (2000). Relationships between mastitis and functional longevity in danish black and white dairy cattle estimated using survival analysis. *J Dairy Sci.* 83, 1064–1071. doi: 10.3168/jds.S0022-0302(00)74970-8

Nilforooshan, M. A., and Edriss, M. A. (2004). Effect of age at first calving on some productive and longevity traits in iranian holsteins of the Isfahan province. *J. Dairy Sci.* 87, 2130–2135. doi: 10.3168/jds.S0022-0302(04)70032-6

Overton, M. W., and Dhuyvetter, K. C. (2020). Symposium review: an abundance of replacement heifers: what is the economic impact of raising more than are needed. *J. Dairy Sci.* 103, 3828–3837. doi: 10.3168/jds.2019-17143

Pascottini, O. B., and Leblanc, S. J. (2020). Modulation of immune function in the bovine uterus peripartum. *Theriogenology* 150, 193–200. doi: 10.1016/j.theriogenology.2020.01.042

Pelt, M. L. V., Meuwissen, T. H. E., Jong, G. D., and Veerkamp, R. F. (2015). Genetic analysis of longevity in dutch dairy cattle using random regression. *J. Dairy Sci.* 98, 4117–4130. doi: 10.3168/jds.2014-9090

Perez-Cabal, M. A., Garcia, C., Gonzalez-Recio, O., and Alenda, R. (2006). Genetic and phenotypic relationships among locomotion type traits, profit, production, longevity, and fertility in spanish dairy cows. *J. Dairy Sci.* 89, 1776–1783. doi: 10.3168/jds.S0022-0302(06)72246-9

Pinedo, P. J., and Vries, A. D. (2010). Effect of days to conception in the previous lactation on the risk of death and live culling around calving—ScienceDirect. *J. Dairy Sci.* 93, 968–977. doi: 10.3168/jds.2009-2408

Probo, M., Pascottini, O. B., Leblanc, S., Opsomer, G., and Hostens, M. (2018). Association between metabolic diseases and the culling risk of high-yielding dairy cows in a transition management facility using survival and decision tree analysis. *J. Dairy Sci.* 101, 9419–9429. doi: 10.3168/jds.2018-14422

Pryce, J. E., Coffey, M. P., and Brotherstone, S. (2000). The genetic relationship between calving interval, body condition score and linear type and management traits in registered Holsteins. *J. Dairy Sci.* 83, 2664–2671. doi: 10.3168/jds.S0022-0302(00)75160-5

Pryce, J. E., Haile-Mariam, M., Verbyla, K., Bowman, P. J., Goddard, M. E., and Hayes, B. J. (2010). Genetic markers for lactation persistency in primiparous australian dairy cows. *J. Dairy Sci.* 93, 2202–2214. doi: 10.3168/jds.2009-2666

Raguz, N., Jovanovac, S., Gantner, V., Meszaros, G., and Solkner, J. (2011). Analysis of factors affecting the length of productive life in croatian dairy cows. *Bulg. J. Agric. Sci.* 17, 232–240.

Rajala-Schultz, P. J., and Gröhn, Y. T. (1999). Culling of dairy cows. Part i. effects of diseases on culling in finnish Ayrshire cows. *Prev. Vet. Med.* 41, 195–208. doi: 10.1016/S0167-5877(99)00046-X

Ramos, P. V. B., Silva, F. F. E., Silva, L. O. C. D., Santiago, G. G., and Menezes, G. R. D. O. (2020). Genomic evaluation for novel stayability traits in Nellore cattle. *Bulg. J. Agric. Sci.* 55, 266–273. doi: 10.1111/rda.13612

Saowaphak, P., Duangjinda, M., Plaengkaeo, S., Suwannasing, R., and Boonkum, W. (2017). Genetic correlation and genome-wide association study (GWAS) of the length of productive life, days open, and 305-days milk yield in crossbred Holstein dairy cattle. *Genet. Mol. Res.* 16: gmr16029091. doi: 10.4238/gmr16029091

Sartori, C., Guzzo, N., and Mantovani, R. (2020). Genetic correlations of fighting ability with somatic cells and longevity in cattle. *Animal* 14, 13–21. doi: 10.1017/S175173111900168X

Sasaki, O., Aihara, M., Hagiya, K., Nishiura, A., Ishii, K., and Satoh, M. (2012). Genetic evaluation of the longevity of the Holstein population in Japan using a weibull proportional hazard model. *Anim. Sci. J.* 83, 95–102. doi: 10.1111/j.1740-0929.2011.00943.x

Sawa, A., Bogucki, M., and Niewiadomski, P. (2016). Cow longevity in herds of different milk production levels and herd size. *Anim. Sci.* 55, 261–266.

Schingoethe, D. J. (2017). A 100-year review: Total mixed ration feeding of dairy cows. *J. Dairy Sci.* 100, 10143–10150. doi: 10.3168/jds.2017-12967

Schuster, J. C., Barkema, H. W., Vries, A. D., Kelton, D. F., and Orsel, K. (2020). Invited review: academic and applied approach to evaluating longevity in dairy cows. *J. Dairy Sci.* 103, 11008–11024. doi: 10.3168/jds.2020-19043

Seeker, L. A., Ilska, J. J., Psifidi, A., Wilbourn, R. V., Underwood, S. L., and Fairlie, J. (2018). Bovine telomere dynamics and the association between telomere length and productive lifespan. *Sci. Rep.* 8:12748. doi: 10.1038/s41598-018-31185-z

Setati, M. M., Norris, D., Banga, C. B., and Benyi, K. (2004). Relationships between longevity and linear type traits in Holstein cattle population of southern africa. *Trop. Anim. Health Prod.* 36, 807–814. doi: 10.1023/B:TROP.0000045965.99974.9c

Sewalem, A., Kistemaker, G. J., Ducrocq, V., and Doormaal, B. J. V. (2005). Genetic analysis of herd life in Canadian dairy cattle on a lactation basis using a weibull proportional hazards model. *J. Dairy Sci.* 88, 368–375. doi: 10.3168/jds.S0022-0302(05)72696-5

Sewalem, A., Kistemaker, G. J., Miglior, F., and Doormaal, B. J. V. (2004). Analysis of the relationship between type traits and functional survival in Canadian holsteins using a weibull proportional hazards model. *J. Dairy Sci.* 87, 3938–3946. doi: 10.3168/jds.S0022-0302(04)73533-X

Sewalem, A., Miglior, F., and Kistemaker, G. J. (2010). Analysis of the relationship between workability traits and functional longevity in Canadian dairy breeds. *J. Dairy Sci.* 93, 4359–4365. doi: 10.3168/jds.2009-2969

Sewalem, A., Miglior, F., Kistemaker, G. J., Sullivan, P., and Doormaal, B. J. V. (2008). Relationship between reproduction traits and functional longevity in Canadian dairy cattle. *J. Dairy Sci.* 91, 1660–1668. doi: 10.3168/jds.2007-0178

Sewalem, A., Miglior, F., Kistemaker, G. J., Sullivan, P., Huapaya, G., and Doormaal, B. J. V. (2007). Short communication: modification of genetic evaluation of herd life from a three-trait to a five-trait model in Canadian dairy cattle[J]. *J. Dairy Sci.* 90, 2025–2028. doi: 10.3168/jds.2006-719

Shabalina, T., Yin, T., and Knig, S. (2020). Influence of common health disorders on the length of productive life and stayability in german Holstein cows. *J. Dairy Sci.* 103, 583–596. doi: 10.3168/jds.2019-16985

Short, T. H., and Lawlor, T. J. (1992). Genetic parameters of conformation traits, milk yield, and herd life in holsteins. *J. Dairy Sci.* 75, 1987–1998. doi: 10.3168/jds.S0022-0302(92)77958-2

Smith, C., and Westgarth, D. R. (1957). The use of survival time in the analysis of neutralization tests for serum antibody surveys. *J. Hyg.* 55, 224–238. doi: 10.1017/S0022172400037128

Stanojević, D., Dragan, R., Bogdanović, V., Raguž, N., Kučević, D., and Popovac, M. (2018). Genetic trend of functional productive life in the population of black and white cattle in Serbia. *Genetika* 50, 855–862. doi: 10.2298/GENSR1803855S

Steri, R., Moioli, B., Catillo, G., Galli, A., and Buttazzoni, L. (2019). Genome-wide association study for longevity in the Holstein cattle population. *Animal* 13, 1350–1357. doi: 10.1017/S1751731118003191

Stokes, M. (2019). Andrea cox: trapped in a vice: the consequences of confinement for young people New Brunswick. *J. Youth Adolesc.* 6, 313–314. doi: 10.1007/s10964-019-01023-0

Strapáková, E., Strapák, P., and Candrák, J. (2019). Genetic trend of length of productive life in Holstein and slovak simmental cattle in Slovakia. *Acta Univ. Agric. Silvic. Mendel. Brun.* 67, 1227–1234. doi: 10.11118/actaun201967051227

Szyda, J., Morek-Kope, M., Komisarek, J., and Zarnecki, A. (2011). Evaluating markers in selected genes for association with functional longevity of dairy cattle. *BMC Genet.* 12:30. doi: 10.1186/1471-2156-12-30

Tassell, C. P. V., Ashwell, M. S., and Sonstegard, T. S. (2000). Detection of putative loci affecting milk, health, and conformation traits in a US holstein population using 105 microsatellite markers. *J. Dairy Sci.* 83:1872. doi: 10.3168/jds.S0022-0302(00)75058-2

Timsit, E., Hallewell, J., Booker, C., Tison, N., Amat, S., Trevor, W., et al. (2017). Prevalence and antimicrobial susceptibility of mannheimia haemolytica, pasteurella multocida, and histophilus somni isolated from the lower respiratory tract of healthy feedlot cattle and those diagnosed with bovine respiratory disease. *Vet. Microbiol.* 208, 118–125. doi: 10.1016/j.vetmic.2017.07.013

Tsuruta, S., Misztal, I., and Lawlor, T. J. (2005). Changing definition of productive life in US holsteins: effect on genetic correlations. *J. Dairy Sci.* 88, 1156–1165. doi: 10.3168/jds.S0022-0302(05)72782-X

Uetake, K., Kilgour, R. J., Ishiwata, T., and Tanaka, T. (2004). Temperament assessments of lactating cows in three contexts and their applicability as management traits. *Anim. Sci. J.* 75, 571–576. doi: 10.1111/j.1740-0929.2004.00229.x

Vacek, M., Štípková, M., Němcová, E., and Bouška, J. (2006). Relationships between conformation traits and longevity of Holstein cows in the Czech Republic. *Czech. J. Anim. Sci.* 51, 327–333. doi: 10.17221/3946-CJAS

Vanraden, P. M. (2002). Selection of dairy cattle for lifetime profit. Proceedings of World Congress on Genetics Applied to Livestock Production Montpellier France August Session. Available at: https://aipl.arsusda.gov/publish/other/2002/7thWCGALP_VanRaden_Commun01-21.pdf (Accessed July 4, 2021).

Vanraden, P. M., and Klaaskate, E. J. H. (1993). Genetic evaluation of length of productive life including predicted longevity of live cows. *J. Dairy Sci.* 76, 2758–2764. doi: 10.3168/jds.S0022-0302(93)77613-4

Vollema, A. R. A. R. (1998). Selection for longevity in dairy cattle. Available at: https://edepot.wur.nl/206306 (Accessed July 4, 2021).

Vollema, A. R., and Groen, A. F. (1996). Genetic parameters of longevity traits of an upgrading population of dairy cattle. *J. Dairy Sci.* 79, 2261–2267. doi: 10.3168/jds.S0022-0302(96)76603-1

Vollema, A. R., and Groen, A. F. (1997). Genetic correlations between longevity and conformation traits in an upgrading dairy cattle population. *J. Dairy Sci.* 80, 3006–3014. doi: 10.3168/jds.S0022-0302(97)76267-2

Vries, A. D., and Marcondes, M. I. (2020). Review: overview of factors affecting productive lifespan of dairy cows. *Animal* 14, s155–s164. doi: 10.1017/S1751731119003264

Vukašinović, N., Moll, J., and Künzi, N. (1995). Genetic relationships among longevity, milk production, and type traits in swiss brown cattle. *Livest. Prod. Sci.* 41, 11–18.

Vukasinovic, N., Moll, J., and Künzi, N. (1997). Analysis of productive life in swiss brown cattle. *J. Dairy Sci.* 80, 2572–2579. doi: 10.3168/jds.S0022-0302(97)76213-1

Wang, D. L. (2020). Causes and prevention measures of bovine brucellosis. *Jilin Anim. Husb. Vet. Med.* 388, 78–80.

Wang, D. W., Cao, Z. J., Li, S. L., Huang, W. M., Wang, Y., Wen, W., et al. (2010). The relationship between content of milk urea nitrogen and reproductive performance of dairy cows. *Acta Zoonutr. Sin.* 22, 1509–1514. doi: 10.3969/j.issn.1006-267x.2010.06.007

Wang, M. Q., Ni, W., Tang, C., Guo, J. H., Zhang, H. M., and Li, M. (2018). Association analysis on the SNP of LF-131C>T and LF-28A>C with milk performance, clinical mastitis and lifetime for chinese Holstein. *J. Agr. Biotechnol.* 26, 811–818.

Wasana, N., Cho, G. H., Park, S. B., Kim, S. D., and Do, C. H. (2015). Genetic relationship of productive life, production and type traits of korean holsteins at early lactations. *Asian-Australas. J. Anim. Sci.* 28, 1259–1265. doi: 10.5713/ajas.15.0034

Weigel, K. A., Hoffman, P. C., Herring, W., and Lawlor Jr., T. J. (2012). Potential gains in lifetime net merit from genomic testing of cows, heifers, and calves on commercial dairy farms. *J. Dairy Sci.* 95:2215–2225. https://doi.org/10.3168/jds.2011-4877

Weigel, K. A., Lawlor, T. J., Raden, V. P. M., and Wiggans, G. R. (1998). Use of linear type and production data to supplement early predicted transmitting abilities for productive life. *J. Dairy Sci.* 81:2040. doi: 10.3168/jds.S0022-0302(98)75778-9

Wilbourn, R. V., Moatt, J. P., Froy, H., Craig, A. W., Daniel, H. N., and Boonekamp, J. J. (2018). The relationship between telomere length and mortality risk in non-model vertebrate systems: a meta-analysis. *Philos. Trans. Roy. Soc. B. Biol. Sci.* 373:20160447. doi: 10.1098/rstb.2016.0447

Wu, H. C. (2007). Relationships between body type score and FHL, 305 milk, CI in Holstein cow. master degree Thesis/dissertation. Shandong: Shandong Agricultural University.

Wu, X. D., Feng, W. Y., Zhou, Q. M., Xu, X., and Zhang, Y. (2018). Causes of hoof disease in dairy cows and common treatments. *Chin. J. Anim. Husb. Vet. Med.* 496:77. doi: 10.3969/J.ISSN.1671-6027.2018.04.061

Yalew, B., Lobago, F., and Goshu, G. (2011). Calf survival and reproductive performance of Holstein-Friesian cows in Central Ethiopia. *Trop. Anim. Health Prod.* 43, 359–365. doi: 10.1007/s11250-010-9697-9

Yang, X. B., and Zhou, G. W. (2010). Diagnosis and treatment of major diseases of dairy cows. *Modern Anim. Husb. Sci. Technol.* 7:176.

Yazdi, M. H., Thompson, R., Ducrocq, V., and Visscher, P. (1999). A comparison of two survival analysis methods with the number of lactations as a discrete time variate. *School Biol. Sci.* 48–52.

Yong-II, C., and Kyoung-Jin, Y. (2014). An overview of calf diarrhea—infectious etiology, diagnosis, and intervention. *J. Vet. Sci.* 15, 1–17. doi: 10.4142/jvs.2014.15.1.1

Zavadilova, L., Nemcova, E., and Stipkova, M. (2011). Effect of type traits on functional longevity of czech Holstein cows estimated from a cox proportional hazards model. *J. Dairy Sci.* 94, 4090–4099. doi: 10.3168/jds.2010-3684

Zavadilová, L., and Stipkova, M. (2012). Genetic correlations between longevity and conformation traits in the czech Holstein population. *Czech. J. Anim. Sci.* 57, 125–136. doi: 10.17221/5566-CJAS

Zavadilova, L., Stipkova, M., Nemcova, E., Bouska, J., and Matejickova, J. (2009). Analysis of the phenotypic relationships between type traits and functional survival in czech fleckvieh cows. *Czech. J. Anim. Sci.* 54, 521–531. doi: 10.17221/29/2009-CJAS

Zhang, Q., Guldbrandtsen, B., Thomasen, J. R., Lund, M. S., and Sahana, G. (2016). Genome-wide association study for longevity with whole-genome sequencing in 3 cattle breeds. *J. Dairy Sci.* 99, 7289–7298. doi: 10.3168/jds.2015-10697

Zhang, S. J., Kou, H. W., Ding, X. T., Liu, X., Cai, W. W., Zhang, Z. J., et al. (2021b). The research progress and application of genomic-wide selection in ruminant genetics and breeding. *Chin. J. Agric. Biotechol.* 29, 571–578. doi: 10.3969/j.issn.1674-7968.2021.03.016

Zhang, H. L., Liu, A. X., Mi, S. Y., Li, X., Luo, H. P., Yan, X. Y., et al. (2020). Review on longevity trait in dairy cattle breeding. *Sci. Agric. Sin.* 53, 4070–4082. doi: 10.3864/j.issn.0578-1752.2020.19.019

Zhang, H., Liu, A., Wang, Y., Luo, H., Yan, X., Guo, X., et al. (2021a). Genetic parameters and genome-wide association studies of eight longevity traits representing either full or partial lifespan in chinese Holsteins. *Front. Genet.* 12:634986. doi: 10.3389/fgene.2021.634986

Zhang, F., Qiu, W., Liu, X. J., Chang, J., Qiu, J. Q., Liu, X.-H., et al. (2011). Effects of dietary protein level on milk performance and urea nitrogen content of lactating dairy cows. *J. Hebei Agric. Sci.* 15:87.

Zhao, J. Q. (2013). Research on molecular markers and productive life, linear type traits of holstein cattle in china. master degree thesis/dissertation. Wuhan: Huazhong Agricultural University.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Hu, Mu, Ma, Wang and Ma. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 August 2021
doi: 10.3389/fgene.2021.697549

Evidence That the Etiology of Congenital Hypopituitarism Has a Major Genetic Component but Is Infrequently Monogenic

Youn Hee Jee[1], Mariam Gangat[2], Olga Yeliosof[2], Adrian G. Temnycky[1], Selena Vanapruks[1], Philip Whalen[1], Evgenia Gourgari[3], Cortney Bleach[4], Christine H. Yu[5], Ian Marshall[2], Jack A. Yanovski[1], Kathleen Link[6], Svetlana Ten[7], Jeffrey Baron[1†] and Sally Radovick[2†]*

[1] Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, United States, [2] Division of Pediatric Endocrinology Rutgers Robert Wood Johnson Medical School Child Health Institute of New Jersey, New Brunswick, NJ, United States, [3] Division of Pediatric Endocrinology, MedStar Georgetown University Hospital, Washington, DC, United States, [4] Division of Pediatric Endocrinology, Walter Reed National Military Medical Center, Bethesda, MD, United States, [5] Section of Adult and Pediatric Endocrinology and Metabolism, University of Chicago, Chicago, IL, United States, [6] Division of Pediatric Endocrinology, Pediatric Subspecialists of Virginia, Fairfax, VA, United States, [7] Pediatric Endocrinology, Richmond University Medical Center, Staten Island, NY, United States

Purpose: Congenital hypopituitarism usually occurs sporadically. In most patients, the etiology remains unknown.

Methods: We studied 13 children with sporadic congenital hypopituitarism. Children with non-endocrine, non-familial idiopathic short stature (NFSS) (n = 19) served as a control group. Exome sequencing was performed in probands and both unaffected parents. A burden testing approach was used to compare the number of candidate variants in the two groups.

Results: First, we assessed the frequency of rare, predicted-pathogenic variants in 42 genes previously reported to be associated with pituitary gland development. The average number of variants per individual was greater in probands with congenital hypopituitarism than those with NFSS (1.1 vs. 0.21, mean variants/proband, P = 0.03). The number of probands with at least 1 variant in a pituitary-associated gene was greater in congenital hypopituitarism than in NFSS (62% vs. 21%, P = 0.03). Second, we assessed the frequency of rare, predicted-pathogenic variants in the exome (to capture undiscovered causes) that were inherited in a fashion that could explain the sporadic occurrence of the proband's condition with a monogenic etiology (*de novo* mutation, autosomal recessive, or X-linked recessive) with complete penetrance. There were fewer monogenic candidates in the probands with congenital hypopituitarism than those with NFSS (1.3 vs. 2.5 candidate variants/proband, P = 0.024). We did not find

OPEN ACCESS

Edited by:
Mehdi Pirooznia, National Heart, Lung, and Blood Institute (NHLBI), United States

Reviewed by:
M. Dattani, University College London, United Kingdom
Zirui Dong, The Chinese University of Hong Kong, China

***Correspondence:**
Youn Hee Jee
jeeyh@mail.nih.gov
orcid.org/0000-0001-7227-5925

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: *19 April 2021*
Accepted: *12 July 2021*
Published: *11 August 2021*

Citation:
Jee YH, Gangat M, Yeliosof O, Temnycky AG, Vanapruks S, Whalen P, Gourgari E, Bleach C, Yu CH, Marshall I, Yanovski JA, Link K, Ten S, Baron J and Radovick S (2021) Evidence That the Etiology of Congenital Hypopituitarism Has a Major Genetic Component but Is Infrequently Monogenic. Front. Genet. 12:697549. doi: 10.3389/fgene.2021.697549

any candidate variants (0 of 13 probands) in genes previously reported to explain the phenotype in congenital hypopituitarism, unlike NFSS (8 of 19 probands, $P = 0.01$).

Conclusion: Our findings provide evidence that the etiology of sporadic congenital hypopituitarism has a major genetic component but may be infrequently monogenic with full penetrance, suggesting a more complex etiology.

Keywords: congenital hypopituitarism, monogenic, digenic, ectopic posterior pituitary gland, combined pituitary hormone deficiencies

INTRODUCTION

The embryonic development of the pituitary gland is a complex process involving formation of Rathke's pouch, interactions with the developing ventral diencephalon, and differentiation of multiple pituitary cell types (Bancalari et al., 2012). These structural and cellular developmental processes are orchestrated by complex molecular mechanisms involving the interplay of multiple signaling pathways and transcription factors such as *HESX1, SOX2/3, TBX2/3, LHX2/3/4, SIX3/6, PITX1/PITX2, PROP1, POU1F1*, and *TPIT* (Rizzoti, 2015). Abnormalities in this developmental program can result in congenital hypopituitarism, a disorder characterized by presence of one or more pituitary hormone deficiencies. Often these functional pituitary deficits are accompanied by structural pituitary abnormalities, such as a hypoplastic or absent anterior pituitary gland and/or an ectopic posterior pituitary gland.

Although much has been learned about pituitary development from mouse models, the etiology can only be identified in a minority of patients with congenital hypopituitarism. In some affected patients, monogenic etiologies have been identified, including genetic defects that are often recessive in *HESX1* (Thomas et al., 2001), *PROP1* (Wu et al., 1998; Correa et al., 2019), *POU1F1* (Turton et al., 2005; Birla et al., 2019), *LHX3* (Rajab et al., 2008; Jullien et al., 2019), and *LHX4* (Tajima et al., 2007). However, a monogenic etiology can be identified in fewer than 10% of cases (Castinetti et al., 2015); in the remainder, the etiology is generally unclear. These idiopathic congenital cases are usually sporadic; familial cases of isolated congenital hypopituitarism are less commonly observed (Zwaveling-Soonawala et al., 2018), suggesting that single-gene defects may not be the dominant cause. In some patients, pathogenic variants are suspected but are also found in unaffected relatives, suggesting incomplete penetrance due to additional unknown contributing factors (Bashamboo et al., 2016, 2017; Cohen et al., 2017; Hovinga et al., 2018; Babu et al., 2019; Dateki et al., 2019). These observations suggest that the etiology of congenital hypopituitarism is often more complex than a simple monogenic defect. A possible explanation is that, as has been described for idiopathic hypogonadotropic hypogonadism congenital hypopituitarism may often have a digenic or oligogenic etiology. Indeed, cases of digenic inheritance have been reported (McCormack et al., 2017). A second possibility is that there are major non-genetic causative factors. For example, an association between congenital hypopituitarism and adverse perinatal events with breech presentation has been observed (Maghnie et al., 1991), suggesting an etiological role for birth asphyxia (Kikuchi et al., 1988).

In this study, we asked two questions regarding the etiology of congenital hypopituitarism. First, in patients with sporadic congenital hypopituitarism, is there a major genetic component to the etiology? Second, in these patients, are monogenic causes common, involving as yet undiscovered genes? To address the first question—whether there is a major genetic component underlying sporadic congenital hypopituitarism—we studied a group of patients with non-familial congenital hypopituitarism and used exome sequencing to determine the frequency of rare, predicted-pathogenic variants in 42 genes that have been implicated in pituitary gland development (Zwaveling-Soonawala et al., 2018). As a control group, we used patients with non-familial short stature (NFSS) of unknown etiology who had been studied in parallel with the hypopituitary subjects, using the same methods for exome sequencing, data processing, and sequence analysis. These control subjects had been evaluated to exclude pituitary disease and thus would be expected to have a frequency of pituitary-related variants similar to that of the normal population.

To address the second question—whether sporadic congenital hypopituitarism often has a monogenic cause, involving as yet undiscovered genes—we studied the same patient population. However, in this second analysis, we searched for variants in any gene (not just those known to be involved in pituitary development) that was inherited in a pattern that would explain the apparent sporadic occurrence. Thus, we counted only rare, predicted-pathogenic variants that showed a recessive inheritance (homozygous or compound heterozygous in the proband with one variant inherited from each parent), an X-linked recessive inheritance (male probands only), or arose *de novo* (not present in either parent). We compared the frequency of such candidate variants in subjects with congenital hypopituitarism to the frequency in subjects with NFSS. In this analysis, NFSS subjects allowed us to compare the relative frequency of potential monogenic causes in the two conditions.

MATERIALS AND METHODS

Subjects and Sample Collection

Thirteen patients with congenital hypopituitarism (age 4–31 years, 5 males) and their unaffected biological parents were studied. Based on MRI evaluation, 10 probands had hypoplastic or absent anterior pituitary glands with ectopic posterior

pituitaries and 3 probands had hypoplastic or absent anterior pituitary gland and absence of the bright spot that corresponds to the posterior pituitary gland. Affected subjects had either isolated GH deficiency (n = 1) or combined with other pituitary hormone deficiencies (n = 12; **Table 1**). Growth hormone deficiency was diagnosed with either provocative testing using 2 different stimuli or low IGF-1 in the context of other pituitary hormone deficiencies and pituitary abnormalities on MRI. No affected subjects had a history of consanguinity or a family history of pituitary disease. No patients had a history of adverse perinatal events including birth asphyxia or breech delivery. TSH deficiency was diagnosed based on a low serum free T4 with serum TSH < 10 mIU. ACTH deficiency was diagnosed with ACTH stimulation based on a peak serum cortisol < 18 μg/dL after receiving 250 micrograms of $ACTH_{1-24}$ intravenously. FSH/LH deficiency was diagnosed when subjects failed to enter puberty. One subject had diabetes insipidus.

Nineteen subjects with NFSS and their unaffected biological parents were also studied. None of these subjects (age 3–42 years, male = 12) had a history of pituitary hormone deficiencies. On clinical and biochemical evaluation, they had no evidence of central hypothyroidism. Similarly, all NFSS subjects either passed growth hormone provocative testing or showed no evidence of growth hormone deficiency to justify provocative testing. There was no history of consanguinity. All affected subjects' parents had normal heights suggesting monogenic inheritance (**Table 2**). At the time of presentation, the affected subjects had no clinical findings revealing a commonly known genetic cause of diminished linear growth evident either to their health care providers or to the study investigators. The characteristics of these subjects are provided in detail in **Table 2**. Nine patients had isolated short stature whereas 10 patients had other accompanying abnormal features.

Genomic DNA from white blood cell in all patients was extracted for genetic testing. The study (clinicaltrials.gov/ct2/show/NCT02311322) was approved by the NICHD IRB. All adult subjects and parents of minors provided written informed consent and children provided written assent.

TABLE 1 | Characteristics of subjects with congenital hypopituitarism.

Age†	Pituitary hormone deficiencies	Anterior pituitary gland on MRI	Posterior pituitary on MRI
10–15	GH, partial ACTH, TSH	HAP	EP
16–20	GH, ACTH, TSH, LH/FSH	HAP	EP
16–20	GH, ACTH, TSH, LH/FSH	AAP	EP
6–10	GH	HAP	EP
20–25	GH, ACTH, TSH, LH/FSH	AAP	EP
15–20	GH, ACTH, TSH, LH/FSH	HAP	X
20–25	GH, ACTH, TSH, LH/FSH	HAP	EP
16–20	GH, ACTH, TSH, LH/FSH	AAP	X
30–31	GH, ACTH, TSH, LH/FSH	HAP	EP
0–5	GH, ACTH, TSH	HAP	EP
15–20	GH, ACTH, TSH, LH/FSH	AAP	EP
6–10	GH, ACTH, TSH, LH/FSH	HAP	X
0–5	GH, ACTH, TSH	HAP	EP

HAP, hypoplastic anterior pituitary gland; AAP, aplastic anterior pituitary gland; EP, ectopic posterior pituitary gland; X, Not visualized on pituitary MRI. †Age (y) range was provided per journal policy and sex of individual subjects was omitted to protect privacy per journal policy.

TABLE 2 | Characteristics of subjects with non-familial short stature.

Age†	Height SDS at diagnosis	PH/MPH¶ (SDS)	Associated dysmorphic features
40–45	**−2.1**	−0.5/−0.1/**−0.3**	Disproportionate short stature, Noonan-like facies
6–10	**−2.1**	−1.1/0.1/**−0.6**	None
6–10	**−4.63**	−0.3/0.5/**0.1**	Multiple dysmorphic features
6–10	**−2.9**	0.1/−0.7/**−0.2**	None
35–40	**−2**	−1.2/0.4/**−0.4**	None
5–10	**−2.4**	0.5/−0.1/**0.7**	Speech delay
35–40	**−1.9**	−0.2/0.1/**0.1**	None
15–20	**−2.1**	1.9/0.4/**1.1**	Disproportionate short stature
5–10	**−3.8**	0.5//1.1/**0.7**	Joint laxity, muscular build
5–10	**−2.6**	0.5/−0.5/**0.00**	Disproportionate short stature, facial dysmorphism, Chiari I malformation, developmental delay, Shawl scrotum, developmental delay
10–15	**−1.9** (on GH)	0.1/−1.4/**−0.6**	Early puberty
5–10	**−2.7**	−0.8/−1.1/**−0.9**	None
0–5	**−4.6**	0.5/−0.1/**0.3**	None
6–10	**−1.75**	−0.76/0.7/**−0.1**	Midface hypoplasia Disproportionate short stature
6	**−2.6**	−0.2/0.7/**0.1**	Dysmorphic facial features
8	**−2**	−0.6/−0.3/**−0.5**	None
8	**−2**	−0.6/0.3/**−0.1**	None
8	**−2.2**	0.9/−0.3/**0.3**	None
3	**−5.07**	−2/−2.1/**−2**	Developmental delay, dysmorphic facial features

¶PH/MPH: Father's height/mother's height/Mid-parental height. †Age (y) range was provided per journal policy and sex of individual subjects was omitted to protect privacy per journal policy.

SNP Array and Exome Sequencing and Data Analysis

SNP Array and Exome Sequencing

SNP array was performed as previously described (Jee et al., 2017). SNP array data were analyzed for copy-number variations (CNVs) previously reported to be associated with congenital hypopituitarism (Correa et al., 2018), as well as to confirm paternity and rule out parental consanguinity.

Exome sequencing was performed on probands and the parents as trios at the National Institutes of Health Sequencing Center. The detailed method was previously published (Jee et al., 2017). In brief, Illumina sequencing libraries were generated from 100 ng genomic DNA using the Accel-NGS 2S DNA Library Kit (Swift Biosciences) on a Beckman Coulter Biomek FX robot. The median insert size was approximately 350 bp. Libraries were tagged with unique dual index DNA barcodes to allow pooling of libraries and minimize the impact of barcode hopping. Libraries were pooled in groups of 8 for exome enrichment using the xGen Exome Research Panel v1.0 (IDT). Multiple enriched pools

were combined for sequencing on the NovaSeq 6000 (Illumina) to obtain at least 35 million 150-base read pairs per individual library. Raw data were processed using RTA version 3.4.4 for base calling. All subjects sequencing data were processed in parallel for alignment and variant calling. The combined data were formatted in a single .vs file file and analyzed using VarSifter (Teer et al., 2012) which enables a search for variants in a specific gene and analysis of genotypes using Boolean logic.

Data Analysis 1: Frequency of Candidate Variants in Genes Associated With Pituitary Development

We first identified variants in 42 genes (**Supplementary Table 1**) previously reported to be associated with pituitary development, including 40 genes compiled by Zwaveling-Soonawala et al. (2018) and 2 additional genes that were found associated with pituitary hormone deficiencies in humans. These 42 genes include: (1) genes previously reported in patients with congenital hypopituitarism, (2) genes for which a genetic mouse model showed pituitary abnormalities, or (3) genes encoding proteins that could interact with signaling pathways important for pituitary gland development. Variants were considered potentially pathogenic candidates if they met all of the following criteria: (1) sequencing coverage > 10 reads and most probable genotype (MPG)/coverage > 0.5 (Adams et al., 2012), (2) confirmed to be present by visual inspection of Binary Alignment Map (BAM) files, (3) population frequency < 1% in the Genome Aggregation Database (gnomAD) v2.1.1 data set (GRCh37hg19), and homozygous variants found in < 2 subjects in gnomAD (which includes approximately 150,000 subjects), and (4) altered the predicted amino acid sequence of the encoded protein (i.e., missense, non-sense, frameshift/non-frameshift insertions or deletions, and splicing variants), and (5) predicted to be pathogenic by at least 2 out of 3 prediction algorithms (SIFT, MutationTaster, PolyPhen2) (Adzhubei et al., 2010; Schwarz et al., 2014; Vaser et al., 2016). CADD score was not included as one of the prediction criteria but is provided as **Supplementary Information**. For this analysis, variants that met the above criteria were sought in probands regardless of inheritance pattern in order to broadly screen for any genetic contribution, including incomplete penetrance and di/oligogenic causes (**Supplementary Figure 1**). This analysis was performed separately in subjects with congenital hypopituitarism and in subjects with NFSS, and the frequency of rare, predicted-pathogenic variants was compared in the two groups. The latter group, which had no evidence of pituitary disease, served as a control group for this burden testing approach.

Data Analysis 2: Frequency of Candidate Variants for a Monogenic Etiology

We next sought genetic variants in any protein-coding gene in the genome, not just genes associated with pituitary development. We used the same 5 criteria used in Data Analysis 1 to identify rare, predicted-pathogenic variants. We next narrowed the list to include only variants that were inherited in a pattern consistent with a monogenic etiology in these families that had an affected proband and two unaffected parents. These patterns included: (1) an autosomal recessive inheritance in which the proband had a homozygous or compound heterozygous variant in a gene with one allele inherited from each parent; (2) X-linked recessive inheritance in which a male proband inherited the variant from his mother; or (3) *de novo* occurrence in which a heterozygous variant was found in the proband but was absent in both parents (**Supplementary Figure 1**). This analysis was performed separately in subjects with congenital hypopituitarism and in subjects with NFSS, and the frequency of rare, predicted-pathogenic variants was compared in the two groups to compare the frequency of potential monogenic inheritance in the two conditions.

Statistical Analysis

Statistical analyses were performed using SPSS (v25, IBM, Armonk, NY). The average number of variants per proband was compared in the two groups (congenital hypopituitarism vs. NFSS) using the Mann-Whitney *U*-test. The percent of probands with at least 1 variant was compared in the two groups with the Fisher exact test.

RESULTS

SNP Array Analyses

SNP array analyses confirmed paternity and non-consanguinity in all families and excluded the presence of significant copy-number variations.

Rare, Predicted-Pathogenic Variants in Genes Associated With Pituitary Gland Development Are Enriched in Subjects With Congenital Hypopituitarism

In data analysis 1, using exome sequencing, we searched for rare, predicted-pathogenic sequence variants in 42 genes associated with pituitary gland development. The average number of these variants per proband was greater in subjects with congenital hypopituitarism (n = 13 probands) than in in the control subjects (n = 19 probands) with NFSS (1.1 vs. 0.21 mean variants per proband, P = 0.03, **Figure 1A**). Similarly, the percent of probands with at least 1 variant in any of the 42 pituitary-associated gene was greater in subjects with congenital hypopituitarism than in subjects with NFSS (62% vs. 21%, P = 0.03, **Figure 1B**). Rare, predicted-pathogenic variants were found in 11 genes (*ARID1B, CDON, CHD7, GLI1, GLI4, LHX3, LHX4, SIX1, SIX5, SIX6, SOX3*) in 8 subjects with congenital hypopituitarism (62%) and in 3 genes (*CHD7, LHX4*, and *WNT5A*) in 4 subjects with NFSS (21%) (**Table 3** and **Supplementary Table 1**). Two subjects with congenital hypopituitarism carried more than one rare, predicted-pathogenic variant: one subject with congenital hypopituitarism carried variants in *LHX4* and *CDON* inherited from her father and a variant in *SIX5* from her mother; the other subject with congenital hypopituitarism carried an in-frame variant in the polyalanine region of *SOX3* from her father and variants in *GLI1* and *CHD7* (Gregory, 2020) from her mother. Hughes et al. (2013) showed that changes in the polyalanine expansion could cause a partial loss-of-function in

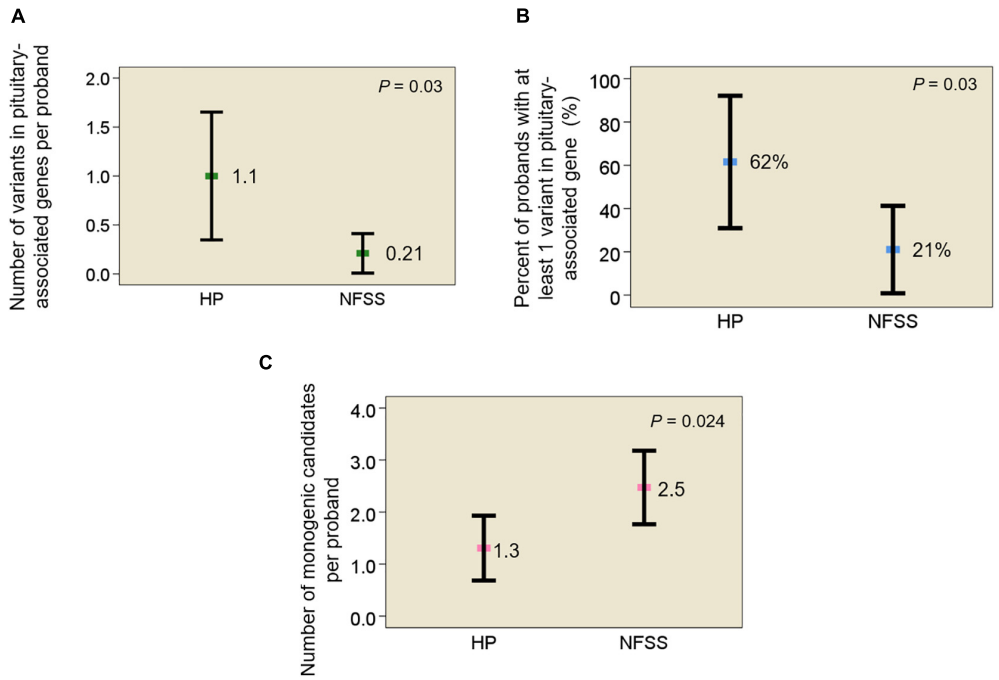


FIGURE 1 | Rare, predicted-pathogenic variants in subjects with congenital hypopituitarism and non-familial short stature. **(A)** Mean number of rare, predicted-pathogenic variants per proband in 42 genes associated with pituitary development. **(B)** Percent of probands with at least 1 variant in any of 42 pituitary-associated genes. **(C)** Mean number of monogenic candidates for the condition (hypopituitarism or non-familial short stature) per proband. Square indicates mean. Error bar shows 95% confidence interval. HP, congenital hypopituitarism; NFSS, non-familial short stature.

the protein and Alatzoglou et al. (2011) reported a patient with hypopituitarism due to deletion in the polyalanine tract (Alatzoglou et al., 2011; Hughes et al., 2013). However, because the subject's father, who had no evidence of hypopituitarism, also carried it, this variant alone does not appear sufficient on its own to cause hypopituitarism; however, it remains possible that this variant could contribute to abnormal development of the pituitary gland in concert with other genetic variants. None of the variants found in subjects with congenital hypopituitarism were inherited in a pattern that could explain the sporadic presentation with a fully penetrant autosomal recessive, X-linked recessive, or *de novo* inheritance. Instead, we observed primarily heterozygous variants inherited from an unaffected parent, suggesting that these variants, if pathogenic, show incomplete penetrance, possibly due to a digenic or oligogenic etiology or to interacting non-genetic factors, as has been proposed (Hong and Krauss, 2012). There was no specific phenotype correlated with the identified variants in these patients.

Candidate Variants That Could Account for a Fully Penetrant Monogenic Etiology Were Infrequently Found in Subjects With Congenital Hypopituitarism

In data analysis 2, we searched the exome sequencing data for variants in any gene (not just those known to be involved in pituitary development) in order to capture undiscovered genetic causes that were inherited in a fashion that could explain the sporadic occurrence of the proband's condition (either hypopituitarism or short stature) with a monogenic, fully penetrant etiology. Thus, we counted only rare, predicted-pathogenic variants that showed a potentially recessive inheritance (homozygous or compound heterozygous in the proband with one allele inherited from each parent), an X-linked recessive inheritance (male probands only), or arose *de novo* (not present in either parent) because these inheritance patterns could explain the presence of the condition in the proband and the absence in the parents. Interestingly, there were fewer of these monogenic candidates in subjects with congenital hypopituitarism than in the subjects with NFSS (1.3 vs. 2.5 candidates per proband, P = 0.024, **Figure 1C** and **Table 4**). Furthermore, none of these variants (**Table 4**) were found in genes that have been reported to cause pituitary hormone deficiencies or anatomic brain abnormalities and indeed variants in many of these genes are associated with disorders that do not involve the pituitary (**Table 4**), suggesting that most of these variants are not responsible for the hypopituitarism. In contrast, in subjects with NFSS, we found variants in multiple genes that have been found to cause monogenic short stature, including *BRF1* (Cerebellofaciodental Syndrome) (Jee et al., 2017), *QRICH1* (Ververi-Brady Syndrome) (Lui et al., 2019), *FBN1* (acromelic dysplasia), *HUWE1* (X-linked mental retardation syndrome), *SRCAP* (Floating-Harbor syndrome), *ACAN* (short stature and advanced bone age) (Tatsi et al., 2017), *ZEB* (Mowat-Wilson Syndrome), *and CUL7* (3-M syndrome). The causality of the variants found in these genes was supported

TABLE 3 | Genetic variants found in 42 known pituitary-associated genes.

Gene (variant, hg19)	Variant type	Inheritance from	Population frequency in gnomAD	Evidence for a role in pituitary gland
Congenital hypopituitarism				
CDON (chr11: 125887183 C > T)	Missense	Mother	Not found	Reported to be associated with pituitary stalk interruption (Bashamboo et al., 2016)
ARID1B (chr6: 157454234 C > T)	Missense	Mother	0.004	Reported to be associated with growth hormone deficiency in human and mice (Yu et al., 2015; Celen et al., 2017)
SIX6 (chr14: 60976501 G > A)	Missense	Mother	0.007	Expressed in pituitary gland during development and interacts with PITX1 and PITX2 (Jean et al., 1999; Xie et al., 2015)
SIX1 (chr14: 61113110 G > A)	Missense	Mother	0.003	Reported to induce pituitary placode (Zimmer et al., 2016)
LHX4 (chr1: 180235637 G > A)	Missense	Father	Not found	Reported in patients with CH (Cohen et al., 2017)
CDON (chr11: 125859642 C > T)	Non-sense	Father	0.00006	Reported to be associated with pituitary stalk interruption (Bashamboo et al., 2016)
SIX5 (chr19: 46268833 C > T)	Missense	Mother	0.0002	Member of SIX family and high expression in mouse and human pituitary gland (http://biogps.org/#goto)
ARID1B (chr6: 157522148 C > A)	Missense	Mother	0.00001	Reported to be associated with growth hormone deficiency in human and mice (Yu et al., 2015; Celen et al., 2017)
GLI1 (chr12: 57865830-AA-A)	frameshift	Mother	0.000006	Involved in pituitary gland development in zebrafish (Wang et al., 2010)
SOX3 (chrX: 139586486-CAGCGGCGGCGGCCG CGGCAGC-C)	Non-frameshift substitution	Father	Not found	Reported in patients with CH (Alatzoglou et al., 2011)
CHD7 (chr8:61693942_61693947 insGCAAAA)	Non-frameshift Insertion	Mother	Not found	Reported in patients with CHARGE syndrome and CH (Gregory et al., 2013)
GLI4 (chr8: 144358408 G > C)	Missense	Father	0.0001	Mediates Sonic Hedgehog signaling (Villavicencio et al., 2000)
LHX3 (chr9: 139091601 G > T)	Missense	Father	0.0002	Reported in patients with CH (Jullien et al., 2019)
Non-familial short stature				
CHD7 (chr8: 61774876 A > G)	Missense	Father	0.00001	Reported in patients with CHARGE syndrome and CH (Gregory et al., 2013)
LHX4 (chr1: 180243593 C > T)	Missense	Mother	0.0002	Reported in patients with CH (Cohen et al., 2017)
LHX4 (chr1: 180217489 A > G)	Missense	Mother	0.0001	Reported in patients with CH (Cohen et al., 2017)
WNT5A (chr3: 55508427 T > C)	Missense	Mother	0.00001	Involved in mouse pituitary gland development (Cha et al., 2004)

CH, congenital hypopituitarism.

by bioinformatic analysis and either a highly specific clinical phenotype consistent with the known disorder or a consistent but less specific phenotype combined with functional studies. Thus, candidate variants in genes previously reported to explain the phenotype were identified in 0 of the 13 trios with congenital hypopituitarism and in 8 of 19 trios (42%) with NFSS (P = 0.01; **Table 2**). These findings suggest that a monogenic inheritance is substantially less common in sporadic congenital hypopituitarism than in sporadic short stature.

DISCUSSION

The first major finding of this study is that rare, predicted-pathogenic variants in genes associated with pituitary development were enriched in subjects with congenital hypopituitarism compared to a control group of subjects with NFSS. In the 13 subjects with congenital hypopituitarism in our study, rare variants that are predicted to be pathogenic were found in 10 genes implicated in pituitary development including 8 genes (*ARID1B, CDON, CHD7, LHX3, LHX4, PAX6, SIX6*, and *SOX3*) reported to be associated with pituitary hormone deficiencies in humans. Some of these variants may have been found incidentally and not be causative. However, only 0.26 such variants per proband were found in the control group, which suggests that only a few of the variants found in the subjects with hypopituitarism were incidental. Thus, the findings from this first burden analysis suggest that the etiology of sporadic congenital hypopituitarism has a major genetic component.

The second major finding from this study is that the number of monogenic candidate variants was lower in subjects with congenital hypopituitarism than in those with NFSS. For this analysis, we searched for rare, predicted-pathogenic variants that were inherited in a fashion (autosomal recessive, X-linked recessive, or *de novo* occurrence) that could explain the condition (either congenital hypopituitarism or NFSS) as a monogenic, fully penetrant disorder. This analysis was not restricted to genes involved in pituitary development but potentially included all protein-coding genes in the genome in order to capture undiscovered genetic causes. In the subjects with NFSS, we found an average of 2.5 monogenic candidates per proband including multiple genes previously reported to cause monogenic short stature, suggesting that NFSS frequently has a monogenic etiology. In contrast, in subjects with congenital hypopituitarism, we found fewer monogenic candidates, 1.3 per proband, and none of these had previously been shown to cause hypopituitarism or play an important role in pituitary or brain development. Taken together, the two analyses suggest that congenital hypopituitarism has a major

TABLE 4 | Candidate monogenic variants.

Congenital hypopituitarism			
Inheritance mode	**Gene**	**Known associated disease**	**Gene function/tissue expression/mouse model**
De novo AR (compound het) XLR	*TTLL4* *TNXB* .	. Ehlers-Danlos syndrome (No CE) .	*TTLL4*: Required for cytoskeletal organization (Ijaz et al., 2017)/NS/NA *TNXB*: extracellular matrix glycoproteins/NS/abnormal skin tensile strength
De novo AR XLR	*KLHDC4* . .	. . .	*KLHDC4*: Orphan member of the kelch repeat superfamily, possibly involved in cell proliferation and migration (Lian et al., 2016)/NS/NA
De novo AR (compound het) XLR	*MARCH3* *PER2* *FLG* .	. Familial advanced sleep phase syndrome (No CE) Ichthyosis Vulgaris (No CE)	*MARCH3*: E3 ubiquitin ligase (Fatehchand et al., 2016)/NS/available but pituitary phenotype was not investigated (Lin et al., 2018) *PER2*: circadian pace maker/high expression in rat pituitary gland/disrupted circardian rhythm and cancer prone (Fu et al., 2002) *FLG*: intermediate filament-associated protein/NS/percutaneous allergen priming (Fallon et al., 2009)
De novo AR (compound het) XLR	*CRHR1* . *GPC3*	. . Type 1 Simpson-Golabi-Behmel syndrome (No CE)	*CRHR1*: CRH receptor/high expression in human pituitary gland in human/*Crh1*(-/-) mice show upregulated Acth receptor in pituitary and adrenal gland (Müller et al., 2001) *GPC3*: control of cell division and growth regulation/NS/controls limb patterning and skeletal development (Paine-Saunders et al., 2000)
De novo AR (homozygous) *DCAKD*	. *TTLL6* .	. . .	. *TTLL6*: polyglutamylase enzyme and regulates cilia structure and motility (Lee et al., 2012)/NS/NA *DCAKD*: Parkinson disease high risk loci (Barbu et al., 2020)/NS/NA
De novo AR	. .	. .	. .
De novo AR	. .	. .	. .
De novo Compound het	. *PDE4DIP*	. .	. *PDE4DIP*: phosphohydrolases, involved in signal transduction and hydrolyze 3' cyclic phosphate bonds in 3', 5'-cGMP and 3',5'-cAMP (Shapshak, 2012)/NS/NA
De novo AR (compound het)	. *CEP128*	. .	. *CEP128*: regulates TGF-β/BMP singling at the primary cilium (Mönnich et al., 2018)/NS/NA
De novo AR (homozygous)	. *MET*	. Autosomal recessive deafness (No CE)	. *MET*: receptor tyrosine kinase/NS/tumorigenesis (Graveel et al., 2009)
De novo AR (homozygous) X-linked	. *PRRG2* .	. .	. *PRRG2*: serves as binding partner for multiple proteins (Yazicioglu et al., 2013)/NS/NA
De novo AR (homozygous) AR (compound het)	*GJB5* *ACADVL* *RGPD3*	. Very long-chain acyl CoA dehydrogenase deficiency (No CE) .	*GJB5*: involved in trophoblast stem cell differentiation (Kibschull et al., 2014)/NS/available but pituitary phenotype was not investigated (Zheng-Fischhöfer et al., 2007) *ACADVL*: Catalytic enzyme/NS/hepatic steatosis (Cox et al., 2001) *RGPD3*: GTPase activator, reported to be associated with craniofacial morphology (Wu et al., 2019)/high expression in human pituitary gland/NA
De novo AR	. .	. .	. .
Non-familial short stature			
Inheritance	**Gene**	**Known associated disease or role in linear growth**	
De novo AR (compound het)	**Potential new genetic cause** *SMARCD2* *TMBIM1*	**Likely role in linear growth** Autosomal recessive neutropenia (No CE) No known association with linear growth	
De novo AR (homozygous) (compound het)	. *APLP2* *ZFHX3*	. No known association with linear growth No known association with linear growth	
De novo AR (compound het) XLR	. ***BRF1* (causative variant)** *MCM3AP* .	. **Cerebellofaciodental syndrome (CE)** Autosomal recessive peripheral neuropathy (No CE) .	

(Continued)

TABLE 4 | Continued

Non-familial short stature		
Inheritance	**Gene**	**Known associated disease or role in linear growth**
De novo	.	.
AR (compound het)	*SDSL*	Congenital neutropenia (no CE)
De novo	**Potential new genetic cause**	**Likely role in linear growth**
AR (compound het)	*FZD6*	Autosomal receive Nail disorder (no CE)
XLR	*HECW2*	Neurodevelopmental disorder (no CE)
	WDR52	Autosomal recessive spermatic failure (no CE)
	.	.
De novo	***QRICH1* (causative variant)**	**Ververi-Brady syndrome (CE)**
AR	*PQLC2*	No known association with linear growth
De novo	*FBXO41*	No known association with linear growth
AR (compound het)	*PEX6*	Zellweger syndrome (no CE)
De novo	*PDCD6IP*	No known association with linear growth
AR (compound het)	**Potential new genetic cause**	**Likely role in linear growth**
XLR	*SRRM2*	No known association with linear growth
	.	.
De novo	***FBN1* (causative variant)**	**Acromelic dysplasia (CE)**
AR	.	.
XLR	.	.
De novo	*ZNF506*	No known association with linear growth
AR	.	.
XLR	***HUWE1* (causative variant)**	**X-linked syndromic mental retardation, Turner type (CE)**
De novo	***SRCAP* (causative variant)**	**Floating-Harbor syndrome (CE)**
AR	.	.
XLR	.	.
De novo	.	.
AR	*MMP8*	No known association with linear growth
XLR	*SPTB*	Type 2 spherocytosis (No CE)
De novo	*DDX46*	No known association with linear growth
AR (homozygous)	*MRGPRX1*	No known association with linear growth
AR (compound het)	*FTCD*	Autosomal recessive glutamate formiminotransferase deficiency (no CE)
	MCOLN3	No known association with linear growth
	RYR1	Neuromuscular disease (no CE)
	MLH1	Autosomal recessive Mismatch mismatch repair cancer syndrome (no CE)
De novo	***ACAN* (causative variant)**	**Short stature with advanced bone age (CE)**
AR	.	.
XLR	*BCOR*	Syndromic microphthalmia (no CE)
	NAP1L2	No known association with linear growth
	ACTRT1	No known association with linear growth
	MAGEC2	No known association with linear growth
De novo	***ZEB2* (causative variant)**	**Mowat-Wilson syndrome (CE)**
AR	.	.
XLR	*COL4A5*	X-linked Alport syndrome (no CE)
De novo	.	.
AR	.	.
XLR	.	.
De novo	*CYFIP2*	Epileptic encephalopathy (no CE)
AR (compound het)	*IL7R*	Severe combined immunodeficiency (no CE)
	NRAP	No known association with linear growth
De novo	**Potential new genetic cause**	**Likely role in linear growth**
AR	*LRP1B*	No known association with linear growth
XLR	.	.
De novo	.	.
AR (homozygous)	*PLBD1*	No known association with linear growth
	HEG1	No known association with linear growth
AR (compound het)	***CUL7* (causative variant)**	**3-M syndrome (CE)**
XLR	*APOOL*	No known association with linear growth

AR, autosomal recessive; compound het, compound heterozygous; XLR, X-linked recessive; No CE, the subject had no clinical evidence for the disorder listed; CE, the subject had clinical evidence for the disorder listed. Causal genes and strong candidate genes are in bold; NS, non-specific expression (reference: http://biogps.org/#goto); NA, not available. Mouse phenotype was searched at http://www.informatics.jax.org.

genetic component but may have a more complex etiology than a simple monogenic disorder with complete penetrance. However, a large-scale study would be needed to confirm our findings. One possibility consistent with these observations is that congenital hypopituitarism often has a digenic or oligogenic inheritance, as has been suggested (Hamdi-Rozé et al., 2020), although data for this hypothesis has been reported only in a small number of families to date (Zwaveling-Soonawala et al., 2018). In our study, 2 out of 13 families showed a possible oligogenic etiology. Another possibility is that congenital hypopituitarism is often caused by a combination of an underlying genetic defect and a non-genetic insult. This possibility has been demonstrated in mice in which a combination of *Cdon* genetic ablation and prenatal exposure to ethanol impaired development of Rathke's pouch (Hong and Krauss, 2012). A non-sense mutation in *CDON* (Bashamboo et al., 2016) was previously reported in a patient with pituitary stalk interruption syndrome, however the variant was also inherited from the unaffected parent, as we observed in one of our patients, suggesting the presence of other factor(s) contributing to the pathogenesis of the disease. Congenital hypopituitarism also has been associated with breech position and adverse perinatal events (Maghnie et al., 1991) suggesting a role for non-genetic factors; however, we did not observe the association in our cohort.

A strength of the study design is that it avoids several possible confounding factors that could affect the number of variants found in the two groups of subjects. Subjects in both groups were sequenced at a single sequencing center using the same methodology, and their data were aligned, genotyped, and variant-called in parallel. In addition, we sought to avoid false positive and negative variant calls by manually examining the BAM files for all observed rare variants with coverage > 10 sequencing reads and MPG/coverage ratio of > 0.5. However, this study did not have a control population of healthy subjects and parents who had undergone exome sequencing analysis using all the same methodologies as the subjects with hypopituitarism. An additional limitation is the small size of the study. Far more subjects would have been required to perform burden testing for individual genes, that is, to show an increased number of variants in one specific gene in the case-vs.-control analysis. For this reason, we instead performed burden testing on aggregate sets of genes—either genes associated with pituitary development or genes that were inherited in a pattern that matched the pedigree—thus greatly increasing the power to detect differences between the case and control groups. We chose not to compare variant frequencies with those in available public databases which would have introduced multiple confounding variables (Guo et al., 2018). These databases combine exome sequence data obtained using various sequencing platforms, and thus would not match the sequencing platform used for our subjects with hypopituitarism. Similarly, the sequence data would not have been jointly processed and variant-called with the samples from our hypopituitary samples, introducing additional potential systematic biases (Guo et al., 2018). In addition, coverage for each gene is not known, especially for the 42 genes that we examined in this study. Consequently, if these databases had been used, any observed differences in variant burden might be due to methodological differences between the case and control groups, rather than to real genetic differences. Furthermore, examining BAM files of available databases as a quality check on variant calls is not feasible with public databases. Therefore, instead, we compared data from patients with congenital hypopituitarism to a group of subjects with NFSS that had been evaluated for short stature and were not found to have evidence of pituitary disease and thus would be expected to have a frequency of pituitary-associated variants similar to the general population. Moreover, NFSS is often a monogenic disorder with full penetrance and thus serves as a useful comparator to congenital hypopituitarism for the analysis of monogenic candidates. A study design limitation is that exome sequencing data can identify only coding region variants or intronic variants that affect splicing, and thus misses non-coding regulatory variants, such as those in promoter regions, enhancers, or microRNAs. Further study exploring these regulatory regions may reveal additional important genetic cases of congenital hypopituitarism. An additional limitation is that the number of trios examined in our study may not be sufficient to represent the full spectrum of congenital hypopituitarism, which is likely to have a heterogenous etiology.

Our study findings are consistent with the findings of Zwaveling-Soonawala et al. (2018) who studied 20 patients with isolated pituitary stalk interruption syndrome and found multiple variants in genes associated with pituitary development or CNS development, which appeared to contribute to a polygenic etiology. McCormack et al. (2017) also reported the digenic inheritance of pathogenic variants in *PROKR2* and *WDR11* in a child with pituitary stalk interruption syndrome supporting the concept of a digenic cause for congenital hypopituitarism, and abnormal pituitary development. However, our study is novel in that it includes a control group that was sequenced and analyzed in parallel to allow rigorous comparisons of variant frequencies.

CONCLUSION

In conclusion, we report that rare predicted-pathogenic variants in genes known to be associated with pituitary development are enriched in patients with sporadic congenital hypopituitarism, suggesting that the etiology of congenital hypopituitarism has an important genetic component. However, none of the variants found was inherited in a pattern that could explain the sporadic disorder with a fully penetrant monogenic etiology. Even when we widened our search beyond genes known to be associated with pituitary development, to capture undiscovered genes, we found few variants consistent with a fully penetrant monogenic etiology compared to subjects with NFSS. Thus, taken together, the findings suggest that genetic variants affecting pituitary development play a major role in the etiology of congenital hypopituitarism but that the disorder may not be a simple single-gene defect, at least in protein-coding regions.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found online in dbGaP using accession number phs001617.v1.p1.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the National Institute of Child Health and Development. Written informed consent to participate in this study was provided by participants and the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

YJ, JB, and SR contributed to the conception and design of the study. YJ organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. All authors contributed to evaluating subjects, manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the NICHD grant U01HD086838 and the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development grants ZIA-HD000640 and ZIA-HD00641.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.697549/full#supplementary-material

Supplementary Figure 1 | Diagram of study design. **(A)** Data analysis 1: rare, predicted-pathogenic variants were sought in 42 genes associated with pituitary gland development. **(B)** Data analysis 2: rare, predicted-pathogenic variants were sought in any gene in the genome (to capture undiscovered causes) that was inherited in a fashion that could explain the sporadic occurrence of the proband's condition with a monogenic etiology (*de novo* mutation, autosomal recessive, or X-linked recessive). HP, congenital hypopituitarism; NFSS, non-familial short stature.

Supplementary Table 1 | Genetic variants found in 42 known pituitary-associated genes: ARID1B, ARNT2, BMP2, BMP4, CHD7, CDON, FGF8, FGF10, FGF18, FGFR1, GATA2, GLI1, GLI2, GLI3, GLI4, GLI5, GLI6, GPR161, HESX1, IGSF1, LHX3, LHX4, NR5A1, OTX2, PAX6, PITX1, PITX2, POU1F1, PROP1, SHH, SIX1, SIX2, SIX3, SIX4, SIX5, SIX6, SOX1, SOX2, SOX3, TBX19, TGIF, WNT5A.

REFERENCES

Adams, D. R., Sincan, M., Fajardo, K. F., Mullikin, J. C., Pierson, T. M., Toro, C., et al. (2012). Analysis of DNA sequence variants detected by high-throughput sequencing. *Hum. Mutat.* 33, 599–608. doi: 10.1002/humu.22035

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Alatzoglou, K., Kelberman, D., Cowell, C. T., Palmer, R., Arnhold, I. J. P., Melo, M. E., et al. (2011). Increased transactivation associated with SOX3 polyalanine tract deletion in a patient with hypopituitarism. *J. Clin. Endocrinol. Metab.* 96, E685–E690.

Babu, D., Fanelli, A., Mellone, S., Muniswamy, R., Wasniewska, M., Prodam, F., et al. (2019). Novel GLI2 mutations identified in patients with combined pituitary hormone deficiency (CPHD): evidence for a pathogenic effect by functional characterization. *Clin. Endocrinol. (Oxf)* 90, 449–456.

Bancalari, R. E., Gregory, L. C., McCabe, M. J., and Dattani, M. T. (2012). Pituitary gland development: an update. *Endocr. Dev.* 23, 1–15. doi: 10.1159/000341733

Barbu, M. C., Spiliopoulou, A., Colombo, M., McKeigue, P., Clarke, T. K., Howard, S. M., et al. (2020). Expression quantitative trait loci-derived scores and white matter microstructure in UK Biobank: a novel approach to integrating genetics and neuroimaging. *Transl. Psychiatr.* 10, 55.

Bashamboo, A., Bignon-Topalovic, J., Moussi, N., McElreavey, K., and Brauner, R. (2017). Mutations in the human ROBO1 gene in pituitary stalk interruption syndrome. *J. Clin. Endocrinol. Metab.* 102, 2401–2406. doi: 10.1210/jc.2016-1095

Bashamboo, A., Bignon-Topalovic, J., Rouba, H., McElreavey, K., and Brauner, R. A. (2016). Nonsense Mutation in the hedgehog receptor CDON associated with pituitary stalk interruption syndrome. *J. Clin. Endocrinol. Metab.* 101, 12–15. doi: 10.1210/jc.2015-2995

BIRLA, S., Vijayakumar, P., Sehgal, S., Bhatnagar, S., Pallavi, K., and Sharma, A. (2019). Characterization of a novel POU1F1 mutation identified on screening 160 growth hormone deficiency patients. *Horm. Metab. Res.* 51, 248–255. doi: 10.1055/a-0867-1026

Castinetti, F., Reynaud, R., Quentien, M.-H., Jullien, N., Marquant, E., Rochette, C., et al. (2015). Combined pituitary hormone deficiency: current and future status. *J. Endocrinol. Invest.* 38, 1–12. doi: 10.1007/s40618-014-0141-2

Celen, C., Chuang, J. C., Luo, X., Nijem, N., Walker, A. K., Chen, F., et al. (2017). Arid1b haploinsufficient mice reveal neuropsychiatric phenotypes and reversible causes of growth impairment. *eLife* 6:e25730.

Cha, K. B., Douglas, K. R., Potok, M. A., Liang, H., Jones, S. N., and Camper, S. A. (2004). WNT5A signaling affects pituitary gland shape. *Mech. Dev.* 121, 183–194.

Cohen, E., Maghnie, M., Collot, N., Leger, J., Dastot, F., Polak, M., et al. (2017). Contribution of LHX4 mutations to pituitary deficits in a cohort of 417 unrelated patients. *J. Clin. Endocrinol. Metab.* 102, 290–301. doi: 10.1210/jc.2016-3158

Correa, F. A., Jorge, A. I., Nakaguma, M., Canton, A. P., Costa, S. S., Funari, M. F., et al. (2018). Pathogenic copy number variants in patients with congenital hypopituitarism associated with complex phenotypes. *Clin. Endocrinol. (Oxf).* 88, 425–431. doi: 10.1111/cen.13535

Correa, F. A., Nakaguma, M., Madeira, J. L. O., Nishi, M. Y., Abrão, M. G., Jorge, A. L., et al. (2019). Combined pituitary hormone deficiency caused by PROP1 mutations: update 20 years post-discovery. *Arch. Endocrinol. Metab.* 63, 167–174. doi: 10.20945/2359-3997000000139

Cox, K. B., Hamm, D. A., Millington, D. S., Matern, D., Vockley, K., Rinaldo, P., et al. (2001). Gestational, pathologic and biochemical differences between very long-chain acyl-CoA dehydrogenase deficiency and long-chain acyl-CoA dehydrogenase deficiency in the mouse. *Comp. Study Hum. Mol. Genet.* 10, 2069–2077. doi: 10.1093/hmg/10.19.2069

Dateki, S., Watanabe, S., Mishima, H., Shirakawa, T., Morikawa, M., Kinoshita, E., et al. (2019). Homozygous splice site ROBO1 mutation in a patient with a novel syndrome with combined pituitary hormone deficiency. *J. Hum. Genet.* 64, 341–346. doi: 10.1038/s10038-019-0566-8

Fallon, P. G., Sasaki, T., Sandilands, A., Campbell, L. E., Saunders, S. P., Mangan, N. F., et al. (2009). A homozygous frameshift mutation in the mouse Flg gene facilitates enhanced percutaneous allergen priming. *Nat. Genet.* 41, 602–608. doi: 10.1038/ng.358

Fatehchand, K., Ren, L., Elavazhagan, S., Fang, H., Mo, X., Vasilakos, J. P., et al. (2016). Toll-like receptor 4 ligands down-regulate Fcγ receptor IIb (FcγRIIb) via MARCH3 protein-mediated ubiquitination. *J. Biol. Chem.* 291, 3895–3904. doi: 10.1074/jbc.m115.701151

Fu, L., Pelicano, H., Liu, J., Huang, P., and Lee, C. (2002). The circadian gene Period2 plays an important role in tumor suppression and DNA damage response in vivo. *Cell* 111, 41–50. doi: 10.1016/s0092-8674(02)00961-3

Graveel, C. R., DeGroot, J. D., Su, Y., Koeman, J., Dykema, K., Leung, S., et al. (2009). Met induces diverse mammary carcinomas in mice and is associated with human basal breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 106, 12909–12914. doi: 10.1073/pnas.0810403106

Gregory, L. C. (2020). Mehul tulsidas dattani. the molecular basis of congenital hypopituitarism and related disorders. *J. Clin. Endocrinol. Metab.* 105:dgz184. doi: 10.1210/clinem/dgz184

Gregory, L. C., Gevers, E. F., Baker, J., Kasia, T., Chong, K., Josifova, D. J., et al. (2013). Structural pituitary abnormalities associated with CHARGE syndrome. *J. Clin. Endocrinol. Metab.* 98, E737–E743.

Guo, M. H., Plummer, L., Chan, Y. M., Hirschhorn, J. N., and Lippincott, M. F. (2018). Burden testing of rare variants identified through exome sequencing via publicly available control data. *Am J. Hum.Genet.* 103, 522–534.

Hamdi-Rozé, H., Ware, M., Guyodo, H., Rizzo, A., Ratié, L., Rupin, L., et al. (2020). Disrupted hypothalamo-pituitary axis in association with reduced SHH underlies the pathogenesis of NOTCH-deficiency. *J. Clin. Endocrinol. Metab.* 105, dgaa249.

Hong, M., and Krauss, R. S. (2012). Cdon mutation and fetal ethanol exposure synergize to produce midline signaling defects and holoprosencephaly spectrum disorders in mice. *PLoS Genet.* 8:e1002999. doi: 10.1371/journal.pgen.1002999

Hovinga, I. C. L. K., Giltay, J. C., van der Crabben, S. N., Steyls, A., van der Kamp, H. J., and Paulussen, A. D. C. (2018). Extreme phenotypic variability of a novel GLI2 mutation in a large family with panhypopituitarism and polydactyly: clinical implications. *Clin. Endocrinol. (Oxf).* 89, 378–380. doi: 10.1111/cen.13760

Hughes, J., Piltz, S., Rogers, N., McAninch, D., Rowley, L., and Thomas, P. (2013). Mechanistic insight into the pathology of polyalanine expansion disorders revealed by a mouse model for X linked hypopituitarism. *PLoS Genet.* 9:e1003290. doi: 10.1371/journal.pgen.1003290

Ijaz, F., Hatanaka, Y., Hatanaka, Y., Tsutsumi, K., Iwaki, T., Umemura, K., et al. (2017). Proper cytoskeletal architecture beneath the plasma membrane of red blood cells requires Ttll4. *Mol. Biol. Cell* 28, 535–544. doi: 10.1091/mbc.e16-02-0089

Jean, D., Bernier, G., and Gruss, P. (1999). Six6 (Optx2) is a novel murine Six3-related homeobox gene that demarcates the presumptive pituitary/hypothalamic axis and the ventral optic stalk. *Mech. Dev.* 84, 31–40. doi: 10.1016/s0925-4773(99)00068-4

Jee, Y. H., Sowada, N., Markello, T. C., Rezvani, I., Borck, G., and Baron, J. (2017). BRF1 Mutations in a family with growth failure, markedly delayed bone age, and central nervous system anomalies. *Clin. Genet.* 91, 739–747. doi: 10.1111/cge.12887

Jullien, N., Romanet, P., Philippon, M., Quentien, M. H., Beck-Peccoz, P., Bergada, I., et al. (2019). Heterozygous LHX3 mutations may lead to a mild phenotype of combined pituitary hormone deficiency. *Eur. J. Hum. Genet.* 27, 216–225. doi: 10.1038/s41431-018-0264-6

Kibschull, M., Colaco, K., Matysiak-Zablocki, E., Winterhager, E., and Lye, S. J. (2014). Connexin31.1 (Gjb5) deficiency blocks trophoblast stem cell differentiation and delays placental development. *Stem Cells Dev.* 23, 2649–2660. doi: 10.1089/scd.2014.0013

Kikuchi, K., Fujisawa, I., Momoi, T., Yamanaka, C., Kaji, M., Nakano, Y., et al. (1988). Hypothalamic-pituitary function in growth hormone-deficient patients with pituitary stalk transection. *J. Clin. Endocrinol. Metab.* 67, 817–823. doi: 10.1210/jcem-67-4-817

Lee, J. E., Silhavy, J. L., Zaki, M. S., Schroth, J., Bielas, S. L., Marsh, S. E., et al. (2012). CEP41 is mutated in Joubert syndrome and is required for tubulin glutamylation at the cilium. *Nat. Genet.* 44, 193–199. doi: 10.1038/ng.1078

Lian, Y. F., Yuan, J., Cui, Q., Feng, Q. S., Xu, M., Bei, J. X., et al. (2016). Upregulation of KLHDC4 predicts a poor prognosis in human nasopharyngeal carcinoma. *PLoS One* 11:e0152820. doi: 10.1371/journal.pone.0152820

Lin, H., Gao, D., Hu, M. M., Zhang, M., Wu, X. X., Feng, L., et al. (2018). MARCH3 attenuates IL-1β-triggered inflammation by mediating K48-linked polyubiquitination and degradation of IL-1RI. *Proc. Natl. Acad. Sci. U.S.A.* 115, 12483–12488. doi: 10.1073/pnas.1806217115

Lui, J. C., Jee, Y. H., Lee, A., Yue, S., Wagner, J., Donnelly, D. E., et al. (2019). QRICH1 mutations cause a chondrodysplasia with developmental delay. *Clin. Genet.* 95, 160–164. doi: 10.1111/cge.13457

Maghnie, M., Larizza, D., Triulzi, F., Sampaolo, P., Scotti, G., and Severi, F. (1991). Hypopituitarism and stalk agenesis: a congenital syndrome worsened by breech delivery? *Horm. Res.* 35, 104–108. doi: 10.1159/000181883

McCormack, S. E., Li, D., Kim, Y. J., Lee, J. Y., Kim, S. H., Rapaport, R., et al. (2017). Digenic inheritance of PROKR2 and WDR11 mutations in pituitary stalk interruption syndrome. *J. Clin. Endocrinol. Metab.* 102, 2501–2507. doi: 10.1210/jc.2017-00332

Mönnich, M., Borgeskov, L., Breslin, L., Jakobsen, L., Rogowski, M., Doganli, C., et al. (2018). CEP128 localizes to the subdistal appendages of the mother centriole and regulates TGF-β/BMP signaling at the primary cilium. *Cell Rep.* 22, 2584–2592. doi: 10.1016/j.celrep.2018.02.043

Müller, M. B., Preil, J., Renner, U., Zimmermann, S., Kresse, A. E., Stalla, G. K., et al. (2001). Expression of CRHR1 and CRHR2 in mouse pituitary and adrenal gland: implications for HPA system regulation. *Endocrinology* 142, 4150–4153. doi: 10.1210/endo.142.9.8491

Paine-Saunders, S., Viviano, B. L., Zupicich, J., Skarnes, W. C., and Saunders, S. (2000). glypican-3 controls cellular responses to Bmp4 in limb patterning and skeletal development. *Dev. Biol.* 225, 179–187. doi: 10.1006/dbio.2000.9831

Rajab, A., Kelberman, D., de Castro, S. C. P., Biebermann, H., Shaikh, H., Pearce, K., et al. (2008). Novel mutations in LHX3 are associated with hypopituitarism and sensorineural hearing loss. *Hum. Mol. Genet.* 17, 2150–2159. doi: 10.1093/hmg/ddn114

Rizzoti, K. (2015). Genetic regulation of murine pituitary development. *J. Mol. Endocrinol.* 54, R55–R73.

Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods.* 11, 361–362. doi: 10.1038/nmeth.2890

Shapshak, P. (2012). Molecule of the month, PDE4DIP. *Bioinformation* 8, 740–741. doi: 10.6026/97320630008740

Tajima, T., Hattori, T., Nakajima, T., Okuhara, K., Tsubaki, J., and Fujieda, K. A. (2007). Novel missense mutation (P366T) of the LHX4 gene causes severe combined pituitary hormone deficiency with pituitary hypoplasia, ectopic posterior lobe and a poorly developed sella turcica. *Endocr. J.* 54, 637–641. doi: 10.1507/endocrj.k06-200

Tatsi, C., Gkourogianni, A., Mohnike, K., DeArment, D., Witchel, S., Andrade, A. C., et al. (2017). Aggrecan mutations in nonfamilial short stature and short stature without accelerated skeletal maturation. *J. Endocr. Soc.* 1, 1006–1011. doi: 10.1210/js.2017-00229

Teer, J. K., Green, E. D., Mullikin, J. C., and Biesecker, L. G. (2012). VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer. *Bioinformatics* 28, 599–600. doi: 10.1093/bioinformatics/btr711

Thomas, P. Q., Dattani, M. T., Brickman, J. M., McNay, D., Warne, G., Zacharin, M., et al. (2001). Heterozygous HESX1 mutations associated with isolated congenital pituitary hypoplasia and septo-optic dysplasia. *Hum. Mol. Genet.* 10, 39–45. doi: 10.1093/hmg/10.1.39

Turton, J. P. G., Reynaud, R., Mehta, A., Torpiano, J., Saveanu, A., Woods, K. S., et al. (2005). Novel mutations within the POU1F1 gene associated with variable combined pituitary hormone deficiency. *J. Clin. Endocrinol. Metab.* 90, 4762–4770. doi: 10.1210/jc.2005-0570

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9. doi: 10.1038/nprot.2015.123

Villavicencio, E. H., Walterhouse, D. O., and Iannaccone, P. M. (2000). The sonic hedgehog–patched–gli pathway in human development and disease. *Am. J. Hum. Genet.* 67, 1047–1054. doi: 10.1016/s0002-9297(07)62934-6

Wang, Y., Martin, J. F., and Bai, C. B. (2010). Direct and indirect requirements of Shh/Gli signaling in early pituitary development. *Dev. Biol.* 348, 199–209. doi: 10.1016/j.ydbio.2010.09.024

Wu, W., Cogan, J. D., Pfäffle, R. W., Dasen, J. S., Frisch, H., O'Connell, S. M., et al. (1998). Mutations in PROP1 cause familial combined pituitary hormone deficiency. *Nat. Genet.* 18, 147–149. doi: 10.1038/ng0298-147

Wu, W., Zhai, G., Xu, Z., Hou, B., Liu, D., Liu, T., et al. (2019). Whole-exome sequencing identified four loci influencing craniofacial morphology in northern Han Chinese. *Hum. Genet.* 138, 601–611. doi: 10.1007/s00439-019-02008-6

Xie, H., Hoffmann, H. M., Meadows, J. D., Mayo, S. L., Trang, C., Leming, S. S., et al. (2015). Homeodomain proteins SIX3 and six6 regulate gonadotrope-specific genes during pituitary development. *Mol. Endocrinol.* 29, 842–855. doi: 10.1210/me.2014-1279

Yazicioglu, M. N., Monaldini, L., Chu, K., Khazi, F. R., Murphy, S. L., Huang, H., et al. (2013). Cellular localization and characterization of cytosolic binding partners for Gla domain-containing proteins PRRG4 and PRRG2. *J. Biol. Chem.* 288, 25908–25914. doi: 10.1074/jbc.m113.484683

Yu, Y., Yao, R., Wang, L., Fan, Y., Huang, X., Hirschhorn, J., et al. (2015). De novo mutations in ARID1B associated with both syndromic and non-syndromic short stature. *BMC Genomics* 16:701. doi: 10.1186/s12864-015-1898-1

Zheng-Fischhöfer, Q., Kibschull, M., Schnichels, M., Kretz, M., Petrasch-Parwez, E., Strotmann, J., et al. (2007). Characterization of connexin31.1-deficient mice reveals impaired placental development. *Dev. Biol.* 312, 258–271. doi: 10.1016/j.ydbio.2007.09.025

Zimmer, B., Piao, J., Ramnarine, K., Tomishima, M. J., Tabar, V., and Studer, L. (2016). Derivation of diverse hormone-releasing pituitary cells from human pluripotent stem cells. *Stem Cell Rep.* 6, 858–872. doi: 10.1016/j.stemcr.2016.05.005

Zwaveling-Soonawala, N., Alders, M., Jongejan, A., Kovacic, L., Duijkers, F. A., Maas, S. M., et al. (2018). Clues for polygenic inheritance of pituitary stalk interruption syndrome from exome sequencing in 20 patients. *J. Clin. Endocrinol. Metab.* 103, 415–428. doi: 10.1210/jc.2017-01660

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer MD declared a past co-authorship with one of the authors JB to the handling editor.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Jee, Gangat, Yeliosof, Temnycky, Vanapruks, Whalen, Gourgari, Bleach, Yu, Marshall, Yanovski, Link, Ten, Baron and Radovick. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Pediatrics

ORIGINAL RESEARCH
published: 23 August 2021
doi: 10.3389/fped.2021.724196

Molecular Characterization of α- and β-Thalassaemia Among Children From 1 to 10 Years of Age in Guangxi, A Multi-Ethnic Region in Southern China

Sheng He[1†], Dongming Li[1†], Shang Yi[1], Xiuning Huang[1], Chaofan Zhou[1], Biyan Chen[1], Yangjin Zuo[1], Li Lin[1], Faqin Chen[2] and Hongwei Wei[1*]*

[1] Key Laboratory of Basic Research for Guangxi Birth Defects Control and Prevention, Guangxi Zhuang Autonomous Region Women and Children Care Hospital, Nanning, China, [2] Department of Laboratory Medicine, Youjiang Medical University for Nationalities, Baise, China

OPEN ACCESS

Edited by:
Blanka Rogina,
University of Connecticut Health Center, United States

Reviewed by:
Emanuele Micaglio,
IRCCS Policlinico San Donato, Italy
Jernej Kovac,
University Medical Center Ljubljanaju, Slovenia

***Correspondence:**
Faqin Chen
chenfaqin1965@163.com
Hongwei Wei
weihongwei1965@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Pediatrics

Received: 12 June 2021
Accepted: 27 July 2021
Published: 23 August 2021

Citation:
He S, Li D, Yi S, Huang X, Zhou C, Chen B, Zuo Y, Lin L, Chen F and Wei H (2021) Molecular Characterization of α- and β-Thalassaemia Among Children From 1 to 10 Years of Age in Guangxi, A Multi-Ethnic Region in Southern China. Front. Pediatr. 9:724196. doi: 10.3389/fped.2021.724196

Background: Thalassemia is one of the most common genetic diseases in southern China. However, population in different regions or different population has their own spectrums of thalassemia. To investigate the prevalence and spectrum features of thalassemia among children in Guangxi. Hematology and genetic analysis were performed on 71,459 children aged 1–10 years in various regions of Guangxi.

Results: A total of 11,821 children were diagnoses with thalassemia including 7,615 (10.66%) subjects of α-thalassemia, 3,507 (4.90%) subjects of β-thalassemia, and 699 (0.98%) cases with both α- and β-thalassemia. Nine α-thalassemia mutations and 30 genotypes were identified among the α-thalassemia children. The $--^{SEA}$ and $--^{SEA}/\alpha\alpha$ were the most frequent mutation and genotype, respectively. One α-thalassemia fusion gene and a rare 2.4 kb deletion both causing α^{+}-thalassemia were identified, respectively. Thirteen β-thalassemia mutations and 31 genotypes were characterized among the β-thalassemia children, with the most common mutation CD41-42 (-CTTT) accounting for 46.05% of the β-mutations. Two rare mutations IVS-II-5 (G>C), and IVS-I-2 (T>C) were firstly identified. Furthermore, 92 genotypes were identified among 699 children with both α- and β-thalassemia.

Conclusions: Our findings highlight the great heterogeneity and the extensive spectrum of thalassemia among children in Guangxi, which provide an available reference for prevention of thalassemia in this area.

Keywords: thalassemia, prevalence and spectrum, children, Guangxi, China

INTRODUCTION

The thalassemias are a group of inherited hemoglobin disorders characterized by microcytic hypochromic anemia resulting from reduced or absent synthesis of one or more of the globin chains of hemoglobin (1). According to the type of globin involved, thalassemias are divided into α-, β-, δ-thalassemia, with α- and β-thalassemia most widely distributed worldwide (2). Clinical phenotype of people with thalassemia varies from almost asymptomatic to a lethal hemolytic

anemia. Carriers of heterozygous globin gene mutations or patients with mild thalassaemia may not have symptoms of anemia. However, patients with severe α-thalassemia (i.e., hydrops fetalis) always die *in utero* or soon after birth and sometimes leading to the mortality of the pregnant mother (3, 4). β-thalassemia major patients have a severe form of anemia requiring lifelong transfusion and may have shortened life span (3). To date, there is no effective treatment for patients with thalassemia major, except bone marrow transplantation (2). Thalassemia major has caused a huge burden on society and seriously affected the quality of life of people in developing countries (3, 5). Therefore, it is critical to implement prenatal diagnosis and genetic counseling to prevent the birth of children with thalassemia major on the basis of grasping the molecular epidemiological characteristics of the frequency and distribution of thalassemia.

Approximately 5% of the global population are carriers of thalassemia (2, 6). Thalassemias are particularly frequent in Mediterranean countries, the Middle East, Africa, and Southeast Asia (2). In China, thalassemia is widely distributed in the southern bank of the Yangtze River, particularly in the three southern provinces of Guangdong, Guangxi, and Jiangxi (7–9). Previous studies showed that the frequency of thalassemia reached about 20% in Guangxi province, a multi-ethnic region. However, these studies only focused on the adult population, and there has been no report about the spectrum thalassemia among children (8, 10, 11). Since the spectrum, and frequencies of α- and β-thalassaemia alleles vary considerably with geographic locations and different populations (12), this study was performed to characterize α- and β-thalassaemia mutations at the molecular level among children in Guangxi province.

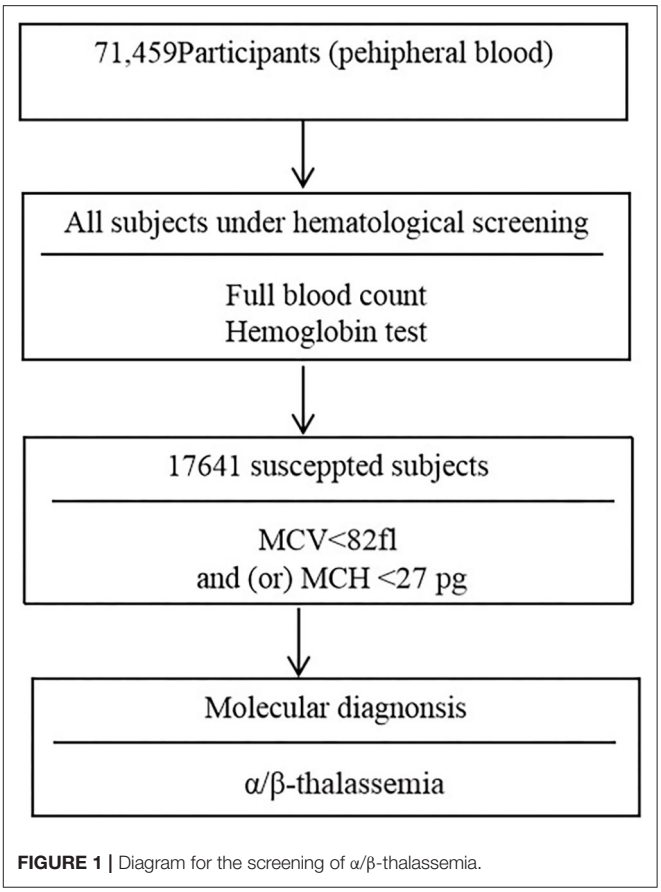


FIGURE 1 | Diagram for the screening of α/β-thalassemia.

MATERIALS AND METHODS

Human Subjects

This study population included 71,459 aged from 1 to 10-year-old children in difference regions of Guangxi for hemoglobinopathy screening between January 1, 2011 and December 31, 2019. These subjects visited medical units for routine healthy examination including blood tests, and the discarded blood samples were used for further study. Information records of nationality, sex, age, dialect, Guangxi aborigines or not and written consent forms were available in Chinese to ensure comprehensive understanding of the study objectives. Informed consent was signed by the participants or their guardians. All studies were approved by the Ethic Committee of Guangxi Zhuang Autonomous Region Women and Children Care Hospital. All procedures were carried out in accordance with ethical guidelines for human subject's research.

Hematological Analysis

Two ml volume of peripheral venous blood samples were collected into an EDTA anticoagulated tube. Peripheral blood counts and red blood cell incidences were determined using a SYSMEX XE800i automatic blood cell analyzer (Kobe, Japan). Hemoglobin electrophoresis was applied to analyze the percentage of hemoglobin Hbs A, A_2, F, and any abnormal hemoglobin variant. The children with microytosis [low mean corpuscular volume (MCV) values (<82 fl) and/or low mean cell hemoglobin (MCH) values (<27 pg)] after the exclusion of iron deficiency anemia were consider suspected thalassemia children. Moreover, children with low HbA_2 (<2.5%) and high HbA_2 (≥3.5%) values were considered possible α-thalassemia carriers and β-thalassemia carriers, respectively (7, 13). The flowchart of screening strategy used in this study is illustrated in **Figure 1**.

Genetic Analysis

Genomic DNA of subjects with microcytosis were isolated using a DNA blood extraction kit (Tiangen Bio-Tech Co. Ltd., Beijing, China) according to the manufacturer's instructions. The four known α-thalassemia deletionas [- -^{SEA} (NG_000006.1:g.26264_45564del19301), -$\alpha^{3.7}$ (NG_000006.1:g.34164_37967del3804), -$\alpha^{4.2}$ (AF221717), - -^{THAI} (NG_000006.1:g/10664_44164del33501)], three α-thalassemia mutations [Hb Constant Spring, Hb CS (*HBA2*:c.427T>C), Hb Quong Sze (Hb QS, *HBA2*:c.377T>C), and Hb Westmead (Hb WS, *HBA2*:c.369C>G)] and 17 known β-thalassemia mutations most commonly seen in the Chineses population [CD41-42(-TCTT) (*HBB*:c.126_129delCTTT), IVS-II-654(C→T) (*HBB*:c.316-197C>T), CD17(A→T) (*HBB*:c.52A>T),−28(A→G) (*HBB*:c.-78A>G), CD26 (G→ A)

TABLE 1 | Distribution of genotypes among the 7,615 α-thalassemia children in Guangxi.

Genotype	Number	Phenotype	Frequency (%)
Heterozygote	6,734		88.43
$--^{SEA}/\alpha\alpha$	4,075	α^0/α	53.51
$-\alpha^{3.7}/\alpha\alpha$	1,157	α^+/α	15.19
$\alpha^{CS}\alpha/\alpha\alpha$	622	α^+/α	8.17
$-\alpha^{4.2}/\alpha\alpha$	488	α^+/α	6.41
$\alpha^{WS}\alpha/\alpha\alpha$	250	α^+/α	3.28
$\alpha^{QS}\alpha/\alpha\alpha$	88	α^+/α	1.16
$--^{THAI}/\alpha\alpha$	53	α^0/α	0.70
$-\alpha^{2.4}/\alpha\alpha$	1	α^+/α	0.01
Compound heterozygote	799		10.49
$--^{SEA}/-\alpha^{3.7}$	281	α^0/α^+	3.69
$--^{SEA}/-\alpha^{4.2}$	143	α^0/α^+	1.88
$--^{SEA}/\alpha^{CS}\alpha$	129	α^0/α^+	1.69
$--^{SEA}/\alpha^{WS}\alpha$	72	α^0/α^+	0.95
$-\alpha^{3.7}/\alpha^{CS}\alpha$	40	α^+/α^+	0.53
$-\alpha^{3.7}/-\alpha^{4.2}$	31	α^+/α^+	0.41
$-\alpha^{3.7}/\alpha^{WS}\alpha$	26	α^+/α^+	0.34
$--^{SEA}/\alpha^{QS}\alpha$	17	α^0/α^+	0.22
$-\alpha^{4.2}/\alpha^{CS}\alpha$	14	α^+/α^+	0.18
$-\alpha^{4.2}/\alpha^{WS}\alpha$	12	α^+/α^+	0.16
$\alpha^{WS}\alpha/\alpha^{CS}\alpha$	11	α^+/α^+	0.15
$-\alpha^{3.7}/\alpha^{QS}\alpha$	7	α^+/α^+	0.09
$\alpha^{WS}\alpha/\alpha^{QS}\alpha$	6	α^+/α^+	0.08
$--^{THAI}/-\alpha^{3.7}$	5	α^0/α^+	0.07
$-\alpha^{4.2}/\alpha^{QS}\alpha$	1	α^+/α^+	0.01
$--^{THAI}/-\alpha^{4.2}$	1	α^0/α^+	0.01
$--^{THAI}/\alpha^{WS}\alpha$	2	α^0/α^+	0.03
$--^{THAI}$ αfusion gene	1	α^0/α^+	0.01
Homozygote	82		1.08
$-\alpha^{3.7}/-\alpha^{3.7}$	48	α^+/α^+	0.63
$\alpha^{CS}\alpha/\alpha^{CS}\alpha$	14	α^+/α^+	0.18
$-\alpha^{4.2}/-\alpha^{4.2}$	13	α^+/α^+	0.17
$\alpha^{WS}\alpha/\alpha^{WS}\alpha$	7	α^+/α^+	0.09
Total	7,615		100.00

α, normal production of the α-globin polypeptide chain; α0, no production of the α-globin polypeptide chain; α+, reduced production of the α-globin polypeptide chain.

(*HBB*:c.79G>A), CD71-72 (+A), (*HBB*:c.216_217insA), CD43 (G→T) (*HBB*:c.130G>T),−29 (A→G) (*HBB*:c.-79A>G), CD14-15 (+G) (*HBB*:c.45_46insG), CD27-28 (+ C) (*HBB*:c.84_85insC),−32 (C→A) (*HBB*:c.-82C>A),−30 (T→C) (*HBB*:c.-80T>C), IVS-I-1(G→T) (*HBB*:c.92+1G>T), IVS-I-5 (G→ C) (*HBB*:c.92+5G>A), CD31 (-C) (*HBB*:c.94delC), and CapM (-AAAC) (*HBB*:c.-50A>C) were identified as described by He et al. (13). Multiplex ligation-dependent probe amplification (MLPA) (Service XS^{TM} Leiden, the Netherlands) was performed to detect gross deletions in the α- or β-globin gene cluster in case of new mutations. DNA sequencing was applied to detect unknown and rare globin mutations.

Statistical Analysis

Statistical analysis was conducted with SSPS 17.0 software. The prevalence of different thalassemia alleles was calculated from modified Hardy-Weinberg formula.

RESULTS

The Characteristic of Thalassemia Among Children in Guangxi Province

From 2011 to 2019, a total of 17,641 (24.69%) children with microcytosis were found. 11,821 out of the 17,641 suspected children were diagnosed with thalassemia, including 7,615 (10.64%) subjects of α-thalassemia, 3,507 (4.90%) subjects of β-thalassemia and 699 (0.98%) subjects with both α- and β-thalassemia. Among the 7,615 α-thalassemia children, 6,734 (88.43%) cases had heterozygote mutations and 881 cases (11.57%) were with homozygote and compound heterozygote mutations. The - $--^{SEA}/\alpha\alpha$ was the most frequent genotype, accounting for more than half of all α-thalassemia genotypes (53.51%). The other most common genotypes of α-thalassemia were - $-\alpha^{3.7}/\alpha\alpha$, $\alpha^{CS}\alpha/\alpha\alpha$, - $-\alpha^{4.2}/\alpha\alpha$, $--^{SEA}/-\alpha^{3.7}$, and $\alpha^{WS}\alpha/\alpha\alpha$ (**Table 1**).

Distribution of Thalassemia Genotypes Among Children

Thirty-five genotypes were found in 3,507 β-thalassemia children, including 3,419 heterozygotes, 31 mutant homozygotes, and 57 compound heterozygotes, accounting for 97.49, 0.88, and 1.63%, respectively (**Table 2**). $\beta^{CD41-42}/\beta^N$ was the most prevalent genotype, accounting for 45.11% of all β-thalassemia genotypes. Most of the remaining genotypes were β^{CD17}/β^N, $\beta^{IVS-II-654}/\beta^N$, $\beta^{CD71-72}/\beta^N$, β^{-28}/β^N, β^{CD26}/β^N, $\beta^{IVS-I-1}/\beta^N$, and β^{CD43}/β^N. Overall, these eight genotypes accounted for 96.72% of all β-thalassemia genotypes (**Table 2**). Six hundred and ninety nine children with 92 genotypes carried both α and β-globin mutations. Among these children, four top frequent types were $--^{SEA}$ /αα combined with $\beta^{CD41-42}/\beta^N$ (17.60%), $-\alpha^{3.7}/\alpha\alpha$ combined with $\beta^{CD41-42}/\beta^N$ (9.88%), $-^{SEA}/\alpha\alpha$ combined with β^{CD17}/β^N (9.02%), and - $\alpha^{3.7}/\alpha\alpha$ combined with β^{CD17}/β^N (7.30%) (**Supplementary Table 1**).

Allele Frequency of α-Thalassemia and β-Thalassemia Among Children

Tables 3, **4** showed the frequency of a specific type of mutation in all α (or β) mutant chromosomes in thalassemia children. Twenty-four mutations were identified including nine alpha gene mutations and 15 beta gene mutations. The - $--^{SEA}$ was the most frequent mutation, accounting for 54.23% of all α mutant chromosomes. The other high frequency mutations were $-\alpha^{3.7}$, $\alpha^{CS}\alpha$, α^{WS} α, and $-\alpha^{4.2}$, with allele frequencies of 19.76, 9.92, 8.58, and 5.40%, respectively (**Table 3**). Two rare mutations: a fusion and $-\alpha^{2.4}$ were identified in these children. Of β-globin mutant chromosomes, eight mutations accounted for as much as 99.22% of all β-thalassemia defects. These mutations in the order of frequency were CD41-42 (-CTTT) (46.07%), CD17 (A>T) (28.44%), IVS-II-654(C>T) (7.10%),−28 (A>G) (4.84%), CD26

TABLE 2 | Distribution of genotypes among the 3,507 β-thalassemia children in Guangxi.

Genotype	Cases	Phenotype	Frequency (%)
Heterozygote	3,419		97.49
$\beta^{CD41-42}/\beta^{N}$	1,582	β^{0}/β^{N}	45.11
β^{CD17}/β^{N}	974	β^{0}/β^{N}	27.77
$\beta^{IVS-II-654}/\beta^{N}$	248	β^{0}/β^{N}	7.07
$\beta^{CD71-72}/\beta^{N}$	160	β^{0}/β^{N}	4.56
β^{-28}/β^{N}	157	β^{+}/β^{N}	4.48
β^{CD26}/β^{N}	145	β^{0}/β^{N}	4.14
$\beta^{IVS-I-1}/\beta^{N}$	79	β^{0}/β^{N}	2.25
β^{CD43}/β^{N}	47	β^{0}/β^{N}	1.34
β^{-29}/β^{N}	10	β^{+}/β^{N}	0.28
β^{27-28}/β^{N}	10	β^{0}/β^{N}	0.28
$\beta^{CD14-15}/\beta^{N}$	4	β^{0}/β^{N}	0.11
$\beta^{VS-II-5}/\beta^{N}$	1	β^{+}/β^{N}	0.03
β^{VS-I-5}/β^{N}	1	β^{+}/β^{N}	0.03
β^{CAP+1}/β^{N}	1	β^{+}/β^{N}	0.03
Homozygote	31		0.88
$\beta^{CD41-42}/\beta^{CD41-42}$	19	β^{0}/β^{0}	0.54
$\beta^{CD17}/\beta^{CD17}$	9	β^{0}/β^{0}	0.26
$\beta^{IVS-I-1}/\beta^{IVS-I-1}$	3	β^{0}/β^{0}	0.09
Compound heterozygote	57		1.63
$\beta^{CD41-42}/\beta^{CD17}$	14	β^{0}/β^{0}	0.40
$\beta^{CD41-42}/\beta^{-28}$	7	β^{0}/β^{+}	0.20
$\beta^{41-42}/\beta^{CD26}$	5	β^{0}/β^{+}	0.14
$\beta^{CD41-42}/\beta^{IVS-II-654}$	4	β^{0}/β^{+}	0.11
β^{CD17}/β^{-28}	4	β^{0}/β^{+}	0.11
$\beta^{CD71-72}/\beta^{CD17}$	4	β^{0}/β^{0}	0.11
$\beta^{CD41-42}/\beta^{IVS-I-1}$	4	β^{+}/β^{0}	0.11
$\beta^{CD17}/\beta^{CD26}$	3	$\beta^{0}\beta^{0}$	0.09
$\beta^{IVS-II-654}/\beta^{CD17}$	3	β^{0}/β^{0}	0.09
$\beta^{CD71-72}/\beta^{CD26}$	1	β^{0}/β^{+}	0.03
$\beta^{CD17}/\beta^{IVS-I-1}$	1	β^{0}/β^{0}	0.03
$\beta^{CD43}/\beta^{CD17}$	1	β^{+}/β^{0}	0.03
$\beta^{CD41-42}/\beta^{CD71-72}$	1	β^{0}/β^{0}	0.03
$\beta^{CD41-42}/\beta^{CD27/28}$	1	β^{0}/β^{+}	0.03
$\beta^{IVS-II-654}/\beta^{CD26}$	1	β^{0}/β^{+}	0.03
$\beta^{IVS-II-654}/\beta\beta^{IVS-II-5}$	1	β^{0}/β^{+}	0.03
$\beta^{CD41-42}/\beta^{CAP}$	1	β^{0}/β^{+}	0.03
β^{-28}/β^{CD26}	1	β^{0}/β^{+}	0.03
Total	3,507		100.00

β^{N}, normal production of the β-globin polypeptide chain; β^{0}, no production of the β-globin polypeptide chain; β^{+}, impaired production of the β-globin polypeptide chain.

(Hb E) (G>A) (4.58%), CD71-72 (+A) (4.23%), IVS-I-1 (G>T) (2.66%), and CD43 (G>T) (1.30%). The other seven mutations with frequencies of no more than 1% were−29 (A>G), CD27-28 (+ C), CD14-15 (+G), IVS-II-5 (G>C), CAP (A>C), IVS-I-5 (G>C), IVS-I-2 (T>C) (**Table 4**).

TABLE 3 | Allele frequency of α-thalassemia among the 71,459 children in Guangxi.

Mutation	HGVS name	Type	Allele (*n*)	Frequency (%)
$--^{SEA}$	NG_000006.1:g.26264_45564del19301	α^{0}	5,008	54.23
$-\alpha^{3.7}$	NG_000006.1:g.34164_37967del3804	α^{+}	1,825	19.76
$\alpha^{CS}\alpha$	*HBA2*:c.427T>C	α^{+}	916	9.92
$-\alpha^{4.2}$	AF221717	α^{+}	792	8.58
$\alpha^{WS}\alpha$	*HBA2*:c.369C>G	α^{+}	499	5.40
$\alpha^{QS}\alpha$	*HBA2*:c.377T>C	α^{+}	127	1.38
$--^{THAI}$	NG_000006.1:g/10664_44164del33501	α^{0}	66	0.71
$-\alpha^{2.4}$	*HBA1*:g36859_39252del2392	α^{+}	1	0.01
α^{fusion}		α^{+}	1	0.01
Total			9,235	100.00

α, normal production of the α-globin polypeptide chain; α^{0}, no production of the α-globin polypeptide chain; α^{+}, reduced production of the α-globin polypeptide chain.

TABLE 4 | Allele frequency of β-thalassemia among the 71,459 children in Guangxi.

Mutation	HGVS name	Type	Allele (*n*)	Frequency (%)
CD41-42 (-TTCT)	*HBB*:c.126_129delCTTT	β^{0}	1,991	46.07
CD17 (A>T)	*HBB*:c.52A>T	β^{0}	1,229	28.44
IVS-II-654 (C>T)	*HBB*:c.316-197C>T	β^{+}	307	7.10
−28 (A>G)	*HBB*:c.-78A>G	β^{+}	209	4.84
CD26 (G>A)	*HBB*:c.79G>A	β^{+}	198	4.58
CD71-72 (+A)	*HBB*:c.216_217insA	β^{0}	183	4.23
IVS-I-1 (G>T)	*HBB*:c.92+1G>T	β^{0}	115	2.66
CD43 (G>T)	*HBB*:c.130G>T	β^{0}	56	1.30
−29 (A>G)	*HBB*:c.-79A>G	β^{+}	13	0.30
CD27-28 (+C)	*HBB*:c.84_85insC	β^{0}	11	0.25
CD14-15 (+G)	*HBB*:c.45_46insG	β^{0}	4	0.09
IVS-II-5 (G>C)	*HBB*:c.315+5G>C	β^{0}	2	0.05
CAP (A>C)	*HBB*:c.-50A>C	β^{+}	2	0.05
IVS-I-5 (G>A)	*HBB*:c.92+5G>A	β^{+}	1	0.02
IVS-I-2 (T>C)	*HBD*:c.92+2T>C	(deltabeta) zero	1	0.02
Total			4,322	100.00

β^{N}, normal production of the β-globin polypeptide chain; β^{0}, no production of the β-globin polypeptide chain; β^{+}, impaired production of the β-globin polypeptide chain.

DISCUSSION

In the present study, we have firstly reported the prevalence of thalassemia at molecular level and the related mutation among children in Guangxi, a southwestern province in China. The data indicated there was a high prevalence of thalassemia among children in this region. The prevalence of α-, β-, and α + β-thalassemia was 10.64, 4.90, and 0.98%, respectively. The overall

prevalence was 16.52%. It is was slightly lower than the average level of Guangxi province (19.52%) (8). The difference may be due to the population and the strategy selected. Previous studies were based on epidemiological surveys in mixed populations including neonates, children, adults, and pregnant women (8, 10, 11). Whereas, the research subjects in our study were only focused on the children from 1 to 10-year-old in various regions of the province and can more objectively reflect the rate among the general children. Traditional methods using hematological test as a primary screening for thalassemia is a limitation in our study. Some silent thalassemia carriers with normal or borderline red cell indices and/or HbA_2 levels might be missed and were not detected (14, 15). Recently, screening for both α- and β-thalassemia genes by next-generation sequencing (NGS) has been introduced. Limitations of our study, such as possible missed carriers during primary screening will be further addressed in the future (15, 16). When compared with the overall prevalence with other regions in China, the frequencies of thalassemia was higher than that in Guangdong (11.07%) (8), Yunnan (9.7%) (17), Fujian (4.57%) (18), Sichuan (3.43%) (19), and Jiangxi province (9.49%) (9) of China.

Previous studies have shown that α-thalassemia is mainly caused by three types of gene deletions ($--^{SEA}$, $-\alpha^{3.7}$, and $-\alpha^{4.2}$) and two types of gene non-deletion mutations ($\alpha^{CS}\alpha$ and $\alpha^{WS}\alpha$) in Chinese population (7, 8, 16). In this study, five deletions and 3 non-deletions with 26 genotypes were identified among α-thalassemia children. The $- -^{SEA}$ and $- -^{SEA}/\alpha\alpha$ were the most frequent α-thalassemia mutation and genotype, respectively, which is consistent with the observations in Guangxi adult population and Yunnan children (8, 14). In addition, the Thailand deletion ($--^{THAI}$) which was previously described in Southeast Asian and Taiwan aboriginals (20, 21), and be regarded as rare mutation in mainland China was identified in 65 children, with the allele frequency of 0.7% in our study, which indicates that it should be considered a common mutation. Moreover, we identified an α-thalassemia fusion gene in this study. This fusion gene was formed by the fusion of α2 and $\Psi\alpha1$ sequence of α-globin gene, which changed the $3'$UTR of the α2 gene and gave rise to a polyadenylation signal mutation. A polyadenylation signal mutation could produce an extensive transcript of the α2 gene and cause α^+-thalassemia (22), which was confirmed by the hematology parameters of the child. These findings enriched the gene mutation database of thalassemia among children in Guangxi province and might have great significance for genetic counseling, thalassemia genetic diagnosis, and prevention.

In contrast of α-thalassemia, most cases of β-thalassemia are caused by point mutations which lead to a reduction (β^+) or complete absence (β^0) of β-globin chain synthesis from the affected allele (23). At present, more than 300 mutations in the β-globin gene have been characterized world-wide (15). In China, 17 common and more than 23 rare β-globin gene mutations have been found, while the type frequency and spectrum of these mutations vary considerably with geographical location and different populations (9). In this study, 13 mutations with 35 genotypes were identified in the 3,507 β-thalassemia children. CD41-42(-TTCT) is the most frequent β-thalassemia mutation with an allele frequency of 46.07%, which is similar to the situation reported in other regions such as the Guilin Region (52.02%) (24), Guangxi Province (48.37%) (11), Guangdong province (36.4%) (7). Whereas, it was different from Guizhou province, Meizhou city of Guangdong Province, and Yunnan province (17, 25, 26). CD17 (A>T) was the most frequent β-thalassemia mutation in Baise Region and Guizhou province (13, 25). IVS-II-654 and CD26 (β^E) are the principal mutations of β-thalassemia in Meizhou city of Guangdong Province (26), and in Yunnan region, respectively [18]. Additionally, two rare mutations IVS-II-5 (G>C) and IVS-I-2 (T>C) were identified in the children, which was the first report of the abnormal hemoglobin in the Chinese population according to our knowledge. The distribution of the mutation spectrum with the geographical locations reflected the heterogeneity of β-thalassemia and highlighted the screen for beta thalassemia in the different populations and locations.

Ninety-two genotypes were identified among 699 children with both α- and β-thalassemia, which indicates the molecular background of α and β-thalassemia mutation among the children in Guangxi is more complex and extensive heterogeneity than that in other areas of China. Guangxi is a multi-ethnic province, where nearly 40 different ethnic groups live, thus the heterogeneity may be caused by population admixture, marriage patterns, and migrations. With a better knowledge of the thalassemia mutations present in different populations, it may be possible to develop a more rational approach to population screening for control and counseling of disease in an area (7). Our study has analyzed mutations among children that have shown patterns in Guangxi which can be used for quicker and more convenient identification of mutations while conducting newborn screening and prenatal diagnosis, and provide dates for a large population screening plan of thalassemia in Guangxi province.

CONCLUSION

In conclusion, our study has demonstrated the great heterogeneity and the extensive spectrum of α-thalassemia and β-thalassemia mutations among children in Guangxi. The findings provide the valuable information for pre-marital and pre-pregnancy screening, prenatal diagnostic services, and designing appropriate prevention programs to control the incidence of severe thalassemia in this area.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The study was approved by the Ethic Committee of Guangxi Zhuang Autonomous Region Women and Children Care Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SH and DL conceived the experiments and wrote the manuscript. SY and XH performed data analyzed and summarized results performed. CZ and BC collected the data. YZ and LL performed part of genetic tests. FC and HW conceived the idea of the study. All authors have read and approved the manuscript.

FUNDING

This study was supported by Natural Science Foundation of China (81760615), Natural Science Foundation of Guangxi (GuikeZy1949016 and Gui 14124004-1-5), and the Health Department of Guangxi Province (S201613).

ACKNOWLEDGMENTS

We thanks all the people who participated in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fped.2021.724196/full#supplementary-material

REFERENCES

1. Thar AT, Weatherall DJ, Cappellini MD. Thalassaemia. *Lancet.* (2018) 391:155–67. doi: 10.1016/S0140-6736(17)31822-6
2. Vichinsky EP. Clincial manifestations of α-thalassemia. *Cold Spring Harb Perspect Med.* (2013) 3:a011742. doi: 10.1101/cshperspect.a011742
3. Weatherall DJ, Clegg JB. Inherited haemoglobin disorders: an increasing global health problem. *Bull World Health Organ.* (2001) 79:704–12. doi: 10.1590/S0042-96862001000800005
4. Weatherall DJ. Keynote address: the challenge of thalassemia for the developing countries. *Ann N Y Acad Sci.* (2005) 1054:11–7. doi: 10.1196/annals.1345.002
5. Kattamis A, Fomi GL, Aydinok Y, Viprakasit V. Changing patterns in the epidemiology of β-thalassemia. *Eur J Haematol.* (2020) 105:692–703. doi: 10.1111/ejh.13512
6. Angastiniotis M, Lobitz S. Thalassemias: an overview. *Int J Neonatal Screen.* (2019) 5:16. doi: 10.3390/ijns5010016
7. Xu XM, Zhou YQ, Lou GX, Liao C, Chen PY, Lu JP, et al. The prevalence and spectrum of α and β-thalassemia in Guangdong Province: implications for the further health burden and population screening. *J Chin Pathol.* (2014) 57:517–22. doi: 10.1136/jcp.2003.014456
8. Zheng CG, Liu M, Du J,Chen K, Yang YG, Yang Z. Molecular spectrum of α- and β-globin mutations detected in the population of Guangxi Zhuang Autonomous Region, people's republic of China. *Hemoglobin.* (2011) 35:28–39. doi: 10.3109/03630269.2010.547429
9. Lin M, Zhong TY, Chen YG, Wang JZ, Wu JR, Lin F, et al. Molecular epidemiological characterization and health burden of thalassemia in Jiangxi Province, P.R. China. *PLoS ONE.* (2014) 9:e101505. doi: 10.1371/journal.pone.0101505
10. Pan HF, Long G F, Li Q, Feng YN, Lei ZY, Wei HW, et al. Current status of thalassemia in minority populations in Guangxi, China. *Clin Genet.* (2007) 71:419–26. doi: 10.1111/j.1399-0004.2007.00791.x
11. Xiong F, Sun M, Zhang X, Cai R, Zhou Y, Luo J, et al. Molecular epidemiological survey of haemoglobinopathies in Guangxi Zhuang Autonomous Region of southern China. *Clin Genet.* (2010) 78:139–48. doi: 10.1111/j.1399-0004.2010.01430.x
12. Yatim NF, Rahim MA, Menon K, Al-Hassan FM, Ahmad R, Manocha AB, et al. Molecular characterization of α- and β-Thalassaemia among Malay patients. *Int J Mol Sci.* (2014) 15:8835–45. doi: 10.3390/ijms15058835
13. He S, Qin Q, Yi S, Wei Y, Lin L, Chen SK, et al. Prevalence and genetic analysis of α- and β-thalassemia in Baise region, a multi-ethnic region in Southern China. *Gene.* (2017) 619:71–5. doi: 10.1016/j.gene.2016.02.014
14. Piel FB, Weatherall DJ. The alpha-thalassemias. *N Engl J Med.* (2014) 371:1908–16. doi: 10.1056/NEJMra1404415
15. Shang X, Peng ZY, Ye YH, Asan, Zhang XH, Chen Y, et al. Rapid targeted next-generation sequencing platform for molecular screening and clinical genotyping in subjects with hemoglobinopathies. *EBioMedicine.* (2017) 23:150–9. doi: 10.1016/j.ebiom.2017.08.015
16. Zhao JH, Li J, Lai QH, Yu YP. Combined use of gap-PCR and next-generation sequencing improves thalassaemia carrier screening among premarital adults in China. *J Clin Pathol.* (2020) 73:488–92. doi: 10.1136/jclinpath-2019-206339
17. Zhang J, Zhu BS, He J, Zeng XH, Su J, Xu XH, et al. The spectrum of α- and β-thalassemia mutations in Yunnan Province of Southwestern China. *Hemoglobin.* (2012) 36:464–73. doi: 10.3109/03630269.2012.717327
18. Xu LP, Huang H, Yan W, Lin Z, Yuan L. Molecular epidemiological analysis of α- and β- thalassemia in Fujian province. *Chin J Med Genet.* (2013) 30:403–6. doi: 10.3760/cma.j.issn.1003-9406.2013.04.005
19. Yu X, Yang LY, Yang HT, Liu CG, Cao DC, Shen W, et al. Molecular epidemiological investigation of thalassemia in the Chengdu region, Sichuan Province, Southwest China. *Hemoglobin.* (2015) 39:393–7. doi: 10.3109/03630269.2015.1070733
20. Fischel-Ghodsian N, Vickers MA, Ship M, Winichagoon P, Higgs DR. Characterization of two deletions that remove the entire human zeta-globin gene complex (-THAI and -FIL). *Br J Haematol.* (1988) 70:233. doi: 10.1111/j.1365-2141.1988.tb02469.x
21. Ko TM, Chen TA, Hsieh MI, Tseng LH, Hsieh FJ, Chuang SM, et al. Alpha-thalassemia in the four major aboriginal groups in Taiwan. *Hum Genet.* (1993) 92:79–80. doi: 10.1007/BF00216149
22. Proudfoot NJ, Maniatis T. The structure of a human α-globin pseudogene and its relationship to α-globingene duplication. *Cell.* (1980) 21:537–44. doi: 10.1016/0092-8674(80)90491-2
23. Origa R. β-thalassemias. *Genet Med.* (2017) 19:609–19. doi: 10.1038/gim.2016.173
24. Tang WJ, Zhang CL, Lu FF, Tang J, Lu Y, Cui X, et al. Spectrum of α- and β-globin gene mutations in Guilin Region of southern China. *Clin Bio.* (2015) 48:1068–72. doi: 10.1016/j.clinbiochem.2015.06.008
25. Huang SW, Liu XM, Li GF, Sun L, Wu X, Wang RL. Spectrum of β-thalassemia mutations in Guizhou Province, PR China, including first observation of codon 121 (GAA>TAA) in Chinese population. *Clin Biochem.* (2013) 46:1865–8. doi: 10.1016/j.clinbiochem.2013.09.014
26. Wu H, Zhu Q, Zhong H, Yu Z, Zhang Q, Huang Q. Analysis of genotype distribution of thalassemia and G6PD defificiency among Hakka population in Meizhou city of Guangdong province. *J Clin Lab Anal.* (2020) 34:e23140. doi: 10.1002/jcla.23140

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 He, Li, Yi, Huang, Zhou, Chen, Zuo, Lin, Chen and Wei. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 30 September 2021
doi: 10.3389/fgene.2021.713230

From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries

Binglan Li[1] *and Marylyn D. Ritchie*[2,3]*

[1] Department of Biomedical Data Science, Stanford University, Stanford, CA, United States, [2] Department of Genetics, University of Pennsylvania, Philadelphia, PA, United States, [3] Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States

OPEN ACCESS

Edited by:
Jordi Pérez-Tur,
Consejo Superior de Investigaciones Científicas (CSIC), Spain

Reviewed by:
Ping Zeng,
Xuzhou Medical University, China
Chong Wu,
Florida State University, United States
Heather E. Wheeler,
Loyola University Chicago, United States

***Correspondence:**
Marylyn D. Ritchie
marylyn@pennmedicine.upenn.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 22 May 2021
Accepted: 27 July 2021
Published: 30 September 2021

Citation:
Li B and Ritchie MD (2021) From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. Front. Genet. 12:713230. doi: 10.3389/fgene.2021.713230

Since their inception, genome-wide association studies (GWAS) have identified more than a hundred thousand single nucleotide polymorphism (SNP) loci that are associated with various complex human diseases or traits. The majority of GWAS discoveries are located in non-coding regions of the human genome and have unknown functions. The valley between non-coding GWAS discoveries and downstream affected genes hinders the investigation of complex disease mechanism and the utilization of human genetics for the improvement of clinical care. Meanwhile, advances in high-throughput sequencing technologies reveal important genomic regulatory roles that non-coding regions play in the transcriptional activities of genes. In this review, we focus on data integrative bioinformatics methods that combine GWAS with functional genomics knowledge to identify genetically regulated genes. We categorize and describe two types of data integrative methods. First, we describe fine-mapping methods. Fine-mapping is an exploratory approach that calibrates likely causal variants underneath GWAS signals. Fine-mapping methods connect GWAS signals to potentially causal genes through statistical methods and/or functional annotations. Second, we discuss gene-prioritization methods. These are hypothesis generating approaches that evaluate whether genetic variants regulate genes via certain genetic regulatory mechanisms to influence complex traits, including colocalization, mendelian randomization, and the transcriptome-wide association study (TWAS). TWAS is a gene-based association approach that investigates associations between genetically regulated gene expression and complex diseases or traits. TWAS has gained popularity over the years due to its ability to reduce multiple testing burden in comparison to other variant-based analytic approaches. Multiple types of TWAS methods have been developed with varied methodological designs and biological hypotheses over the past 5 years. We dive into discussions of how TWAS methods differ in many aspects and the challenges that different TWAS methods face. Overall, TWAS is a powerful tool for identifying

complex trait-associated genes. With the advent of single-cell sequencing, chromosome conformation capture, gene editing technologies, and multiplexing reporter assays, we are expecting a more comprehensive understanding of genomic regulation and genetically regulated genes underlying complex human diseases and traits in the future.

Keywords: TWAS, GWAS, summary statistics, data integration, eQTL, functional annotation, fine-mapping

INTRODUCTION

For the last two decades, genome-wide association studies (GWAS) have been a successful approach for associating single nucleotide polymorphism (SNP) loci to a variety of complex human traits. In fact, as of July 2021, the NHGRI-EBI GWAS catalog includes more than 167,000 SNPs associated with human diseases and traits (Buniello et al., 2019). The abundant discoveries of SNP associations with complex human diseases have led to significant enthusiasm and growth in interdisciplinary, translational medicine studies. Translational medicine aims to translate genomic discoveries of complex human diseases to clinical settings to achieve precision medicine (Collins and Varmus, 2015) and to improve the overall quality of health care. The expedition from bench to bedside investigates genetically determined disease susceptibility and inter-individual variability in treatment response to develop genomics-informed diagnosis and prognosis tools as well as individually tailored treatment plans. However, the majority (∼90%) of statistically significant GWAS signals are located in non-coding regions of the human genome (Maurano et al., 2012). Thus, connecting these non-coding variants to downstream affected genes is a nontrivial task. The gap between non-coding GWAS signals and affected genes hinders the translation of GWAS discoveries to clinical settings.

Increased volume and improved precision of omics data, newly invented molecular technologies, and recently developed bioinformatics algorithms, together reveal novel avenues in translational medicine to walk from GWAS signals to downstream affected genes. Non-coding regions of the human genome, including intergenic and intronic regions, can act as regulatory elements that have effects on transcriptional or translational activities of genes. Several classes of widely-studied functional elements include enhancers, promoters, transcription factor binding sites (TFBS), CCCTC-binding factor (CTCF); and these functional elements can host genetic variants, like expression quantitative trait loci (eQTLs), splicing quantitative trait loci (sQTLs), and protein quantitative trait loci (pQTLs), which participate in various transcriptional and translational regulatory mechanisms [Visel et al., 2007; The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014; Andersson et al., 2014; Roadmap Epigenomics Consortium et al., 2015; Sun et al., 2018; ENCODE Project Consortium et al., 2020; GTEx Consortium, 2020]. Each class of functional element describes a type of regulatory mechanism by which genetic variants may modulate genes. The goals of many developed bioinformatics methods in the post-GWAS era are to identify genetically regulated genes from GWAS discoveries by integrating functional genomics knowledge. Transcriptome-wide association studies (TWAS) are one type of data integrative bioinformatics method that aims to identify genes that lead to manifestation of complex human traits due to genetically regulated transcriptional activity.

Transcriptome-wide association studies has gained popularity over the years due to its distinct ability to perform gene-level association analyses and generate interpretable transcription hypotheses between genes and complex diseases and traits. Here, we first review updates in functional genomics. We also summarize bioinformatics methods that embrace functional genomics data to identify complex trait-associated genes. Then, we dive into the specifics of TWAS and assess the pros and cons of several developed TWAS methods. Next, we discuss several influential factors in the experimental design of TWAS that may potentially sway interpretation of results. Finally, we review challenges for TWAS and opportunities to maximize the utility of TWAS in the future.

OVERVIEW

The technological advances to identify genomic regulation provide opportunities to prioritize genetically regulated genes from GWAS signals from new perspectives. Fine-mapping of GWAS causal signals has relied heavily on linkage disequilibrium (LD). A common practice following GWAS is to map genetic variants to the residing genes, or nearby genes based on haplotypes and LD structures derived from the study cohort or from a fully sequenced reference panel of presumably similar ancestry [such as an ancestrally similar subset of the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015)]. This approach has led to identification of complex disease and trait-associated loci, but does not recognize the widespread, complex transcriptional regulatory mechanisms which do not necessarily take place in genes' proximity (Heidari et al., 2014; Javierre et al., 2016; Pan et al., 2018).

Genetic variants, regardless of their chromosome locations relevant to genes, can modulate transcriptional activities of target genes up to several mega base pairs (Mbp) away if located in regulatory elements, such as enhancers and transcriptional factor (TF) binding sites, or having suggestive effects on genes, like expression quantitative trait loci (eQTLs) (Javierre et al., 2016; GTEx Consortium, 2020). The distal genomic regulations are accomplished via formations of chromatin loops. As more knowledge about three-dimensional (3D) genome structure becomes available through chromosome conformation capture (3C) technology and its derivatives (Davies et al., 2017), it becomes well-recognized that chromatin looping plays an

important role in controlling transcriptional activities (Dixon et al., 2012; Rao et al., 2014). Chromatin looping allows distal regulatory elements to skip intervening genes to contact distant target genes. For example, using the 3C-carbon copy (5C) approach, Sanyal et al. (2012) observed that only ∼7% of chromatin looping interactions took place between an element (putative enhancers, promotors or CTCF binding sites) and the nearest transcription start site (TSS) in the pilot regions that represented 1% of the human genome in GM12878, K562, and HeLa-S3 cell lines (ENCODE Project Consortium, 2012). Even though Sanyal et al. (2012) inspected only a small proportion of the human genome, the frequency of distal regulatory interactions is profound. Proximity to genes or short-range *cis*-LD structures may not be sufficient tools to pinpoint causal genes of complex traits and diseases given the continuously updating knowledge of genomic regulation. Integration of genetic regulatory knowledge with GWAS results has become necessary to capture the complexity of biological regulatory mechanisms and prioritize genes from GWAS signals. **Figure 1** provides an overview of some of the strategies for post-GWAS gene-mapping procedures.

As of today, there have been various statistical and computational methods that incorporate functional genomics data to unveil complex trait-related genes. In this review, we categorize these methods into two types. First we describe the fine-mapping approach. Second we discuss the gene-prioritization approach.

Fine-Mapping for Post-GWAS Analysis

Fine-mapping is one common option for post-GWAS analyses seeking to identify causal variants or genes for complex diseases or traits (Schaid et al., 2018; Broekema et al., 2020). Traditionally, fine-mapping of potential causal variants relies heavily on LD structures and haplotypes blocks based on the premise that causal variants and tag variants have a non-random chance to be inherited together due to co-segregation during meiotic recombination (**Table 1**). Recently, there have also been multiple studies on alternative functional fine-mapping strategies that aim to identify potential causal functional elements, instead of a single variant, tagged by GWAS signals. These functional fine-mapping studies investigate downstream affected genes by understanding the likely impacted biological regulatory mechanisms. This shift of focus in GWAS fine-mapping is transformative for studies which are perplexed by non-coding GWAS signals and their connections to downstream affected genes (**Table 1**).

Fine-mapped GWAS signals may occur outside of coding regions and be situated in a distant non-coding functional element. Identification of non-coding causal functional elements is imperative for understanding the functional roles of GWAS variants. Examples of non-coding functional roles are enhancers, promoters, TF binding sites, candidate *cis*-regulatory elements (ccREs), and DNaseI hypersensitive sites. The identification of functional elements underlying GWAS pave the way to engage chromosome conformation information to locate the downstream target genes interacting with the functional regions of interest. The Washington Epigenome Browser (Zhou et al., 2011; Li et al., 2019) and 3D genome browser (Wang Y. et al., 2018) host several different kinds of cell line-specific or tissue-specific 3C, 5C, Hi-C, or capture Hi-C data. Both browsers provide necessary visualization tools to inspect the 3D chromatin loop-aided interactions for genomic regions of interest. FUMA developed by Watanabe et al. (2017) is another data integrative computational tool to assist functional annotation of fine-mapped GWAS variants and functional regions. Watanabe et al. (2017) assembles positional, eQTL, and chromosome confirmation mappings in FUMA. FUMA offers interactive visual aids for post-GWAS functional annotation and prioritization of potential complex trait-related genes based on multiple types of functional genomics data.

Table 1 lists exemplary methods of two major types of fine-mapping approaches. The statistical mapping focuses on the statistical approaches and models. The functional mapping focuses on the varied ways of using different functional genomic data for fine-mapping purposes. These two types of fine-mapping approaches are not mutually exclusive. A fine-mapping method can also fall into both categories depending on the method or study design. To summarize, fine-mapping methods integrate various types of omics data to deduct possible variant-gene relationships and biological mechanisms underpinning complex diseases or traits.

Gene-Prioritization for Post-GWAS Analysis

The capability of high-throughput sequencing technologies to quantify intermediate molecular traits, such as gene expression levels and protein abundance, enables the estimation of statistical significance of molecular mechanisms behind complex diseases and traits. Here, we discuss three different types of gene-prioritization methods that to evaluate how genetic variants can modify complex disease risk by exerting effects on an intermediate molecular trait.

One such integrative gene-prioritization method is colocalization (**Table 1**; Hukku et al., 2021). In general, colocalization analyzes the co-occurring patterns between QTLs (for example, eQTLs) and GWAS signals. Colocalization assesses the biological hypothesis of whether a causal locus or a genetic variant contribute to both the intermediate molecular changes and the complex trait of interest. A GWAS signal that is colocalized with a QTL is more likely to be functional. Colocalization analyses can be performed at a locus level or at a SNP level.

The locus-level colocalization methods assume that a group of SNPs in a tight LD region contain both a causal eQTL and a causal disease GWAS signal (**Table 1**). One will observe no marginal effect of a causal eQTL by conditioning on the most significant disease GWAS signal, and vice versa (Nica et al., 2010). An alternative method states that one will observe a maximum joint likelihood of associations if the two traits of interest are driven by the same causal variant (Chun et al., 2017).

The SNP-level colocalization methods focus on quantifying the probability of colocalization signals of two distinct traits surrounding a suspected causal variant (hence, at the single SNP/variant resolution) (**Table 1**). Several

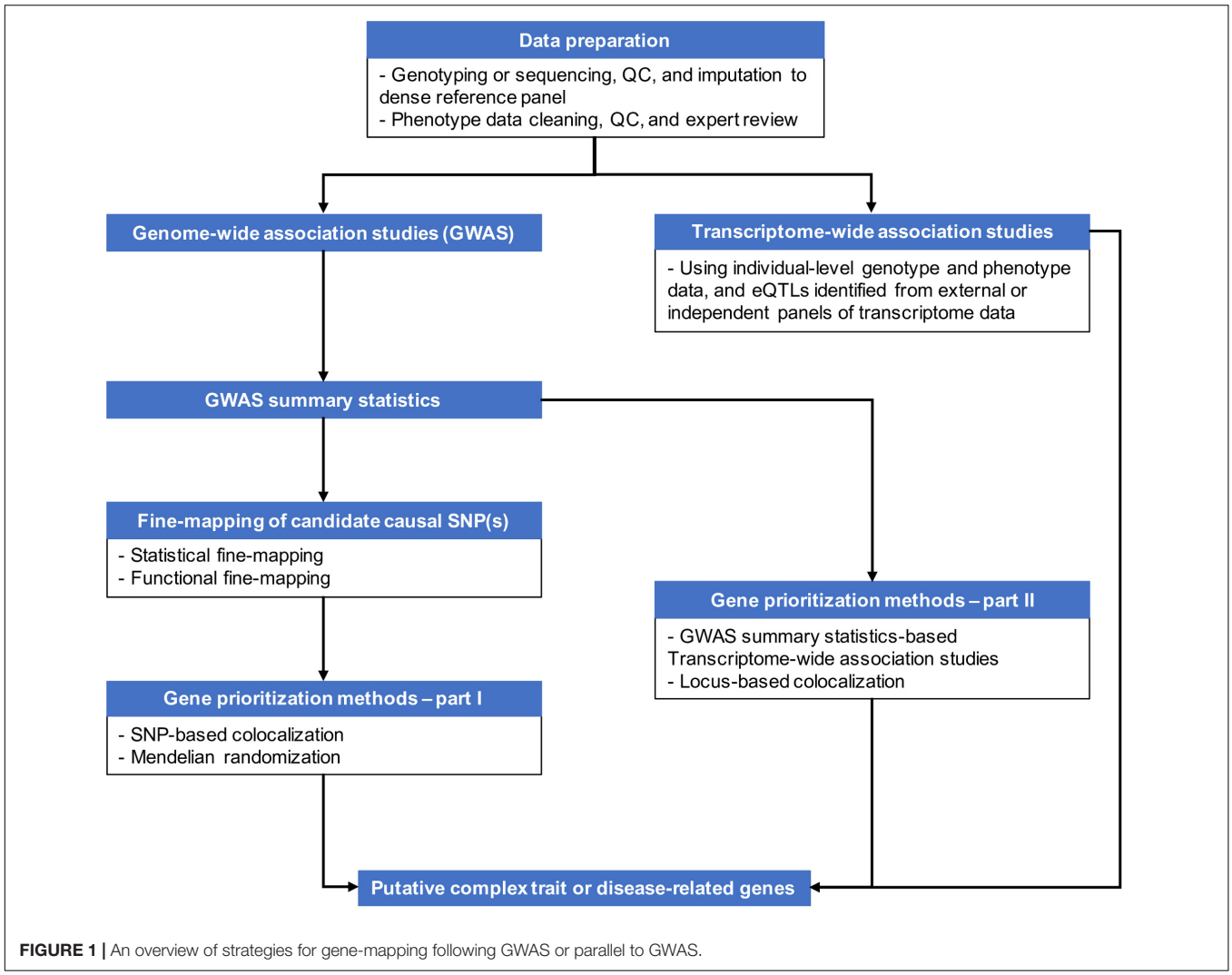


FIGURE 1 | An overview of strategies for gene-mapping following GWAS or parallel to GWAS.

TABLE 1 | Toolbox of gene-mapping methods and gene-prioritization methods (see **Table 2** for TWAS).

Gene mapping approaches		Type	Examples
Fine-mapping	Statistical mapping[a]	Heuristic approaches	Haploview Barrett et al., 2005, LocusZoom Pruim et al., 2010
		Penalized regression	LASSO Tibshirani, 1996, Elastic Net Zou and Hastie, 2005
		Bayesian methods	CAVIAR Hormozdiari et al., 2014, PAINTOR Kichaev et al., 2014
	Functional mapping[b]	Integrative annotation tools to infer functions	VEP McLaren et al., 2016, ANNOVAR Wang et al., 2010, HaploReg Ward and Kellis, 2011; Ward and Kellis, 2016, RegulomeDB Boyle et al., 2012, ENCODE SCREEN ENCODE Project Consortium et al., 2020, INFERNO Amlie-Wolf et al., 2018
		Visual annotation tools of 3D genome interactions	3D genome browser Wang Y. et al., 2018, WashU genome browser Zhou et al., 2011; Li et al., 2019, FUMA Watanabe et al., 2017
Colocalization[c]	Locus-level	RTC Nica et al., 2010, JLIM Chun et al., 2017	
	Variant-level	eCAVIAR Hormozdiari et al., 2016, coloc Giambartolomei et al., 2014, ENLOC Wen et al., 2017	
Mendelian randomization[d]		SMR Zhu et al., 2016, MR-JTI Zhou et al., 2020	

[a] *For detailed review of statistical fine-mapping, see Schaid et al. (2018);* [b] *For detailed review of functional fine-mapping, see Broekema et al. (2020);* [c] *See Hukku et al. (2021) for a detailed review of colocalization methods;* [d] *See Davies et al. (2018) for practical guidelines for clinical implementation of MR.*

exemplary SNP-level colocalization methods include eCAVIAR (Hormozdiari et al., 2016), COLOC (Giambartolomei et al., 2014), ENLOC (Wen et al., 2017), and fastENLOC (Pividori et al., 2020).

Mendelian Randomization (MR) is another approach, which makes causal inference between a modifiable exposure and complex disease risk (Holmes et al., 2017). The modifiable exposure can be blood concentrations of low-density lipoprotein cholesterol (LDL-c). The complex disease can be coronary heart disease (CHD). LDL-c related genetic variants are used in the process as instrumental variables to estimate the causal effects of LDL-c on CHD risk. One rising MR approach harnesses eQTLs to investigate whether one or more genetic variants influence both gene expression and a complex trait at the same time. This approach estimates, for example, if a *PCSK9* eQTL regulates *PCSK9* gene expression levels to impact blood LDL-c levels (Taylor et al., 2019; Richardson et al., 2020). eQTL-instrumented MR analyses are an innovative means to investigate LDL-related genes, which may further contribute to CHD risk. However, the success and accurate interpretation of MR results depend on three key assumptions (Holmes et al., 2017; Davies et al., 2018). Following the *PCSK9* eQTL and LDL-c example: (1) the genetic variant must be associated with gene expression levels; (2) there cannot be unmeasured confounding effects between the genetic variant and LDL-c; and (3) the genetic variant affects LDL-c only through their effects on gene expression levels.

Transcriptome-wide association study is a gene-based association approach first developed by Gamazon et al. (2015). TWAS integrates GWAS data with eQTL information to identify transcriptionally regulated genes underlying complex traits and diseases. TWAS first imputes the genetically regulated gene expression levels by combining individual-level genotype data or GWAS summary statistics with externally estimated eQTLs. At the second step, TWAS assesses the associations between imputed gene expression levels and a complex trait or disease (see section "Introduction to TWAS").

Transcriptome-wide association studies and mendelian randomization are similar in the way that TWAS is equivalent to a two-stage weighted allele score-based MR. The first stage estimates the aggregate effect of multiple instrumental variables on the exposure (for example, eQTLs' aggregate effect on a gene). The second stage regresses the outcome on the fitted values of the exposure from the first stage (for example, regression of continuous or categorial disease-related phenotype on the predicted genetically regulated gene expression levels). More interdisciplinary details can be found in Burgess et al. (2017); Burgess and Thompson (2013), and Pierce and Burgess (2013). The rest of this review focuses on the statistical aspects of TWAS as a gene-based association approach.

Transcriptome-wide association studies have attracted much interest in the field of complex disease due to its ability to perform gene-level association testing. This feature distinguishes TWAS from variant-based analytic approaches, such as some of the aforementioned fine-mapping, colocalization, or MR. These variant-based analytic approaches rely greatly on GWAS ability to identify complex trait or disease-related genetic variants. However, detecting variants with small to moderate effects requires considerable sample sizes in order to reach satisfactory statistical power (McCarthy et al., 2008; Manolio et al., 2009). TWAS overcomes this issue by aggregating regulatory effects of multiple eQTLs and directly testing associations between genes and diseases. Moreover, TWAS has a substantially smaller multiple testing burden by performing gene-level tests in comparison with variant-based analyses. Furthermore, TWAS is a flexible bioinformatics tool. TWAS can be used as an accessory to GWAS to support GWAS discoveries; or independently from GWAS (**Figure 1**). Some studies include TWAS as a parallel approach to their GWAS to identify putative causal genes associated with complex disease risk. The following sections focus on the variations of TWAS methods and the influential factors of TWAS studies.

TRANSCRIPTOME-WIDE ASSOCIATION STUDIES (TWAS)

Introduction to TWAS

Transcriptome-wide association studies can be considered a subclass of locus-based methods or multi-marker association approaches that are an alternative to variant-based association methods. The growth of locus-based methods is attributable to the wider recognition and appreciation of the polygenic architecture of complex diseases and traits. In other words, the proportion of disease phenotype variation explained by each genetic variant, on average, is small. Nevertheless, the cumulative effect of genetic variants in many genes, collectively, account for a substantial proportion of inter-individual phenotypic variation. Methodologically, locus-based methods take multiple genetic variants' effects into account to assess the overall contribution of a gene or a genetic region (a more interpretable functional unit in comparison to non-coding variants) to complex disease susceptibility. Meanwhile, advances in high-throughput sequencing technologies have enriched the discovery that genetic variants are tightly involved in regulation of transcription and translation of genetic material. eQTLs are one type of important regulatory variants. Recently, the detection of eQTLs has been aided by even lower cost RNA sequencing (RNA-seq) technology, sophisticated statistical models, increasing computational power, and scientific community efforts to consolidate eQTL research resources.

Similar to the shift in the GWAS field from variants with large effect sizes to variants with moderate to small effect sizes by involving greater sample sizes, eQTL research has gone through the same trend. eQTLs with large effect have elucidated molecular mechanisms behind a variety of complex diseases. For example, a promoter eQTL has a dominant genetic effect on *DARC*, a gene expressing malaria parasite receptor. The specific form of the eQTL interrupts *GATA*-1 binding sites and diminishes *DARC* gene expression in specific erythroid cells, which explains malaria resistance found in a certain West African population (Tournamille et al., 1995). Examples like the *DARC* promoter eQTL with a silencing effect are not common. The ability of the community to assemble even larger study cohorts allows for the observations of additional eQTLs, albeit with smaller effect sizes,

and in a diverse pool of tissue types other than blood. TWAS adopts this polygenic view including multiple small effect eQTLs for exploring the genetic architecture of complex disease risk.

Transcriptome-wide association studies exploit the genotype and phenotype data from GWAS along with reference transcriptome data to conduct gene-level association testing (Gamazon et al., 2015; Gusev et al., 2016; Barbeira et al., 2018, 2019; Hu et al., 2019; Pividori et al., 2020). TWAS tests the hypothesis that one or multiple eQTLs collectively regulate the transcriptional activities of a gene, and the genetically altered gene expression levels result in modulated disease risk.

Provided individual-level genome-wide genotype data, TWAS performs a two-step analysis to test this transcriptional hypothesis. For any given gene, step 1 imputes genetically regulated gene expression levels by combining transcriptional regulatory effects of the eQTLs for a gene under an additive genetic model. Step 1 can be done in multiple tissues of interest separately in each tissue or jointly across tissues (see section about "eQTL Detection"). Various eQTL models are available for step 1 thanks to the efforts of consortia, like GTEx (GTEx Consortium, 2020), BLUEPRINT (Chen et al., 2016), eQTLGEN (Võsa et al., 2018), and MESA (Mogil et al., 2018). Let N denote the sample size of a study cohort and M denote the number of eQTLs in a certain gene. Prediction of the gene's genetically regulated gene expression levels can be expressed as follows:

$$E = X\hat{W} \tag{1}$$

where E is the $N\times 1$ vector of predicted genetically regulated gene expression levels of the gene, X is the $N\times M$ matrix of genotypes of eQTLs, and $\hat{W}$ is the $M\times 1$ vector of eQTLs' regulatory effects on the gene, which are estimated from an independent reference transcriptome data panel. While the first step of TWAS is merely to capture genetic components of gene expression levels, TWAS has shown to have a good prediction accuracy for genes that are highly locally heritable ($h^2 \geq 0.5$) (Gamazon et al., 2015; Li et al., 2018).

The second step is to aggregate the imputed gene expression levels from step 1 with a disease phenotype of interest to estimate the statistical significance of each gene-disease association. Let Y denote the phenotype of a study cohort. Y is the $N\times 1$ vector of phenotype, which can be dichotomous, such as case/control status of a complex disease, or continuous measures of health outcomes, such as blood laboratory values. Step 2 calculates the regression coefficient of the phenotype Y on each genes' predicted gene expression levels E, Given its design, TWAS conducts genomic association analyses with an innate transcriptional regulatory hypothesis.

Transcriptome-wide association studies have several advantages over traditional variant-based genomic analyses. First, TWAS is a gene-based analytic approach that has the potential to extend GWAS toward a functional understanding of disease mechanisms. Second, the two analytic steps in TWAS are decoupled and can be conducted independently. For multi-trait or phenome-wide studies, the first step of predicting gene expression levels only needs to be performed once for a given dataset. Predicted genetically regulated gene expression levels can be then evaluated for statistical association with different disease phenotypes or complex traits at step 2. Meanwhile, the technical independence of step 2 gives ample research opportunities for the development of sophisticated statistical models for gene-disease association analyses. Third, multiple testing burden is lower in TWAS in comparison to a genome-wide variant-based test; here, one only needs to adjust for the number of genes tested in the TWAS. For a given trait, a TWAS only needs to adjust for approximately twenty-thousand genes (this is a Bonferroni p-value threshold of approximately 2.5×10^{-6}). Meanwhile, the number of statistical tests goes up to millions for a GWAS. As such, the multiple testing burden is orders of magnitude heavier in GWAS than in TWAS. The lower multiple testing burden allowed Thériault et al. (2018) to identify the association between *PALMD* and calcific aortic valve stenosis (CAVS) in the QUEBEC-CAVS cohort with a sample size of N = 2,000). The *PALMD*-CAVS association was successfully replicated in the much larger UK Biobank CAVS GWAS (N = 353,000). However, the same association was not statistically significant in the QUEBEC-CAVS GWAS due to the great multiple testing burden relative to the limited GWAS sample size (QUEBEC-CAVS N = 2,000). Fourth, TWAS are tissue-specific. TWAS has the capability to predict tissue-specific genetically regulated gene expression levels and investigate gene-trait associations in disease-related or potentially pathological tissues.

A TWAS study is subject to several influential factors which merit cautious interpretations of results (**Table 2**). These influential factors include: (1) the nature of input GWAS data, in other words, individual-level genotype and phenotype data versus GWAS summary statistics, (2) the eQTL models, and (3) the association method used to estimate gene-trait associations. In the following sections, we expand on each of these factors.

Individual-Level Data-Based TWAS Versus GWAS Summary Statistics-Based TWAS

Transcriptome-wide association studies can take different forms of input data types. The first published TWAS method, PrediXcan, developed by Gamazon et al. (2015), accepts individual-level major variant dosages of eQTLs or genotype calls as input. However, individual-level genotype data are not easily obtainable from published GWAS studies for a TWAS follow-up study. As a solution and an alternative TWAS method, FUSION, developed by Gusev et al. (2016), quickly followed the release of PrediXcan. FUSION imputes the regression statistics between the gene expression level of each gene and a trait (hereafter denoted as z_g) directly from GWAS summary statistics. Let Z denote a vector of standardized SNP-trait effect sizes (z-scores) from a GWAS and only include GWAS SNPs that are also eQTLs in a given eQTL-gene expression model; and Σ denote the covariance matrix among all eQTLs (LD). In FUSION, z_g are imputed as a linear combination of elements of Z with weights $\hat{W}$. When there is no SNP-trait association (no signals), $Z\sim N(0, \Sigma)$ and therefore, z_g has a zero mean and variance $\hat{W}'\Sigma\hat{W}$. For a

TABLE 2 | Summary of TWAS methods.

Comparison	PrediXcan	S-PrediXcan	FUSION	MultiXcan	S-MultiXcan	UTMOST
Input GWAS data type	Individual-level genotype data	GWAS summary statistics	GWAS summary statistics	Individual-level genotype data	GWAS summary statistics	GWAS summary statistics
Statistical models for eQTL identifications	Elastic Net Fine-mapped MASHR-based models Joint-Tissue Imputation (JTI) models	Same as PrediXcan	Bayesian sparse linear mixed model (BSLMM)	Same as PrediXcan	Same as PrediXcan	Group LASSO with specialized regularization
Source reference panels	GTEx, MESA, CommonMind, StarNet, DGN, PsychENCODE	Same as PrediXcan	GTEx, TCGA	Same as PrediXcan	Same as PrediXcan	GTEx, StarNet, BLUEPRINT
eQTL Databases	http://predictdb.org/ https://zenodo.org/record/3842289#.YNVbJBOpGdY	http://predictdb.org/	http://gusevlab.org/projects/fusion/	http://predictdb.org/	http://predictdb.org/	https://github.com/Joker-Jerome/UTMOST
Current GTEx version[a]	GTEx v8	GTEx v8	GTEx v7	GTEx v8	GTEx v8	GTEx v6p
Gene-trait association methods	Linear or logistic regression	Dependent on GWAS method	Dependent on GWAS method	Principal component regression	Singular value decomposition (analogous to MultiXcan)	Generalized Berk-Jones test
Tissue-specificity	Tissue-specific	Tissue-specific	Tissue-specific	Cross-tissue	Cross-tissue	Cross-tissue
Output	Single-tissue gene-trait associations	Single-tissue gene-trait associations	Single-tissue gene-trait associations	Cross-tissue gene-trait associations	Cross-tissue gene-trait associations	Cross-tissue gene-trait associations
Pros	Up-to-date eQTL databases; Accurate representation of test cohort LD	Computationally efficient; Up-to-date eQTL databases;	Computationally efficient	Up-to-date eQTL databases;	Computationally efficient; Up-to-date eQTL databases;	Computationally efficient
Cons	Multiple testing burden; Computationally burdensome in comparison to summary-statistics based TWAS;	Reference LD matrix can introduce noises	Multiple testing burden; Reference LD matrix can introduce noises	Computationally burdensome;	Reference LD matrix can introduce noises	Reference LD matrix can introduce noises;
References using PMID	26258848, 32917697, 33020666	29739930	30926970	30668570	30668570	30804563

[a] *Dated August 2021.*

given gene, the effect of genetically regulated gene expression level on the phenotype can be obtained as follows in FUSION:

$$z_g = \frac{\hat{W}' Z}{\sqrt{\hat{W}' \Sigma \hat{W}}} \quad (2)$$

In comparison to individual-level data-based TWAS, GWAS summary statistics-based TWAS is more computationally efficient and has the ability to analyze a larger GWAS dataset as it is less central processing unit (CPU) and memory intensive. Various GWAS summary statistics-based TWAS methods have emerged since FUSION, including S-PrediXcan (Barbeira et al., 2018) and UTMOST (Hu et al., 2019) (**Table 2**).

The primary difference between TWAS that uses individual-level data and those that use GWAS summary statistics is in the estimation of LD structure for testing populations. The individual-level genotype data are usually not easily accessible from most published GWAS studies, making it difficult to examine the LD structure among eQTLs in each GWAS dataset. GWAS summary statistics-based TWAS circumvents this issue by deriving an LD matrix from a reference set, either the reference panel used for eQTL discovery, or a multi-ancestry, deeply sequenced reference panel like 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015). Nevertheless, seldom does a reference population panel perfectly resemble the population structure of a specific study cohort. The discrepancy between the reference LD matrix and the actual LD structure of a study cohort will likely introduce noise and may lead to false positive or false negative results in GWAS summary statistics-based TWAS, despite a general good concordance between individual-level and summary statistics-based TWAS (Barbeira et al., 2018). The silver lining is the increasing sample sizes in reference population panels for more accurate estimates of an LD structure, which matters for GWAS summary statistics-based studies (Benner et al., 2017).

Overall, individual-level TWAS provides more accurate estimates of gene-trait associations. However, it usually takes up significant computational resources; and individual-level genotype data are not always accessible to the research community. On the other hand, GWAS summary statistics-based TWAS is advantageous in its capability to prioritize genes using only GWAS summary statistics and also computation speeds that are orders of magnitude faster than individual-level TWAS. Nevertheless, as mentioned above GWAS summary statistics-based TWAS can introduce noise to association results as the commonly used reference LD matrix cannot perfectly resemble the LD structure of the study cohort. GWAS summary statistics-based TWAS will require a greater GWAS sample size to achieve satisfactory statistical power. Because of these limitations, GWAS summary statistics-based TWAS generally needs additional validation and careful interpretation.

eQTL Detection

The choice of eQTL database is important in TWAS (see "Statistical models for eQTL identifications" in **Table 2**). Quality of the eQTL databases impacts the prediction accuracy of gene expression levels. Transcriptome and genotype data of higher quality can capture greater proportions of the genetic components of gene expression regulation, identify eQTLs with moderate to small effect sizes, and improve the precision of eQTLs in complex gene regions that share the same locus control region or express multiple isoforms.

The power to detect eQTLs from transcriptome and genotype datasets is partially dependent on the sample size. Over the past decade, not only the sample sizes of reference transcriptome data, but also the diversity of human tissues and cell lines, have grown to support a deeper and broader understanding of genetic architecture of eQTLs. Better quality eQTL data in more diverse tissues have been made publicly available thanks to several consortia, including ScanDB (Gamazon et al., 2010), GTEx (GTEx Consortium, 2020), ImmVar (Ye et al., 2014), BLUEPRINT (Chen et al., 2016), CAGE (Lloyd-Jones et al., 2017), PsychENCODE (Wang D. et al., 2018), eQTLGen (Võsa et al., 2018). ScanDB is one of the earliest centralized eQTL databases that explores eQTLs in 176 HapMap Lymphoblastoid Cell Lines, made up by 87 CEPH from Utah (CEU) and 89 Yoruba from Ibadan (YRI) (Gamazon et al., 2010). Approximately five thousand eQTLs were discovered in the CEU and YRI, respectively, and are hosted on ScanDB website[1] (Duan et al., 2008). Following ScanDB, one of the most well-known eQTL studies is the Genotype-Tissue Expression (GTEx) project that was launched in 2010 (GTEx Consortium, 2015). The latest release version of GTEx (GTEx v8) extended the search of eQTLs in 838 donors (15,201 postmortem biospecimen) for 49 primary human tissues and two cell lines (GTEx Consortium, 2020). GTEx provides tissue-specific eQTLs and splicing quantitative trait loci (sQTLs) for 18,262 protein-coding and 5,006 long intergenic non-coding RNA (lincRNA) genes after biological and statistical quality control. GTEx brings the awareness of widespread eQTL effects that almost all protein coding genes and ∼67% of lincRNA genes have been detected to be under the influence of *cis*-eQTLs in at least one tissue. An even greater eQTL detection sample size than the GTEx project has been assembled through the effort of the eQTLGen consortium[2]. eQTLGen meta-analyzed 31,684 blood samples (majority of European ancestry) from 37 datasets whose gene expression levels were profiled by three gene expression arrays and one RNA-seq platform (Võsa et al., 2018). The magnitude of the sample size allows eQTLGen to identify not only *cis*-eQTLs (within 1 Mbp to a gene), but also *trans*-eQTLs that are more than 5 Mbp away from a gene or on another chromosome. A single-cell version of eQTLGen is expected to further unravel the transcriptional regulatory mechanism behind complex disease and traits in delicate individual immune cell types (van der Wijst et al., 2020).

Interpretation of eQTL effects and TWAS results should consider the fact that transcriptional regulation is a spatiotemporal process that can differ from tissue to tissue and between life and death. Ferreira et al. (2018) found that a proportion of genes displayed drastic transcript-level changes over the postmortem intervals due to postmortem ischemia, regulatory changes, and RNA degradation. Genes that are

[1] http://www.scandb.org/

[2] https://www.eqtlgen.org/

affected by postmortem gene regulation differ from tissue to tissue (Ferreira et al., 2018). While postmortem effects on transcriptome are still largely unknown, postmortem tissues, including blood samples, remain irreplaceable natural resources to explore tissue-specific molecular mechanisms of complex diseases. Given the transcriptional regulatory difference between life and death, it is important to validate the effects of eQTLs and transcriptional changes of genes in complex trait or disease-relevant biospecimens using RNA-seq or high-throughput massively parallel reporter assay (MPRA) (Tewhey et al., 2018).

Methods to detect eQTLs are developed based on different biological hypotheses and statistical models. eQTL detection methods can differ in two parts: (1) the assumptions of the genetic architecture of transcriptional regulation, and (2) adoption of a tissue-by-tissue analytic model versus a cross-tissue method design. Due to a wider acknowledgement of the polygenic genetic architecture of intermediate molecular traits (Zhang et al., 2011; King et al., 2014), eQTL studies have set off to detect multiple potential causal eQTLs at a genetic locus, as opposed to only a single eQTL at a locus as would be done in a monogenic model. For example, Gusev et al. (2016) used Bayesian Sparse Linear Mixed Model (BSLMM) (Zhou et al., 2013) to detect eQTLs that were later used to predict gene expression levels (**Table 2**). BSLMM fits all SNPs nearby a gene into the model and allows two types of genetic components, one sparse (i.e., a small set of eQTLs with large effect sizes) and one vastly polygenic (i.e., all SNPs at a locus having marginal effect sizes). BSLMM attained a better prediction performance than a prediction estimated by merely using the top eQTL at a locus (Gusev et al., 2016). This suggests a non-monogenic genetic architecture of gene expression regulation, which is further supported by another contemporary study by Gamazon et al. (2015) that compared the top SNP (monogenic), polygenic score (polygenic), and elastic net (polygenic). To further understand the sparsity of polygenic genetic architecture behind gene expression, Wheeler et al. (2016) evaluated the contribution of sparse and polygenic components for transcriptional regulations, using BSLMM (sparse and polygenic), LASSO (sparse) and elastic net/ridge (polygenic) regression models. They compared the genetic heritability of gene expression explained by each method to determine the local genetic contribution of eQTLs to gene expression variation. They found that *cis* gene expression regulation was dominated by a small number of genetic variants rather than a large collection of genetic variants of marginal effect sizes. The discovery by Wheeler et al. (2016) strongly suggests a non-monogenic, sparse genetic architecture of *cis* transcriptional regulation. However, research in this area is in general impeded by limited sample sizes of transcriptome data.

Cross-tissue meta-analyses of transcriptome data have gained greater attention due to their capability of overcoming the sample size constraint as seen in the tissue-by-tissue eQTL detection approaches (**Table 2**). Research of cross-tissue eQTL detection is fostered by the discovery that an obvious proportion of *cis*-eQTLs are shared across all tissues and have correlated effect sizes across tissues (Battle et al., 2017). Flutre et al. (2013) introduced a cross-tissue Bayesian model that allows a proportion of eQTLs being shared across tissues and accounts for intra-individual correlations among tissues. Their hierarchical model can estimate heterogeneous effects of eQTLs in different tissues and identify eQTL active tissues. A similar approach is Meta-Tissue by Sul et al. (2013) that adopts a linear mixed model, which specifically leverages the random effects model developed by Han and Eskin (2011), to achieve similar goals as the Flutre et al. (2013). More cross-tissue eQTL detection methods have followed over years, including work by Acharya et al. (2016), RECOV by Duong et al. (2017), a sparse group LASSO model embedded in UTMOST by Hu et al. (2019), and a Joint Tissue Imputation (JTI) approach by Zhou et al. (2020). In general, cross-tissue eQTL detection methods have shown greater power in simulation studies in comparison to tissue-by-tissue approaches and a substantial increase in the numbers of identified eQTLs and eGenes (Genes that are regulated by at least one statistically significant eQTLs) (Han and Eskin, 2011; Flutre et al., 2013; Sul et al., 2013; Acharya et al., 2016; Duong et al., 2017; Hu et al., 2019; Zhou et al., 2020) (see "Statistical models for eQTL identifications" in **Table 2**).

Variety of Gene-Trait Association Methods

In addition to eQTL discovery, integrative cross-tissue analyses flourish in the evaluation of TWAS gene-disease associations (**Table 2**). Earliest design of TWAS, i.e., PrediXcan, investigates gene-trait associations in a tissue-specific manner. Naturally, PrediXcan estimates the statistical significance of association between a disease of interest and predicted gene expression levels tissue-by-tissue. However, tissue-specific TWAS faces four issues. First, limited sample sizes of reference transcriptome data not only restrict statistical power to identify eQTLs, but also TWAS power. This can happen in a way where certain tissues do not have sufficient sample sizes and power to detect eQTLs for a functional gene. As a result, TWAS will not be able to predict the gene's expression levels, let alone test for gene-trait associations in an underpowered tissue. Second, causal tissues of many complex diseases or traits can be unclear or unavailable, making it difficult to determine specific tissues or cell lines on which one should conduct TWAS. Third, when causal tissues are unclear, one might choose to conduct an exploratory TWAS on multiple tissues. This kind of study design invites a substantial multiple testing burden. In an exploratory situation, one will need to correct TWAS association results for 49 primary human tissues or cell lines (available by GTEx), when perhaps only one or two tissues were causal to a complex disease. On the other hand, this test-all-tissue approach also carries an implicit assumption that TWAS will only assign statistical significance to tissues that are biologically relevant to the complex trait of interest. This assumption, however, can be easily violated due to the fourth issue. Fourth, cumulative evidence has suggested that there is shared local genetic architecture of gene expression regulation and similar *cis*-eQTL effect sizes across tissues (Battle et al., 2017; Liu et al., 2017; Ongen et al., 2017). The shared eQTL effects across tissues indicates that TWAS cannot distinguish disease-relevant tissues from irrelevant tissues that share similar gene expression levels from a statistical perspective (Wainberg et al., 2019). Cross-tissue TWAS is thus promoted to resolve some of

these issues with tissue-specific TWAS. Essentially, cross-tissue TWAS methods aggregate evidence across tissues to test the joint effect of gene expression levels on complex diseases or traits.

Different cross-tissue TWAS methods have been developed and provide various options for either individual-level genotype data or GWAS summary statistics (**Table 2**). MultiXcan by Barbeira et al. (2019) is a cross-tissue TWAS method provided within the MetaXcan method package. MultiXcan uses individual-level genotype data to predict gene expression levels in each single tissue and then fits the predictions across tissues against a phenotype in a statistical model to estimate the joint effect of a gene on a complex trait of interest. To avoid inflation of results due to correlated gene expression levels across tissues, MultiXcan adopts the principal component regression which specifically uses the first several orthogonal principal components of the predicted gene expression data matrix as explanatory variables. The GWAS summary statistics version of MultiXcan is called S-MulTiXcan (Barbeira et al., 2019). An alternative to S-MulTiXcan is a method called UTMOST developed by Hu et al. (2019) UTMOST uses a generalized Berk-Jones (GBJ) test which carries out a secondary test to examine if a gene is statistically significantly associated with a disease in at least one of the tested tissues. GBJ tests in UTMOST handles correlated gene expression levels across tissues by taking the covariance among single-tissue TWAS test statistics into account (Sun and Lin, 2020).

Cross-tissue TWAS has advantages and disadvantages in comparison to single-tissue TWAS. Cross-tissue TWAS methods have shown improved power to identify gene-level association in both simulated and natural data (Barbeira et al., 2019; Hu et al., 2019). Nevertheless, cross-tissue TWAS results are not tissue-specific and thus, cannot reveal tissue-specific genetic regulatory mechanisms. Computing resources and time required by cross-tissue TWAS methods are much higher than the corresponding single-tissue counterparts. Despite pros and cons, further validation, such as replication in independent datasets or functional validation, are needed by either single-tissue or cross-tissue TWAS.

Cross-tissue TWAS methods are not restricted to the eQTL models that come with the method. In general, a state-of-the-art eQTL method with better prediction accuracy of gene expression levels is preferred. In other words, cross-tissue TWAS methods such as MultiXcan, S-MulTiXcan (Barbeira et al., 2019) and UTMOST (Sun and Lin, 2020) can use the cross-tissue JTI-based eQTL models (Zhou et al., 2020) that is developed separately. The same principle applies to single-tissue TWAS methods. PrediXcan, S-PrediXcan and FUSION can use, for example, the cross-tissue JTI-based eQTL models which provides an improved prediction accuracy of gene expression levels (Gamazon et al., 2015; Gusev et al., 2016; Barbeira et al., 2018; Zhou et al., 2020).

CHALLENGES

While promising methods for disease gene discovery, TWAS faces several challenges. First, prediction accuracy of gene expression levels is limited by the heritability (h^2) of each gene. The heritability (h^2) of a gene's expression levels determines the upper bound of prediction accuracy by eQTLs. On the one hand, different studies have shown that TWAS can accurately predict the expression levels for genes that are highly locally heritable ($h^2 \geq 0.5$) (Gamazon et al., 2015; Li et al., 2018). And 59% of genes in the DGN whole blood have well estimated local h^2 (FDR < 0.1) (Wheeler et al., 2016). On the other hand, some genes have little to negligible estimated local heritability and should be removed from TWAS to avoid false positives. Nonetheless, much is still unclear about the heritability of gene expression levels across tissues and beyond *cis*-eQTLs.

Thus far, TWAS has only been using *cis*-eQTLs within a certain distance from genes. This is consistent with observations in several studies that the majority of *cis*-eQTLs cluster around the transcription start site of the target gene (Nica et al., 2011; GTEx Consortium, 2015). However, gene can be regulated by both *cis* and *trans*-regulatory elements in the human genome. Many studies seek to identify *trans*-eQTLs, which have been absent in gene expression heritability estimation due to technical limitations. Several previous studies estimated that ∼70% of the genetic heritability of gene expression levels could be attributable to *trans*-eQTLs that are on another chromosome or more than 5 Mb away (Boyle et al., 2017; Liu et al., 2019), indicating the importance of *trans*-eQTLs in transcriptional regulation. However, *trans*-eQTL studies face enormous multiple testing burden. Studies to identify *trans*-eQTLs will need to test all possible intra and inter-chromosome variant-gene pairs. The total number of statistical tests is orders of magnitude greater than that of *cis*-eQTLs, which only considers proximal variant-gene pairs. A great number of samples is thus needed for *trans*-eQTL research to guarantee sufficient statistical power (Westra et al., 2013). Even if *trans*-eQTL data are made available, as in blood-related cell lines by eQTLGen (Võsa et al., 2018), TWAS may still have difficulty utilizing *trans*-eQTLs due to two key factors. First is the possible overlapping effects between the *trans* and *cis*-eQTLs for a target gene. *Trans*-eQTLs likely regulate expression of a *trans*-acting TF, which subsequently functions by binding to a *cis*-regulatory element where a *cis*-eQTL resides (Võsa et al., 2018). Second is the difficulty of calculating LD among eQTLs. The computing time and resources needed for such a task are exponentially greater than that for *cis*-eQTLs.

Another challenge in TWAS is the lack of eQTL data from different ancestry groups, diseases, medical conditions, sex, etc. The majority of samples used for large-scale eQTL studies were of European ancestry. eQTL databases that were prepared by a few earlier TWAS methods were exclusively European ancestry individuals (Gamazon et al., 2015; Gusev et al., 2016; Barbeira et al., 2018). Ancestry-specific eQTL data are available for some ancestry groups, but these resources are generally limited. The Multi-Ethnic Study of Atherosclerosis (MESA) characterized eQTLs in African American (N = 233), Hispanic (N = 352), and European (N = 578) populations, separately (Mogil et al., 2018). However, the MESA genotype and RNA-seq data were collected from only CD14+ monocytes and individuals free of clinical cardiovascular diseases (CVD) at recruitment. Although, individuals with CVD and other medical conditions are likely

to experience different transcriptional regulation from their healthy peers. Overall, much is still to explore about the eQTLs in different ancestries, medical conditions, age, sex, etc. (Piasecka et al., 2018).

It is hard to quantify TWAS power due to the complexity of transcriptional regulation and varied genetic backgrounds of different complex diseases or traits (Veturi and Ritchie, 2018; Li et al., 2021). For example, TWAS power can be influenced by the quality of gene expression prediction (sample sizes used for eQTL detection, concordance between transcriptome reference population and testing populations, coverage of eQTLs in the test dataset, etc.), or genetic factors (e.g., genetic heritability of gene expression levels, heritability of the phenotype, sample size, MAF, etc.). On top of the aforementioned factors, TWAS is also challenged by the fact that causal tissues or cell types are unclear in the majority of complex diseases or traits. Overall, TWAS statistical power is contingent on so many varied factors that it is hard to estimate TWAS power without making a delicate set of assumptions; and one should be careful when interpreting TWAS power.

Transcriptome-wide association studies need fine-mapping. Statistically significant TWAS results indicate only association, but not causation. Statistically significant genes are likely tag genes for other causal genes in its proximity, but achieve the greatest statistical significance due to various reasons (Wainberg et al., 2019). One solution is to fine-map causal genes by leveraging the LD structure among genes. For example, the method FOCUS estimates a set of credible genes that are tagged by a statistically significant gene by analyzing the patterns of eQTLs, GWAS signals and surrounding LD structure (Mancuso et al., 2019). One will have certain degree of statistical confidence (90 or 95% by choice) that causal genes are within the set of credible genes. The fine-mapping capability of FOCUS was supported by its success in recovering *SORT1* gene as one of the LDL risk genes. More work is expected in this field of research (Mancuso et al., 2019; Wu and Pan, 2020).

FUTURE DIRECTIONS

Understanding the genetic architecture of complex diseases and traits is still an ongoing task for the field of translational medicine. The journey from bench science to bed-side care requires the knowledge of causal genes, pathways, and mechanisms behind complex traits. The cumulative number of non-coding GWAS discoveries, time and again, stresses the need to fill the gap between non-coding genetic variants and downstream affected genes in order to uncover complex trait mechanisms. In this review, we categorize two types of methods that integrate GWAS with functional genomics data to bridge the variant-to-gene gap – fine-mapping approaches and gene prioritization approaches. We discuss the background, pros and cons of several classes of developed TWAS methods, influential factors in TWAS analyses, and challenges.

We expect greater endeavors in TWAS and functional genomic studies for a variety of geographical ancestry groups in the next 10 years, including but not limited to African, Asian, Hispanic or Latin, Greater Middle Eastern, Native American, Oceanian, and admixed populations (Lavange et al., 2010; H3Africa Consortium et al., 2014; Kowalski et al., 2019; Choudhury et al., 2020; Gay et al., 2020; Shang et al., 2020). Generation of these eQTL data will require resources and efforts from the research communities in different parts of the world.

High-throughput next-generation sequencing technology and array-based platforms will continue to generate informative functional genomics data. Ripening 3C and 3C-derived technologies will generate more knowledge about chromatin loop-assisted *cis* and *trans* regulatory interactions. Increasing evidence suggests the prevalence of distal regulatory mechanisms that cannot be easily captured with local LD structure (Whalen and Pollard, 2019; GTEx Consortium, 2020). Mumbach et al. (2017) recently developed HiChIP that generates high-resolution contact maps for enhancer-promoter interactions in a human coronary artery disease-related (CAD-related) cell type. They found that ∼89% of the coronary artery disease-associated SNPs skipped at least one gene to reach predicted target genes. The extent to which distal transcriptional regulation occurs is still unknown in the majority of complex human diseases or traits. But genomic regulatory information will be useful to decipher functionality of non-coding variants and map non-coding variants to their downstream affected genes.

Another highly expected sequencing technology by the field of eQTL and TWAS studies is the single-cell RNA sequencing (scRNA-seq) (Tang et al., 2009). Bulk RNA-seq of a tissue sample is the most economical way to obtaining transcriptome data in a large scale, despite the fact that a tissue sample comprises more than one cell type. Different cell types undergo distinguished genetic regulation that makes up their specific cellular identities. A gene's expression levels in a tissue, thus, are likely to differ from a cell type to another cell type. scRNA-seq profiles cell type composition in a tissue at a refined resolution and allows exploration of transcriptome heterogeneity across cell types (Snijder and Pelkmans, 2011). Growing scRNA-seq data and analytic methods will pave a new avenue in eQTL research that performs eQTL studies in various cell types in a tissue (van der Wijst et al., 2018). This will improve precision and accuracy of eQTLs. On the other hand, having a grasp on which causal tissues or cell types are important for a given complex disease will be essential for developing a better understanding of disease mechanism and clinical treatment. scRNA-seq data promise greater statistical power to identify complex trait-relevant tissues or cell types by providing distinguishable transcriptome profiles among cell types (Ongen et al., 2017; Finucane et al., 2018). Several scientific consortia have initiated the effort in generating scRNA-seq data in large sample sizes and multiple tissues, including the Human Cell Atlas (Regev et al., 2017), Single-cell eQTLGen (van der Wijst et al., 2020) and the LifeTime consortium (Rajewsky et al., 2020). At this dawn of single-cell omics sequencing technology, sample sizes and diversity of tissues and cell types will likely continue to be limited.

Even though genes are considered functional and heritable units, there is a shortage of gene-centric functional annotation

models. Existing functional annotation models focus on generating regulatory hypotheses for non-coding variants on a variant-centric basis. For most genes, it is unclear how the gene is regulated by different genetic regulatory elements, despite the fact that an average of 3.9 distal elements interact with the transcription start site (TSS) of a gene (Sanyal et al., 2012). The shortage of gene-centric functional annotation models also prevents locus-based statistical methods from combining *cis* and *trans*-regulation. With the advances in sequencing technologies, we are expecting a better understanding of genomic regulation that incorporates *cis* and *trans*-regulation to investigate how dysregulation of a gene, as a functional unit, contributes to complex diseases or traits.

More than a decade into GWAS research of complex disease, the molecular mechanisms behind most complex diseases remains unclear due to the valley between non-coding GWAS signals and the downstream affected genes. The next two decades await more research that sheds new light on complex disease mechanisms to promote novel therapeutics and precision medicine.

AUTHOR CONTRIBUTIONS

BL conceived the idea, drafted, and revised the manuscript. MR extensively revised the manuscript. Both authors listed have made a substantial, direct and intellectual contributions to the work, and approved it for publication.

FUNDING

This project was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R01 AI077505.

ACKNOWLEDGMENTS

We thank Yoseph Barash, Milton Pividori, and William Bone from the University of Pennsylvania for scientific discussions and feedback.

REFERENCES

1000 Genomes Project Consortium., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Acharya, C. R., McCarthy, J. M., Owzar, K., and Allen, A. S. (2016). Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. *BMC Bioinformat.* 17:257–259. doi: 10.1186/s12859-016-1123-5

Amlie-Wolf, A., Tang, M., Mlynarski, E. E., Kuksa, P. P., Valladares, O., Katanic, Z., et al. (2018). INFERNO: inferring the molecular mechanisms of noncoding genetic variants. *Nucleic Acids Res.* 46, 8740–8753. doi: 10.1093/nar/gky686

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461. doi: 10.1038/nature12787

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Communicat.* 9, 1–20. doi: 10.1038/s41467-018-03621-1

Barbeira, A. N., Pividori, M., Zheng, J., Wheeler, H. E., Nicolae, D. L., and Im, H. K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* 15:e1007889. doi: 10.1371/journal.pgen.1007889

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

Battle, A., Brown, C. D., Engelhardt, B. E., and Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

Benner, C., Havulinna, A. S., Jarvelin, M.-R., Salomaa, V., Ripatti, S., and Pirinen, M. (2017). Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *Am. J. Human Genet.* 101, 539–551. doi: 10.1016/j.ajhg.2017.08.012

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797. doi: 10.1101/gr.137323.112

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 169, 1177–1186. doi: 10.1016/j.cell.2017.05.038

Broekema, R. V., Bakker, O. B., and Jonkers, I. H. (2020). A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* 10:190221. doi: 10.1098/rsob.190221

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi: 10.1093/nar/gky1120

Burgess, S., and Thompson, S. G. (2013). Use of allele scores as instrumental variables for Mendelian randomization. *Int. J. Epidemiol.* 42, 1134–1144. doi: 10.1093/ije/dyt093

Burgess, S., Small, D. S., and Thompson, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Statist. Methods Medical Res.* 26, 2333–2355. doi: 10.1177/0962280215597579

Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., et al. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* 167, 1398.e–1414.e. doi: 10.1016/j.cell.2016.10.026

Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., et al. (2020). High-depth African genomes inform human migration and health. *Nature* 586, 741–748. doi: 10.1038/s41586-020-2859-7

Chun, S., Casparino, A., Patsopoulos, N. A., Croteau-Chonka, D. C., Raby, B. A., De Jager, P. L., et al. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* 49, 600–605. doi: 10.1038/ng.3795

Collins, F. S., and Varmus, H. (2015). A new initiative on precision medicine. *N. Engl. J. Med.* 372, 793–795. doi: 10.1056/NEJMp1500523

Davies, J. O. J., Oudelaar, A. M., Higgs, D. R., and Hughes, J. R. (2017). How best to identify chromosomal interactions: a comparison of approaches. *Nat. Methods* 14, 125–134. doi: 10.1038/nmeth.4146

Davies, N. M., Holmes, M. V., and Davey Smith, G. (2018). Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* 362:k601. doi: 10.1136/bmj.k601

The FANTOM Consortium and the RIKEN PMI and CLST (DGT). (2014). A promoter-level mammalian expression atlas. *Nature* 507, 462–470. doi: 10.1038/nature13182

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380. doi: 10.1038/nature11082

Duan, S., Huang, R. S., Zhang, W., Bleibel, W. K., Roe, C. A., Clark, T. A., et al. (2008). Genetic architecture of transcript-level variation in humans. *Am. J. Human Genet.* 82, 1101–1113. doi: 10.1016/j.ajhg.2008.03.006

Duong, D., Gai, L., Snir, S., Kang, E. Y., Han, B., Sul, J. H., et al. (2017). Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. *Bioinformatics* 33, i67–i74. doi: 10.1093/bioinformatics/btx227

ENCODE Project Consortium, Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583, 699–710. doi: 10.1038/s41586-020-2493-4

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Ferreira, P. G., Muñoz-Aguirre, M., Reverter, F., Sá Godinho, C. P., Sousa, A., Amadoz, A., et al. (2018). The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat. Communicat.* 9, 490–415. doi: 10.1038/s41467-017-02772-x

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* 50, 621–629. doi: 10.1038/s41588-018-0081-4

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* 9:e1003486. doi: 10.1371/journal.pgen.1003486

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098. doi: 10.1038/ng.3367

Gamazon, E. R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E. O., Nicolae, D. L., et al. (2010). SCAN: SNP and copy number annotation. *Bioinformatics* 26, 259–262. doi: 10.1093/bioinformatics/btp644

Gay, N. R., Gloudemans, M., Antonio, M. L., Abell, N. S., Balliu, B., Park, Y., et al. (2020). Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol.* 21, 233–220. doi: 10.1186/s13059-020-02113-0

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., et al. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 10:e1004383–e1004315. doi: 10.1371/journal.pgen.1004383

GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330. doi: 10.1126/science.aaz1776

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252. doi: 10.1038/ng.3506

H3Africa Consortium, Rotimi, C., Abayomi, A., Abimiku, A., Adabayeri, V. M., Adebamowo, C., et al. (2014). Research capacity. Enabling the genomic revolution in Africa. *Science* 344, 1346–1348. doi: 10.1126/science.1251546

Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Human Genet.* 88, 586–598. doi: 10.1016/j.ajhg.2011.04.014

Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., et al. (2014). Genome-wide map of regulatory interactions in the human genome. *Genome Res.* 24, 1905–1917. doi: 10.1101/gr.176586.114

Holmes, M. V., Ala-Korpela, M., and Smith, G. D. (2017). Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat. Rev. Cardiol.* 14, 577–590. doi: 10.1038/nrcardio.2017.78

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* 198, 497–508. doi: 10.1534/genetics.114.167908

Hormozdiari, F., van de Bunt, M., Segre, A. V., Li, X., Joo, J. W. J., Bilow, M., et al. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* 99, 1245–1260. doi: 10.1016/j.ajhg.2016.10.003

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* 51, 568–576. doi: 10.1038/s41588-019-0345-7

Hukku, A., Pividori, M., Luca, F., Pique-Regi, R., Im, H. K., and Wen, X. (2021). Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am. J. Human Genet.* 108, 25–35. doi: 10.1016/j.ajhg.2020.11.012

Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* 167, 1369.e–1384.e. doi: 10.1016/j.cell.2016.09.037

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., et al. (2014). Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genet.* 10:e1004722–e1004716. doi: 10.1371/journal.pgen.1004722

King, E. G., Sanderson, B. J., McNeil, C. L., Long, A. D., and Macdonald, S. J. (2014). Genetic dissection of the Drosophila melanogaster female head transcriptome reveals widespread allelic heterogeneity. *PLoS Genet.* 10:e1004322. doi: 10.1371/journal.pgen.1004322

Kowalski, M. H., Qian, H., Hou, Z., Rosen, J. D., Tapia, A. L., Shan, Y., et al. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* 15:e1008500. doi: 10.1371/journal.pgen.1008500

Lavange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., et al. (2010). Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Ann. Epidemiol.* 20, 642–649. doi: 10.1016/j.annepidem.2010.05.006

Li, B., Verma, S. S., Veturi, Y. C., Verma, A., Bradford, Y., Haas, D. W., et al. (2018). Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. *Pacific Sympos. Biocomput. Pacific Sympos. Biocomput.* 23, 448–459. doi: 10.1142/9789813235533_0041

Li, B., Veturi, Y., Verma, A., Bradford, Y., Daar, E. S., Gulick, R. M., et al. (2021). Tissue specificity-aware TWAS (TSA-TWAS) framework identifies novel associations with metabolic, immunologic, and virologic traits in HIV-positive adults. *PLoS Genet.* 17:e1009464. doi: 10.1371/journal.pgen.1009464

Li, D., Hsu, S., Purushotham, D., Sears, R. L., and Wang, T. (2019). WashU Epigenome Browser update 2019. *Nucleic Acids Res.* 47, W158–W165. doi: 10.1093/nar/gkz348

Liu, X., Finucane, H. K., Gusev, A., Bhatia, G., Gazal, S., O'Connor, L., et al. (2017). Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. *Am. J. Hum. Genet.* 100, 605–616. doi: 10.1016/j.ajhg.2017.03.002

Liu, X., Li, Y. I., and Pritchard, J. K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* 177, 1022.e–1034.e. doi: 10.1016/j.cell.2019.04.014

Lloyd-Jones, L. R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., et al. (2017). The Genetic Architecture of Gene Expression in Peripheral Blood. *Am. J. Hum. Genet.* 100, 228–237. doi: 10.1016/j.ajhg.2016.12.008

Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., et al. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* 51, 675–682. doi: 10.1038/s41588-019-0367-1

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369. doi: 10.1038/nrg2344

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122–114. doi: 10.1186/s13059-016-0974-4

Mogil, L. S., Andaleon, A., Badalamenti, A., Dickinson, S. P., Guo, X., Rotter, J. I., et al. (2018). Genetic architecture of gene expression traits across diverse populations. *PLoS Genet.* 14:e1007586. doi: 10.1371/journal.pgen.1007586

Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* 49, 1602–1612. doi: 10.1038/ng.3963

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., et al. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6:e1000895. doi: 10.1371/journal.pgen.1000895

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* 7:e1002003. doi: 10.1371/journal.pgen.1002003

Ongen, H., Brown, A. A., Delaneau, O., Panousis, N. I., Nica, A. C., GTEx Consortium, et al. (2017). Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* 49, 1676–1683. doi: 10.1038/ng.3981

Pan, D. Z., Garske, K. M., Alvarez, M., Bhagat, Y. V., Boocock, J., Nikkola, E., et al. (2018). Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. *Nat. Communicat.* 9:1512. doi: 10.1038/s41467-018-03554-9

Piasecka, B., Duffy, D., Urrutia, A., Quach, H., Patin, E., Posseme, C., et al. (2018). Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc. Natl. Acad. Sci. U S A.* 115, E488–E497. doi: 10.1073/pnas.1714765115

Pierce, B. L., and Burgess, S. (2013). Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* 178, 1177–1184. doi: 10.1093/aje/kwt084

Pividori, M., Rajagopal, P. S., Barbeira, A., Liang, Y., Melia, O., Bastarache, L., et al. (2020). PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Sci. Adv.* 6:aba2083. doi: 10.1126/sciadv.aba2083

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., et al. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26, 2336–2337. doi: 10.1093/bioinformatics/btq419

Rajewsky, N., Almouzni, G., Gorski, S. A., Aerts, S., Amit, I., Bertero, M. G., et al. (2020). LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* 2020, 1–14. doi: 10.1038/s41586-020-2715-9

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 159, 1665–1680. doi: 10.1016/j.cell.2014.11.021

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., et al. (2017). The Human Cell Atlas. *eLife* 6:503. doi: 10.7554/eLife.27041

Richardson, T. G., Hemani, G., Gaunt, T. R., Relton, C. L., and Davey Smith, G. (2020). A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *Nat. Communicat.* 11, 185–111. doi: 10.1038/s41467-019-13921-9

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature* 489, 109–113. doi: 10.1038/nature11279

Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 19, 491–504. doi: 10.1038/s41576-018-0016-z

Shang, L., Smith, J. A., Zhao, W., Kho, M., Turner, S. T., Mosley, T. H., et al. (2020). Genetic Architecture of Gene Expression in European and African Americans: An eQTL Mapping Study in GENOA. *Am. J. Hum. Genet.* 106, 496–512. doi: 10.1016/j.ajhg.2020.03.002

Snijder, B., and Pelkmans, L. (2011). Origins of regulated cell-to-cell variability. *Nat. Rev. Mol. Cell Biol.* 12, 119–125. doi: 10.1038/nrm3044

Sul, J. H., Han, B., Ye, C., Choi, T., and Eskin, E. (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet.* 9:e1003491. doi: 10.1371/journal.pgen.1003491

Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., et al. (2018). Genomic atlas of the human plasma proteome. *Nature* 558, 73–79. doi: 10.1038/s41586-018-0175-2

Sun, R., and Lin, X. (2020). Genetic Variant Set-Based Tests Using the Generalized Berk-Jones Statistic with Application to a Genome-Wide Association Study of Breast Cancer. *J. Am. Statist. Associat.* 115, 1079–1091. doi: 10.1080/01621459.2019.1660170

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315

Taylor, K., Davey Smith, G., Relton, C. L., Gaunt, T. R., and Richardson, T. G. (2019). Prioritizing putative influential genes in cardiovascular disease susceptibility by applying tissue-specific Mendelian randomization. *Genome Med.* 11, 6–15. doi: 10.1186/s13073-019-0613-2

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., et al. (2018). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* 172, 1132–1134. doi: 10.1016/j.cell.2018.02.021

Thériault, S., Gaudreault, N., Lamontagne, M., Rosa, M., Boulanger, M.-C., Messika-Zeitoun, D., et al. (2018). A transcriptome-wide association study identifies PALMD as a susceptibility gene for calcific aortic valve stenosis. *Nat. Communicat.* 9:988. doi: 10.1038/s41467-018-03260-6

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *J. R. Statist. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tournamille, C., Colin, Y., Cartron, J. P., Le Van, and Kim, C. (1995). Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat. Genet.* 10, 224–228. doi: 10.1038/ng0695-224

van der Wijst, M. G. P., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., LifeLines Cohort, et al. (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* 50, 493–497. doi: 10.1038/s41588-018-0089-9

van der Wijst, M., de Vries, D. H., Groot, H. E., Trynka, G., Hon, C. C., Bonder, M. J., et al. (2020). The single-cell eQTLGen consortium. *eLife* 9:1083. doi: 10.7554/eLife.52155

Veturi, Y., and Ritchie, M. D. (2018). How powerful are summary-based methods for identifying expression-trait associations under different genetic architectures? *Pacific Sympos. Biocomput.* 23, 228–239. doi: 10.1101/045260

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. (2007). VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.* 35, D88–D92. doi: 10.1093/nar/gkl822

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *biorxiv*. doi: 10.1101/447367

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* 51, 592–599. doi: 10.1038/s41588-019-0385-z

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. *Science* 362:eaat8464. doi: 10.1126/science.aat8464

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164–e164. doi: 10.1093/nar/gkq603

Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., et al. (2018). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* 19:151. doi: 10.1186/s13059-018-1519-9

Ward, L. D., and Kellis, M. (2011). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934. doi: 10.1093/nar/gkr917

Ward, L. D., and Kellis, M. (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* 44, D877–D881. doi: 10.1093/nar/gkv1340

Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nat. Communicat.* 8:1826. doi: 10.1038/s41467-017-01261-5

Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* 13:e1006646. doi: 10.1371/journal.pgen.1006646

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* 45, 1238–1243. doi: 10.1038/ng.2756

Whalen, S., and Pollard, K. S. (2019). Most chromatin interactions are not in linkage disequilibrium. *Genome Res.* 29, .118–.343. doi: 10.1101/gr.238022.118

Wheeler, H. E., Shah, K. P., Brenner, J., Garcia, T., Aquino-Michaels, K., GTEx Consortium, et al. (2016). Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. *PLoS Genet.* 12:e1006423–e1006423. doi: 10.1371/journal.pgen.1006423

Wu, C., and Pan, W. (2020). A powerful fine-mapping method for transcriptome-wide association studies. *Hum. Genet.* 139, 199–213. doi: 10.1007/s00439-019-02098-2

Ye, C. J., Feng, T., Kwon, H.-K., Raj, T., Wilson, M. T., Asinovski, N., et al. (2014). Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* 345, 1254665–1254665. doi: 10.1126/science.1254665

Zhang, X., Cal, A. J., and Borevitz, J. O. (2011). Genetic architecture of regulatory variation in Arabidopsis thaliana. *Genome Res.* 21, 725–733. doi: 10.1101/gr.115337.110

Zhou, D., Jiang, Y., Zhong, X., Cox, N. J., Liu, C., and Gamazon, E. R. (2020). A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nat. Genet.* 52, 1239–1246. doi: 10.1038/s41588-020-0706-2

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* 9:e1003264. doi: 10.1371/journal.pgen.1003264

Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E. A., et al. (2011). The Human Epigenome Browser at Washington University. *Nat. Methods* 8, 989–990. doi: 10.1038/nmeth.1772

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48, 481–487. doi: 10.1038/ng.3538

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. Ser. B* 67, 301–320.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Li and Ritchie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 October 2021
doi: 10.3389/fgene.2021.718890

Genetically Predicted Longer Telomere Length May Reduce Risk of Hip Osteoarthritis

Jing Yang[1†], Huiqing Xu[2†], Bingyue Cai[2], Jiahe Wei[2], Lingling Sun[3], Yasong Li[4], Tianle Wang[4] and Yingjun Li[2]*

[1]Zhuji People's Hospital of Zhejiang Province, Zhuji Affiliated Hospital of Shaoxing University, Zhuji, China, [2]School of Public Health, Hangzhou Medical College, Hangzhou, China, [3]Department of Orthopaedics, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China, [4]Department of Rheumatology and Immunology, Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou, China

Objective: This two-sample Mendelian randomization (MR) study aimed to examine the potential causal association of telomere length (TL) with the risk of osteoarthritis (OA).

Method: The summary-level data for OA was derived from the United Kingdom Biobank cohort, including 50,508 individuals of European descent. Eighteen single nucleotide polymorphisms associated with TL were identified as instrumental variables from the most up-to-date TL genome-wide association study (GWAS) involving over 78,592 individuals of European descent. Based on the GWASs data, MR was performed using established statistical analysis methods including the inverse variance weighted, weighted median, MR-Egger, and MR pleiotropy residual sum and outlier.

Results: Genetically determined TL was not associated with the risk of total OA (IVW odds ratio [OR] = 1.00, 95% confidence interval [CI] = 0.83, 1.21). In subgroup analyses stratified by OA site, no evidence in favor of association between genetically determined TL and knee OA was found (IVW OR = 1.18, 95% CI = 0.89, 1.58). However, using WM method, we observed a limited protective effect of longer TL on the risk of hip OA (OR = 0.60, 95% CI = 0.36–0.99), whereas the results of the IVW (p = 0.931) and MR-PRESSO (p = 0.932) showed that TL had no effect on hip OA.

Conclusions: This study does not support a causal association between TL and total OA. A potential protective association between longer TL and hip OA, though possible, remains less certain.

Keywords: hip OA, total OA, telomere length, mendelian randomization, knee OA

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Alessandro Gialluisi,
Istituto Neurologico Mediterraneo Neuromed (IRCCS), Italy
Jiangbo Du,
Nanjing Medical University, China

****Correspondence:***
Yingjun Li
2016034036@hmc.edu.cn

[†]These authors contributed equally to this work.

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: *01 June 2021*
Accepted: *20 September 2021*
Published: *05 October 2021*

Citation:
Yang J, Xu H, Cai B, Wei J, Sun L, Li Y, Wang T and Li Y (2021) Genetically Predicted Longer Telomere Length May Reduce Risk of Hip Osteoarthritis. Front. Genet. 12:718890. doi: 10.3389/fgene.2021.718890

INTRODUCTION

Osteoarthritis (OA) is a chronic disease that causes joint-specific pains and disabilities (Glyn-Jones et al., 2015). The latest data shows that over 303.1 million people are under the burden of increasing medical expenses and a declining quality of life produced by OA (Peat and Thomas, 2021), which involves any joint including hands, hips and knees, leading to irreversible cartilage loss or bone sclerosis (Kalamegam et al., 2018). It's well accepted that the incidence of OA has a close relationship with aging and other risk factors (Sacitharan, 2019).

Telomeres are TTAGGG repeats bound by associated protein complexes that are located at the end of chromosomes to maintain genome stability (Loe et al., 2020). Many studies have confirmed that telomeric repeats shorten with each cell division (Martinez and Blasco, 2018). Therefore, telomere length (TL) is often used as an indicator of aging. Recently, the association between telomeres and multiple age-related diseases including OA has been constantly reported (Martin and Buckwalter, 2001; Michou, 2011; Ahmed et al., 2018). However, the nature and causality of such relationship remain inconclusive. For example, a few cross-sectional studies showed that patients with OA had shorter telomeres than normal controls (Tamayo et al., 2011; Fellows et al., 2017; McAlindon et al., 2018). However, a case-control study done by Rose et al. showed that only genomic DNA damage of a higher degree occurred in OA, but the corresponding telomeres did not shorten after comparing OA cartilage and normal cartilage through autopsies (Rose et al., 2012). In addition, traditional epidemiological studies are prone to bias due to confounding factors and reverse causality, which, to a certain extent, are limited by the study sample size (Larsson et al., 2019); therefore, the causal relationship between TL and the risk of OA cannot be reliably inferred based on the observational results.

Mendelian randomization (MR) is a novel method that follows the law of independent assortment, in which genetic variants are used as instrumental variables (IVs) to assess the causal effect of exposure on outcome (Burgess et al., 2017b; Bowden and Holmes, 2019). Since the genotype of an individual is determined during conception and cannot be changed, the reverse causality between the genetic phenotype and the associated outcome is largely avoided through this method. Furthermore, based on the publicly available databases, the causal relationship between exposure and outcome can be inferred more economically and efficiently through this design.

As far as we know, no MR analysis assessing the causal association between TL and OA has been carried out by now. Here, we performed a two-sample MR study to clarify whether there existed a causal relationship between TL and the risk of OA.

MATERIALS AND METHODS

Data Sources

This study was based on summary data from a genome-wide association study (GWAS) of OA, which contained across 16.5 million variants from the resources of United Kingdom Biobank. The GWAS comprised 10,083 cases and 40,425 controls of predominantly European descent (Zengini et al., 2018). In addition, to investigate the causal link between TL and site-specific OA, summary statistics from the GWAS for knee and hip OA, which were two major subgroups of OA, were added into the study. We defined OA cases based on diagnosis records of hospitals. The diagnosis of OA coding by a hospital in the United Kingdom Biobank was based on the ICD-10 code captured from the data of Hospital Episode Statistics (HES). The sample characteristics in this study were described in **Supplementary Table S1**.

A hitherto largest genome-wide meta-analysis on leukocyte TL was conducted among up to 78,592 individuals of European descent, in which TL was measured through an established quantitative PCR technique and expressed as a ratio of the number of telomere repeats to a single-copy gene (Li et al., 2020). In the meta-analysis, a standardized quality control (QC) for exclusion [random-effect meta-analysis on single nucleotide polymorphisms (SNPs) with a Cochrane's Q p value $<1 \times 10^{-6}$, a minor allele frequency ≥1% and a sample size ≥40%] has been implemented, resulting in a total number of 2,362,330 SNPs. After adjusting the age, gender and cohort-specific covariates, Li et al. identified 20 SNPs associated with TL at genome-wide association significance ($p < 5 \times 10^{-8}$).

All participants provided a written informed consent, which was included in the study together with their aggregated data to be used for scientific publications.

SNP Selection

The MR method relies on three core assumptions (**Figure 1**), which are as follows: the SNPs must be highly associated with the exposure (the relevance assumption); they should not be associated with the confounding factors (the independence assumption); they are not associated with outcome except through their effect on exposure (the exclusion restriction assumption) (Burgess et al., 2017b). SNPs can be treated as IVs only when the three assumptions are met.

In our MR analysis, a linkage disequilibrium (LD) test was performed on each SNP identified as a genetic instrument. The LD threshold used for pruning was 0.05. And one SNPs (rs2853677) was removed at $r^2 > 0.05$ and kb < 5,000. We replaced the specific SNPs not identified in OA GWAS with the proxy in high LD ($r^2 > 0.8$). In addition, we did not exclude palindromic SNPs at MAF <0.3.

Statistical Analysis

In this study, TwoSampleMR and MRPRESSO packages with R software were used to analyze the summarized data (Yavorska and Burgess, 2017). The inverse-variance-weighted (IVW) method was our main MR analysis methodology, through which the causal effect of genetically predicted exposure on related outcome can be estimated (Hartwig et al., 2016). Several sensitivity analyses were carried out through methods including the weighted median (WM) method (Burgess et al., 2017a), the MR-Egger regression method and the MR pleiotropy residual sum and outlier (MR-PRESSO) method (Burgess and Thompson, 2017). To be more specific, the WM estimator can be used to overcome some shortcomings of IVW and gives a uniform estimate of the causal effect even when the invalid instrumental variables accounted for 50% (Bowden et al., 2016). The MR-Egger method is a statistical method through which the estimates are actually robust to horizontal pleiotropy and can reflect causal effect (Bowden et al., 2015). Nonetheless, the estimates through MR-Egger may be inaccurate and can be disturbed by outlying genetic variants. Thus, we corrected for outliers through MR-PRESSO. All statistical analyses were two-sided and considered statistically significant at a p value <0.05. Besides, through leave-one-out analysis by removing every single

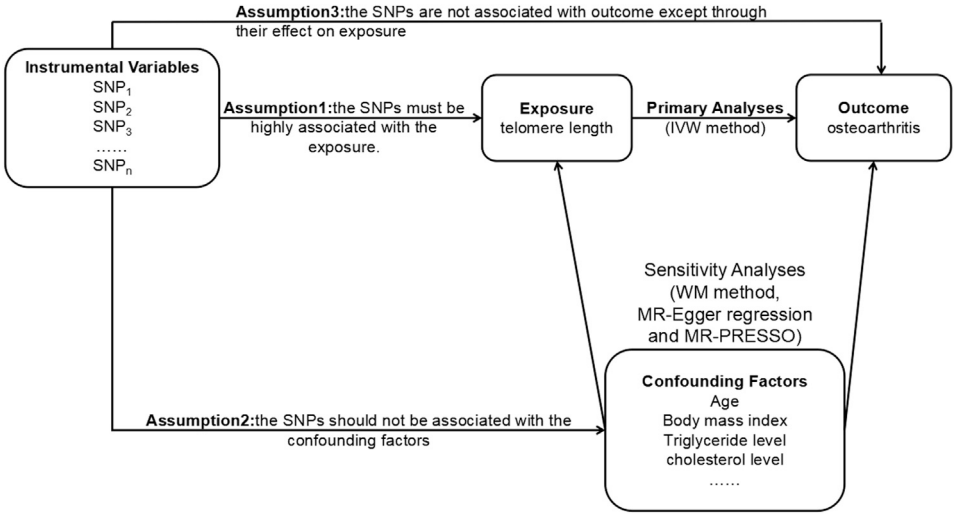


FIGURE 1 | The principles of the MR method. MR, Mendelian randomization; SNP, single nucleotide polymorphism; IVW, inverse variance weighted; WM, weighted median; MR-PRESSO, MR pleiotropy residual sum and outlier.

SNP at turn, the reliability of the results could be made sure. The related R scripts were shown in **Supplementary Text S1**.

In order to reduce the effect of covariates (i.e., confounders) on causality assessments, we manually screened the SNPs by using the Phenoscanner database and excluded the those associated with confounders. Furthermore, the strength of the IVs was assessed by calculating R^2 and the *F*-statistics (Burgess and Thompson, 2011).

RESULTS

SNPs associated with TL of European ancestry are presented in **Table 1** (Zengini et al., 2018). All of them are associated with TL at a level of genome-wide statistical significance ($p < 5 \times 10^{-8}$). After searching them in Phenoscanner database, we found that only rs73624724 (*RTEL1*) showed an association with confounding factors (e.g., body and leg fat percentage). We performed an analysis excluding this SNP. Besides, to verify whether the SNPs met Assumption 3, we analyzed the association between the SNPs and the outcome. No SNP was significantly associated with OA (*p*-value was not much lower than 0.05). Lastly, 18 sentinel variants at 17 genomic loci were taken as IVs for TL. The *F*-statistics of all SNPs were above the threshold of 10, ensuring that they were strong instruments.

As is shown in **Table 2**, the odds ratio (OR) of hip OA per standard deviation (SD) increasing in TL was 0.60 (95% confidence interval [CI] = 0.36, 0.99, $p = 0.049$] through the WM method, while nonsignificant effect estimates were shown through the IVW and MR-PRESSO method. Remarkably, our results showed evidences of weak horizontal pleiotropy (*P* for intercept = 0.050) without heterogeneity (Q = 25.18, $p = 0.091$). Similar results were suggested through MR-Egger adjusting for horizontal pleiotropy based on WM model (OR = 0.34, 95% CI = 0.12, 1.01, $p = 0.071$). Furthermore, genetically predicted TL was not related to the total risks of OA (IVW: OR = 1.00, 95% CI = 0.83, 1.21, $p = 0.989$) and knee OA (IVW: OR = 1.18, 95% CI = 0.89, 1.58, $p = 0.250$), respectively. Similar results were yielded through the WM, MR-Egger and MR-PRESSO method. The MR-Egger analysis suggested that the directional pleiotropy was unlikely to bias the causal effect of TL on total OA and knee OA (*P* for intercept = 0.196 for total OA and *P* for intercept = 0.476 for knee OA). No heterogeneity was detected by Cochran's Q statistic ($p = 0.318$ for total OA and $p = 0.273$ for knee OA). Scatter plots, forest plots and funnel plots of the total OA, knee OA and hip OA are presented in **Supplementary Figure S1–S9**.

The results of leave-one-out sensitivity analysis showed that no single SNP had a significant effect on the pooled results, suggesting the stability of our results (**Supplementary Figure S10–S12**).

DISCUSSION

To the best of knowledge, this is the first MR study examining the relationship between TL and the risk of total OA, as well as OA at specific joint sites (knee and hip). These results suggest that the association between TL and total OA risk is not likely to be causal. Through subgroup analyses, we find some evidences supporting the causal association between TL and hip OA instead of knee OA.

It should be noted that the genetically predicted TL is not associated with the risk of total OA in our MR study. The present study is in disagreement with quite a few previous observational studies. The findings of some case-control studies show that TL is inversely related to the risk of OA (Zhai et al., 2006; Poonpet et al.,

TABLE 1 | Characteristics of SNPs used as genetic instruments for TL in the present MR study.

SNP	Chr	Position	Nearest gene	Effect allele	Non-effect allele	EAF	TL			OA			R^{2a}	*F*-statistic[b]	Remove
							Beta	SE	*p* Value	Beta	SE	*p* Value			
rs3219104	1	226562621	PARP1	C	A	0.83	0.042	0.006	9.60E-11	–0.033	0.022	0.13	0.0006	49	-
rs10936600	3	169,514,585	TERC	T	A	0.24	–0.086	0.006	7.18E-51	0.006	0.018	0.74	0.0026	205	-
rs4691895	4	164,048,199	NAF1	C	G	0.78	0.058	0.006	1.58E-21	–0.022	0.019	0.24	0.0012	93	-
rs7705526	5	1,285,974	TERT	A	C	0.33	0.082	0.006	5.34E-45	0.008	0.017	0.66	0.0024	187	-
rs2853677	5	1,287,194	TERT	A	G	0.59	–0.064	0.006	3.35E-31	–0.007	0.016	0.65	0.0014	114	$r^2 > 0.05$[c]
rs59294613	7	124,554,267	POT1	A	C	0.29	–0.041	0.006	1.17E-13	–0.007	0.017	0.69	0.0006	47	-
rs9419958	10	105,675,946	STN1	C	T	0.86	–0.064	0.007	5.05E-19	0.005	0.023	0.84	0.0011	84	-
rs228595	11	108,105,593	ATM	A	G	0.42	–0.029	0.005	1.43E-08	0.004	0.016	0.81	0.0004	34	-
rs2302588	14	73,404,752	DCAF4	C	G	0.10	0.048	0.008	1.68E-08	0.024	0.026	0.36	0.0005	36	-
rs7194734	16	82,199,980	STMN3	T	C	0.78	–0.037	0.006	6.94E-10	–0.024	0.019	0.19	0.0005	38	-
rs8105767	19	22,215,441	ZNF208	G	A	0.30	0.039	0.005	5.42E-13	0.007	0.017	0.67	0.0008	61	-
rs75691080	20	62,269,750	STMN3	T	C	0.09	–0.067	0.009	5.88E-14	–0.026	0.029	0.37	0.0007	55	-
rs34978822	20	62,291,599	RTEL1	G	C	0.02	–0.140	0.023	7.26E-10	0.104	0.053	0.05	0.0005	37	-
rs73624724	20	62,436,398	RTEL1	C	T	0.13	0.051	0.007	6.33E-12	–0.028	0.023	0.22	0.0007	53	body fat percentage[d]
rs55749605	3	0,1,232,093	SENP7	A	C	0.58	–0.037	0.007	2.45E-08	–0.009	0.016	0.56	0.0004	28	-
rs13137667	4	71,774,347	MOB1B	C	T	0.96	0.077	0.014	2.43E-08	0.036	0.049	0.46	0.0004	30	-
rs34991172	6	25,480,328	CARMIL	G	T	0.07	–0.061	0.011	6.19E-09	0.031	0.028	0.28	0.0004	31	-
rs2736176	6	31,587,561	PRRC2A	C	G	0.31	0.035	0.006	3.53E-10	–0.006	0.017	0.73	0.0004	34	-
rs3785074	16	69,406,986	TERF2	G	A	0.26	0.035	0.006	4.64E-10	0.004	0.017	0.83	0.0004	34	-
rs62053580	16	74,680,074	RFWD3	G	A	0.17	–0.039	0.007	4.08E-08	–0.050	0.021	0.02	0.0004	31	-

Abbreviation: SNP, single nucleotide polymorphism; Chr, chromosome; EAF, effect allele frequency; TL, telomere length; SE, standard error; OA, osteoarthritis.

[a]R^2 was calculated using the following formula: $(2\times EAF\times(1-EAF)\times Beta^2)/[(2\times EAF\times(1-EAF)\times Beta^2) + (2\times EAF\times(1-EAF)\times N\times SE^2)]$, where Beta is the estimated effect on telomere length, N is the sample size of the GWAS for the SNP-telomere length association and SE is the standard error of the estimated effect.

[b]F statistic was calculated using the following formula: $R^2(N-2)/(1-R^2)$, where R^2 is the proportion of variance in telomere length explained by each instrument and N is the sample size of the GWAS for the SNP-telomere length association.

[c]SNP with $r^2 > 0.05$ was removed.

[d]SNP associated with confounding factors (e.g., body fat percentage) was removed.

TABLE 2 | MR results of the association between TL and OA (18 SNPs).

MR method	OR (95%CI)	*p*-value for association[a]	Cochran's Q statistic	*p*-value for heterogeneity	*p*-value for MR egger intercept
OA (10,083/40,425)					
IVW	1.00 (0.83, 1.21)	0.989	19.18	0.318	-
WM	1.01 (0.78, 1.30)	0.948	-	-	-
MR-Egger	0.73 (0.45, 1.20)	0.230	-	-	0.196
MR-PRESSO	1.00 (0.81, 1.19)	0.989	-	-	-
Knee OA (4,462/17,885)					
IVW	1.18 (0.89, 1.58)	0.250	20.02	0.273	-
WM	1.19 (0.81, 1.75)	0.366	-	-	-
MR-Egger	0.90 (0.41, 1.98)	0.800	-	-	0.476
MR-PRESSO	1.18 (0.90, 1.47)	0.266	-	-	-
Hip OA (2,396/9,593)					
IVW	1.02 (0.66, 1.59)	0.931	25.18	0.091	-
WM	0.60 (0.36, 0.99)	**0.049**	-	-	-
MR-Egger	0.34 (0.12, 1.01)	**0.071**	-	-	0.050
MR-PRESSO	1.02 (0.58, 1.46)	0.932	-	-	-

Abbreviation: MR, Mendelian randomization; OR, odds ratio; CI, confidence interval; OA, osteoarthritis; IVW, inverse variance weighted; WM, weighted median; MR-PRESSO, Mendelian randomization pleiotropy residual sum and outlier.

[a]*Bolded* P *represents limited significance.*

2018; Manoy et al., 2020). Similarly, a recently published meta-analysis including ten studies indicates that TL might be a potential biomarker of OA (Xie et al., 2021). On the contrary, as is suggested in another study, there is no association between TL and OA after adjusting age (Tamayo et al., 2010). Taken together, the results of previous epidemiological studies on OA remain contradictory, which may be caused by differences in the joint site, disease progression, population selection or techniques used in the TL measurement. To compute unmeasured confounding in previous observational studies, we reported the E-value, defined as the minimum strength of association on the risk ratio scale that an unmeasured confounder would need to have with both the exposure and the outcome, conditional on the measured covariates, to fully explain away a specific exposure-outcome association (VanderWeele and Ding, 2017). E-value was computed online (https://www.evalue-calculator.com/). The lowest possible E-value is 1. The larger the E-value is, the stronger the unmeasured confounder associations would have to be considered to explain away an effect. More details were shown in **Supplementary Table S2**.

In the MR study on the OA subgroup, we found that TL was related to the risk of hip OA rather than that of knee OA, thus revealing the specificity at joint sites. For hip OA, our results are largely consistent with that of previous observational studies. An indirect association between TL and the risk of hip OA is reported in a study, supporting the conclusion that telomere shortening is related to cartilage degradation (Harbo et al., 2013). A single or a few critically short telomeres are enough to trigger cellular senescence of normal cells, which result in many degenerative and aging-related diseases including OA. The severity of OA is associated with the increase in the number of senescent cells in joint tissues, as the accumulation of senescent chondrocytes will reduce the ability of chondrocytes to maintain and repair cartilage, so that tissues are unable to bear stress (Harbo et al., 2013). Besides, due to the specific role of the hip joint and specific lifestyles, shorter telomeres are less resistant to inflammation and oxidative stress (OXS), leading to an increased formation of oxLDL particles and a decreased level of spermidine, thus resulting in the development and progression of hip pathology (Tootsi et al., 2017). The above may be potential mechanisms that could explain different associations between TL and specific joint sites.

The causal association between longer TL and hip OA, though possible, remains less certain. Both WM and MR-Egger regression method suggested a limited causal association between TL and hip OA, and the IVW and MR-PRESSO method suggested a null causal association. As far as we know, the WM gives a correct estimate of the causal effect as long as at least 50% of the weight comes from effective IVs. Moreover, compared to IVW, it has better finite-sample Type 1 error rates (Bowden et al., 2016). As for the MR-PRESSO, it detects and corrects the outliers in IVW linear regression. In the study, no outliers were highlighted by MR-PRESSO. The MR-Egger regression method may provide estimates for the true causal effect that is consistent even if all the genetic variants are invalid instruments, as long as the instrument strength independent of direct effects (InSIDE) assumption is satisfied (Bowden et al., 2015). Thus, the effect of TL on hip OA was potentially causal, as indicated by the WM and MR-Egger approach.

The primary strength of MR study is less susceptible to confounding and reverse causality, thereby overcoming the defects of observational studies on the association between TL and OA, which provides a strong support to our result (Fan et al., 2020). There are also several limitations. Firstly, only European individuals participated and were included in this study. Thus, our results might not be generalizable to ancestries other than European, and it is essential to verify our results in other populations. Secondly, only summary-level data was used in the MR analysis, which was not sufficient to make further stratified analyses on other specific factors, such as age and

gender, etc. Thirdly, not all MR models showed a statistically significant association between TL and the risk of hip OA. Nevertheless, consistent causal estimates could be provided through the MR-Egger method while taking into account directional pleiotropy. Besides, we conducted an additional sensitivity analysis on potential horizontal pleiotropy, providing evidence for the robustness of our results. Finally, further MR studies as well as longitudinal studies are needed to investigate the association between TL and OA.

In summary, this MR study does not support the causal association between TL and the total OA or knee OA. A potential protective association between longer TL and hip OA, though possible, remains less certain.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JY and HX mainly designed and performed analysis, verifying data and wrote the article; BC and JW performed experiments and analysis; LS, YL, and TW reviewed data and the article; YL supervised the entire project.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant number 81703289).

ACKNOWLEDGMENTS

The authors thank all the participants and researchers for their participation in this MR study. All authors interpreted the data, critically revised the manuscript for important intellectual content, approved the final version of the manuscript, and agreed to be responsible for all aspects of the work. The corresponding author attests that all listed co-authors meet authorship criteria and that no others meeting the criteria have been omitted.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.718890/full#supplementary-material.

REFERENCES

Ahmed, M. S., Ikram, S., Bibi, N., and Mir, A. (2018). Hutchinson-Gilford Progeria Syndrome: A Premature Aging Disease. *Mol. Neurobiol.* 55 (5), 4417–4427. doi:10.1007/s12035-017-0610-7

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression. *Int. J. Epidemiol.* 44 (2), 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* 40 (4), 304–314. doi:10.1002/gepi.21965

Bowden, J., and Holmes, M. V. (2019). Meta-analysis and Mendelian Randomization: A Review. *Res. Syn Meth.* 10 (4), 486–496. doi:10.1002/jrsm.1346

Burgess, S., Bowden, J., Fall, T., Ingelsson, E., and Thompson, S. G. (2017a). Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* 28 (1), 30–42. doi:10.1097/ede.0000000000000559

Burgess, S., Small, D. S., and Thompson, S. G. (2017b). A Review of Instrumental Variable Estimators for Mendelian Randomization. *Stat. Methods Med. Res.* 26 (5), 2333–2355. doi:10.1177/0962280215597579

Burgess, S., and Thompson, S. G. (2011). Avoiding Bias from Weak Instruments in Mendelian Randomization Studies. *Int. J. Epidemiol.* 40 (3), 755–764. doi:10.1093/ije/dyr036

Burgess, S., and Thompson, S. G. (2017). Interpreting Findings from Mendelian Randomization Using the MR-Egger Method. *Eur. J. Epidemiol.* 32 (5), 377–389. doi:10.1007/s10654-017-0255-x

Fan, J., Zhu, J., Sun, L., Li, Y., Wang, T., and Li, Y. (2020). Causal Association of Adipokines with Osteoarthritis: a Mendelian Randomization Study. *Rheumatology (Oxford)* 60, 2808–2815. doi:10.1093/rheumatology/keaa719

Fellows, C. R., Williams, R., Davies, I. R., Gohil, K., Baird, D. M., Fairclough, J., et al. (2017). Characterisation of a Divergent Progenitor Cell Sub-populations in Human Osteoarthritic Cartilage: the Role of Telomere Erosion and Replicative Senescence. *Sci. Rep.* 7, 41421. doi:10.1038/srep41421

Glyn-Jones, S., Palmer, A. J. R., Agricola, R., Price, A. J., Vincent, T. L., Weinans, H., et al. (2015). Osteoarthritis. *Lancet* 386 (9991), 376–387. doi:10.1016/s0140-6736(14)60802-3

Harbo, M., Delaisse, J. M., Kjaersgaard-Andersen, P., Soerensen, F. B., Koelvraa, S., and Bendix, L. (2013). The Relationship between Ultra-short Telomeres, Aging of Articular Cartilage and the Development of Human Hip Osteoarthritis. *Mech. Ageing Develop.* 134 (9), 367–372. doi:10.1016/j.mad.2013.07.002

Hartwig, F. P., Davies, N. M., Hemani, G., and Davey Smith, G. (2016). Two-sample Mendelian Randomization: Avoiding the Downsides of a Powerful, Widely Applicable but Potentially Fallible Technique. *Int. J. Epidemiol.* 45 (6), 1717–1726. doi:10.1093/ije/dyx028

Kalamegam, G., Memic, A., Budd, E., Abbas, M., and Mobasheri, A. (2018). A Comprehensive Review of Stem Cells for Cartilage Regeneration in Osteoarthritis. *Adv. Exp. Med. Biol.* 1089, 23–36. doi:10.1007/5584_2018_205

Larsson, S. C., Michaëlsson, K., and Burgess, S. (2019). Mendelian Randomization in the Bone Field. *Bone* 126, 51–58. doi:10.1016/j.bone.2018.10.011

Li, C., Stoma, S., Lotta, L. A., Warner, S., Albrecht, E., Allione, A., et al. (2020). Genome-wide Association Analysis in Humans Links Nucleotide Metabolism to Leukocyte Telomere Length. *Am. J. Hum. Genet.* 106 (3), 389–404. doi:10.1016/j.ajhg.2020.02.006

Loe, T. K., Li, J. S. Z., Zhang, Y., Azeroglu, B., Boddy, M. N., and Denchi, E. L. (2020). Telomere Length Heterogeneity in ALT Cells Is Maintained by PML-dependent Localization of the BTR Complex to Telomeres. *Genes Dev.* 34 (9-10), 650–662. doi:10.1101/gad.333963.119

Manoy, P., Yuktanandana, P., Tanavalee, A., Tanpowpong, T., Ittipanichpong, T., and Honsawek, S. (2020). Telomere Shortening Is Associated with Poor Physical Performance in Knee Osteoarthritis. *Biomed. Rep.* 13 (4), 27. doi:10.3892/br.2020.1334

Martin, J. A., and Buckwalter, J. A. (2001). Telomere Erosion and Senescence in Human Articular Cartilage Chondrocytes. *J. Gerontol. Ser. A: Biol. Sci. Med. Sci.* 56 (4), B172–B179. doi:10.1093/gerona/56.4.b172

Martínez, P., and Blasco, M. A. (2018). Heart-Breaking Telomeres. *Circ. Res.* 123 (7), 787–802. doi:10.1161/CIRCRESAHA.118.312202

McAlindon, T., Roberts, M., Driban, J., Schaefer, L., Haugen, I. K., Smith, S. E., et al. (2018). Incident Hand OA Is Strongly Associated with Reduced Peripheral Blood Leukocyte Telomere Length. *Osteoarthr. Cartil.* 26 (12), 1651–1657. doi:10.1016/j.joca.2018.08.010

Michou, L. (2011). Genetics of Digital Osteoarthritis. *Jt. Bone Spine* 78 (4), 347–351. doi:10.1016/j.jbspin.2010.09.012

Peat, G., and Thomas, M. J. (2021). Osteoarthritis Year in Review 2020: Epidemiology & Therapy. *Osteoarthr. Cartil.* 29 (2), 180–189. doi:10.1016/j.joca.2020.10.007

Poonpet, T., Saetan, N., Tanavalee, A., Wilairatana, V., Yuktanandana, P., and Honsawek, S. (2018). Association between Leukocyte Telomere Length and Angiogenic Cytokines in Knee Osteoarthritis. *Int. J. Rheum. Dis.* 21 (1), 118–125. doi:10.1111/1756-185x.12988

Rose, J., Söder, S., Skhirtladze, C., Schmitz, N., Gebhard, P. M., Sesselmann, S., et al. (2012). DNA Damage, Discoordinated Gene Expression and Cellular Senescence in Osteoarthritic Chondrocytes. *Osteoarthr. Cartil.* 20 (9), 1020–1028. doi:10.1016/j.joca.2012.05.009

Sacitharan, P. K. (2019). Ageing and Osteoarthritis. *Subcell Biochem.* 91, 123–159. doi:10.1007/978-981-13-3681-2_6

Tamayo, M., Mosquera, A., Rego, I., Blanco, F. J., Gosálvez, J., and Fernández, J. L. (2011). Decreased Length of Telomeric DNA Sequences and Increased Numerical Chromosome Aberrations in Human Osteoarthritic Chondrocytes. *Mutat. Res. Fundam. Mol. Mech. Mutagen.* 708 (1-2), 50–58. doi:10.1016/j.mrfmmm.2011.01.007

Tamayo, M., Mosquera, A., Rego, J. I., Fernández-Sueiro, J. L., Blanco, F. J., and Fernández, J. L. (2010). Differing Patterns of Peripheral Blood Leukocyte Telomere Length in Rheumatologic Diseases. *Mutat. Mutat. Res. Fundam. Mol. Mech. Mutagen.* 683 (1-2), 68–73. doi:10.1016/j.mrfmmm.2009.10.010

Tootsi, K., Märtson, A., Kals, J., Paapstel, K., and Zilmer, M. (2017). Metabolic Factors and Oxidative Stress in Osteoarthritis: a Case-Control Study. *Scand. J. Clin. Lab. Invest.* 77 (7), 520–526. doi:10.1080/00365513.2017.1354255

VanderWeele, T. J., and Ding, P. (2017). Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann. Intern. Med.* 167 (4), 268–274. doi:10.7326/m16-2607

Xie, H., Ma, Y., Shao, M., Kong, J., Zhou, T., Wang, F., et al. (2021). Telomere Length in Patients with Osteoarthritis: a Systematic Review and Meta-Analysis. *Aging Clin. Exp. Res.* doi:10.1007/s40520-021-01944-6

Yavorska, O. O., and Burgess, S. (2017). MendelianRandomization: an R Package for Performing Mendelian Randomization Analyses Using Summarized Data. *Int. J. Epidemiol.* 46 (6), 1734–1739. doi:10.1093/ije/dyx034

Zengini, E., Hatzikotoulas, K., Tachmazidou, I., Steinberg, J., Hartwig, F. P., Southam, L., et al. (2018). Genome-wide Analyses Using UK Biobank Data Provide Insights into the Genetic Architecture of Osteoarthritis. *Nat. Genet.* 50 (4), 549–558. doi:10.1038/s41588-018-0079-y

Zhai, G., Aviv, A., Hunter, D. J., Hart, D. J., Gardner, J. P., Kimura, M., et al. (2006). Reduction of Leucocyte Telomere Length in Radiographic Hand Osteoarthritis: a Population-Based Study. *Ann. Rheum. Dis.* 65 (11), 1444–1448. doi:10.1136/ard.2006.056903

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Yang, Xu, Cai, Wei, Sun, Li, Wang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 13 October 2021
doi: 10.3389/fgene.2021.720507

Molecular Cytogenetics in the Era of Chromosomics and Cytogenomic Approaches

*Thomas Liehr**

Jena University Hospital, Institute of Human Genetics, Friedrich Schiller University, Jena, Germany

Here the role of molecular cytogenetics in the context of yet available all other cytogenomic approaches is discussed. A short introduction how cytogenetics and molecular cytogenetics were established is followed by technical aspects of fluorescence *in situ* hybridization (FISH). The latter contains the methodology itself, the types of probe- and target-DNA, as well as probe sets. The main part deals with examples of modern FISH-applications, highlighting unique possibilities of the approach, like the possibility to study individual cells and even individual chromosomes. Different variants of FISH can be used to retrieve information on genomes from (almost) base pair to whole genomic level, as besides only second and third generation sequencing approaches can do. Here especially highlighted variations of FISH are molecular combing, chromosome orientation-FISH (CO-FISH), telomere-FISH, parental origin determination FISH (POD-FISH), FISH to resolve the nuclear architecture, multicolor-FISH (mFISH) approaches, among other applied in chromoanagenesis studies, Comet-FISH, and CRISPR-mediated FISH-applications. Overall, molecular cytogenetics is far from being outdated and actively involved in up-to-date diagnostics and research.

Keywords: cytogenomics, chromosomics, topologically associating domains (TADs), copy number variations (CNVs), small supernumerary marker chromosomes (sSMCs), chromosomal heteromorphisms (CHMs), glass-needle based chromosome microdissection (midi), chromothripsis

OPEN ACCESS

Edited by:
Mehdi Pirooznia, National Heart, Lung, and Blood Institute (NHLBI), United States

Reviewed by:
Vijaya Kumar Pidugu, National Cancer Institute (NIH), United States
Usha Dutta, Centre for DNA Fingerprinting and Diagnostics (CDFD), India

***Correspondence:**
Thomas Liehr
Thomas.Liehr@med.uni-jena.de

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *04 June 2021*
Accepted: *22 September 2021*
Published: *13 October 2021*

Citation:
Liehr T (2021) Molecular Cytogenetics in the Era of Chromosomics and Cytogenomic Approaches. Front. Genet. 12:720507. doi: 10.3389/fgene.2021.720507

INTRODUCTION

This review is about "molecular cytogenetics" including 1) the historical perspective of its development from cytogenetics, 2) technical aspects, 3) available probe sets, and 4) variants and applications of the basic fluorescence *in situ* hybridization (FISH) approach. According to present zeitgeist, it is trendy to replace the word cytogenetics and/or the application of whole genome oriented molecular genetic approaches, by the designation "cytogenomics". Thus, first a few comments on this point are necessary to understand why a change from the designation "molecular cytogenetics" to "molecular cytogenomics" is not justified by any means, even though "molecular cytogenetics" is clearly a "cytogenomic approach".

Cytogenomics and Chromosomics

In literature, the neologism "cytogenomics" reflects normally only "the changes in technology under its purview" (McGowan-Jordan et al., 2020), which is overall an a bit weak argument to replace the long standing, clearly defined word "cytogenetics" by a new one. Strikingly, a definition for this word coined already in 1999 (Bernheim, 1999) is hard to find in the literature; in 2019, it was referred to as "a general term that encompasses conventional, as well as molecular cytogenetics (FISH,

microarrays) and molecular-based techniques" (Siva et al., 2019). It is here suggested that the word "cytogenomics" should rather be used with the goal to paraphrase a new field of research in genomics and diagnostics in human genetics, with an integrative and comprehensive view. Cytogenomics is, under this definition, nothing else than an equivalent wording for "chromosomics", a designation introduced in 2005 by Prof. Uwe Claussen (Jena, Germany) (Claussen, 2005; Liehr, 2019). He suggested to introduce the term chromosomics being equal to cytogenomics to bring the three-dimensional morphologically of chromosomes into the focus of research, as this is essential for gene regulation. Under this generic term, all chromosome-related studies should be summarized to introduce novel ideas and concepts in biology and medicine, thus having an integrative effect on the field. The latter is an extraordinary thinking approach, as in most other cases a new "omics"-field was introduced to separate the corresponding field from all the others (Bernheim, 1999; Claussen, 2005).

CYTOGENETICS—HISTORICAL ASPECTS

Cytogenetics is the study of chromosomes, which were seen first in 1879; Walter Flemming was at that time the one to introduce the designations "chromatin" and "mitosis" (Flemming, 1879). "Molecular cytogenetics" developed from cytogenetics field, later. In 1888, Heinrich W. Waldeyer introduced the name "stained body" as "chromosome" (Waldeyer, 1888) for what Gregor Mendel already postulated as "Kopplungsgruppen", which refers to "linked up groups" in German (Mendel, 1866). Walter Sutton and Theodor Boveri were then the first to suggest in 1902/03 the chromosome-theory of inheritance (Boveri, 1902; Sutton, 1903).

Human cytogenetic discipline, in particular, underwent different developmental steps—each providing more and better possibilities for the characterization of acquired and constitutional chromosomal aberrations. Reliable identification of such alterations started with banding cytogenetics technique, introduced by Dr. Lore Zech (Uppsala, Sweden) by 1970 (Schlegelberger, 2013). Further approaches, as C-banding (Arrighi and Hsu, 1971) and silver staining of nucleolus organizing (Goodpasture and Bloom, 1975), complemented the cytogenetic method-set by mid to end of the 1970s. GTG-banding (G-bands by Trypsin using Giemsa) (Seabright, 1971) is still considered as the gold standard of chromosomal diagnostics (Schlegelberger, 2013). Even though, without any proof of evidence, cytogenetics is called dead for decades (Salman et al., 2004), it is imperative to remember that each single available "cytogenomic approach" provides unique and complementary possibilities to obtain information from a genome; the latter can be retrieved at single cell-, or millions of cell-level and at different resolutions (Hochstenbach et al., 2019). Cytogenetics has a low resolution of 5–10 mega base pairs, but enables a whole genomic view; it is cost-efficient and single cell oriented; i.e., it is able to pick up small mosaics. Retrospectively one can state that molecular cytogenetics was developed with the following goals: 1) to take still advantage of possibilities of banding cytogenetics, but 2) to overcome the limitation of its low resolution, and 3) to include the possibility to analyze interphase cells, too (Zhang et al., 2018). Between 1969 and 1986, *in situ* hybridization (ISH)

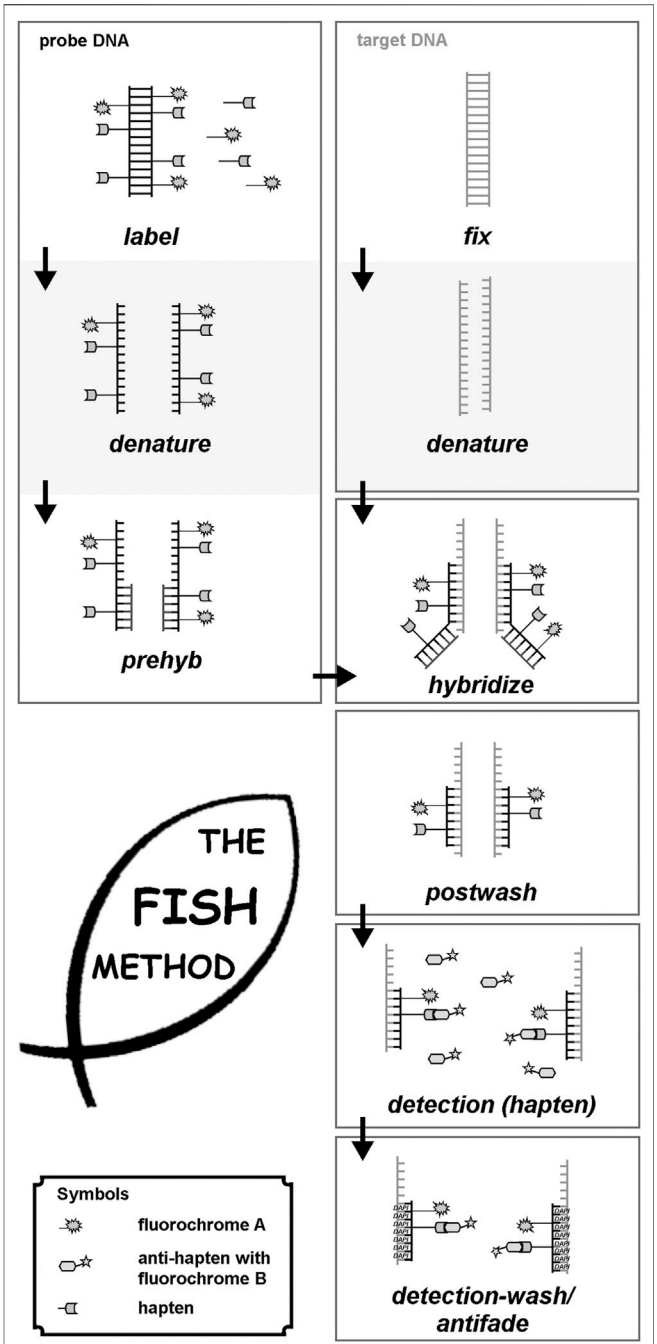


FIGURE 1 | Principle of FISH is given here schematically. First probe- and target DNA are denatured. Probe-DNA is either labeled (commercial probes), or needs to be labeled, e.g., by PCR-approaches. Probe-DNA is pre-hybridized with unlabeled repetitive DNA and then hybridized to the target DNA, being fixed on a glass-slide. After hybridization, postwashing is done to get rid of superfluous probe and blocking DNA. In case a non-fluorescent hapten was used to label the probe-DNA, this has to be detected by an anti-hapten with a fluorochrome. Finally, after washing of the detection-solution, slides can be sealed with antifade and DNA-staining dye (like DAPI = Dipehnylaminoindol), and evaluated under a suited microscope.

could exclusively be performed as a radioactive variant. Nonradioactive probe labeling using biotin as nonradioactive hapten (detectable by fluorochrome-coupled avidin) was developed in 1981, and thus, not earlier than in 1986, the first fluorescence ISH (FISH) on human chromosomes was reported. Besides FISH, also the primed *in situ* hybridization (PRINS) technique was an important molecular cytogenetic approach between 1989 and 2010 (Koch et al., 1995).

MOLECULAR CYTOGENETICS—FISH

FISH, the only remaining approach of molecular cytogenetics, was first available as single (Pinkel et al., 1986) and dual color approach (Hopman et al., 1986); since 1998, it could also be applied in multicolor FISH (Nederlof et al., 1998). The first mile stone in multicolor-FISH was the simultaneous use of all 24 human whole chromosome paints in one experiment (Speicher et al., 1996; Schröck et al., 1996). Besides many other multicolor-FISH (mFISH), probe sets were developed and are summarized elsewhere (Liehr, 2021). FISH is used in multiple ways in diagnostics and research—one of the latest and most interesting developments for both fields maybe at present the molecular combing approach (Florian et al., 2019).

The principle of FISH is simple (Pinkel et al., 1986), and nowadays, it is a well-established approach with hundreds of commercially available and applicable probes and probe sets (Liehr, 2017; 2021) (**Figure 1**). Nonetheless, to find the right laboratory protocols needed initially 1 decade.

Molecular cytogenetics can be performed on different kinds of samples. While in banding cytogenetics it is imperative to have a chromosomal preparation, FISH can be done also on tissue sections and in interphase nuclei. Necessary are always a target-DNA (metaphases or interphases, or for molecular combing (see below) DNA-fibers) and a probe-DNA. The latter has to be labelled with an under a fluorescence microscope detectable hapten (see below). The following steps have to be performed (see also **Figure 1**):

- Denaturation of target- and probe-DNA;
- Incubation of target- and probe-DNA in a hybridization solution at 37°C for several hours (with or without blocking of repetitive DNA-sequences to avoid possible background);
- Washing off superfluous probe-DNA with suited buffers;
- If necessary, detection of the hapten bound to probe-DNA with a fluorophore-labelled antibody; otherwise—if probe-DNA is already fluorescence labelled—addition of antifade-solution and coverslip;
- Evaluation under the fluorescence microscope.
- More technical details can be found elsewhere (Liehr, 2017; 2021).

Types of Target-DNA/Samples in FISH

For FISH experiments, samples need to contain intact, non-degraded high molecular weight target-DNA. All tissues of any species fulfilling this prerequisite can be applied in FISH (Liehr, 2017)—even bacteria can be accessed (Bottari et al., 2006). Accordingly, in multicellular organisms, native cells, extracted nuclei, tissue sections, metaphase chromosomes, or pure DNA can be used as target-DNA. In human, most often used are easily accessible tissues, or such acquired during surgeries, e.g., peripheral blood lymphocytes, bone marrow cells, skin fibroblasts, buccal mucosa, hair root cells, urine derived cells, amniotic fluid, chorion biopsy derived cells, gametes (sperm and oocytes), or tumor cells (including formalin-fixed paraffin-embedded tissues). More details can be found elsewhere (Liehr, 2017; 2021).

Types of Probe-DNA Suited for FISH

On the one hand, there are commercially available probes, especially for molecular cytogenetics based chromosomic research and diagnostics in humans. These probes are usually ready to use and labeled with corresponding fluorophores or non-fluorescent haptens (for review on commercial probes for cancer cytogenetics (Liehr et al., 2015)). The second type of probe-DNA for FISH are in house probes, which need to be labeled, either by PCR-based approaches, Nick-translation, or the so-called Universal Linkage System (ULS) (Liehr, 2017; 2021). In the following, five basic types of probe-DNA applied for FISH are listed.

Locus Specific Probes

Locus-specific probes (LSPs) are normally derived from molecular cloning experiments. Accordingly, genetic vectors, including all kinds of plasmids, bacterial and yeast artificial chromosomes, or others are suited if they contain the wanted insert of species-specific DNA to be targeted by FISH, with inserts of a minimal size of 12 kb (Liehr, 2017; 2021). Alternatively, also contiguous probes may be used (Smith et al., 1997), or for mapping purposes, even smaller single copy probes (Nguyen et al., 2019).

Repetitive Probes

Repetitive DNA can be easily visualized in FISH experiments. Thus, repetitive probes, targeting centromeres, telomeres, or other repetitive, e.g., interspersed satellite-DNAs, result in strong and easily evaluable signals. Interestingly, at least one repetitive DNA (D4Z4) localized in 4q24 has some meaning in human genetic diagnostics and can be traced by molecular combing (Nguyen et al., 2019).

Partial Chromosome Paints

Partial chromosome paints (pcps) can be established by glass needle-based chromosome microdissection (midi) (Al-Rikabi et al., 2019). Pcps simultaneously stain at least 1 or 2 euchromatic chromosomal subbands and are normally not larger than a chromosome arm.

Whole Chromosome Paints

A whole chromosome paint (wcp), staining an entire chromosome can either be established by midi (Ferguson-Smith et al., 2005) or by chromosome flow sorting (Sabile

TABLE 1 | Literature and reviews on molecular cytogenetics in routine diagnostics for the major fields of application in human genetic diagnostics.

Routine diagnostics	References
Prenatal FISH	Weise and Liehr (2008), Pellestor et al. (2011), Liehr (2017), chapter "Commercial FISH-probes" Sala et al. (2019)
Postnatal FISH	Liehr (2017), chapter "Commercial FISH-probes" Liehr and Hamid Al-Rikabi (2018)
Tumor cytogenetic FISH in leukemia and lymphoma	Liehr et al. (2015), Liehr (2017), chapter "Commercial FISH-probes" Cui et al. (2016)
FISH in solid tumors	Cheng et al. (2017), Liehr (2017), chapter "Commercial FISH-probes" Liehr (2017), "interphase FISH in diagnostics"

et al., 1997). Besides, interspecies hybrids (e.g., mouse/human somatic cell hybrid) have been used as sources of species-specific wcp probes (Sabile et al., 1997).

Whole Genome Probes

Even whole genomic DNA can be applied in FISH. This only is informative when using a trick: in a comparative genomic hybridization (CGH) setting (Kallioniemi et al., 1992), two whole genomes labelled in two different colors are co-hybridized to normal human blood-derived metaphases. This approach can be used as CGH in comparative cytogenomics in evolution-research (Majka et al., 2017), or as molecular karyotyping or array-CGH (aCGH) in human genetic diagnostics (Pinkel et al., 1998).

Molecular Cytogenetic Probe Sets

It is possible to combine the just listed different FISH-probes in two-to multicolor-FISH probe sets (Liehr, 2017; 2021). As it is impossible to list all of yet done combinations, in the following sections, only some probe sets together with their applications in the concert of cytogenomic approaches are included. Their impact on chromosomic research and human genetic diagnostics is discussed, too.

MOLECULAR CYTOGENETICS IN ROUTINE DIAGNOSTICS

Even though there are (elsewhere (Liehr, 2020) in more detail discussed) problems of getting sufficient reimbursement for routine FISH-diagnostics, molecular cytogenetic is and remains of constant, and even growing importance in many fields of genetics. Fields of applications include pre- and postnatal as well as tumor diagnostics on cytogenetically worked up cells, with interphase-, as well as metaphase-FISH being performed. Also, FISH is routinely done in FFPE (formalin fixed, paraffin embedded) material for solid tumor diagnostics in pathology (Liehr, 2017).

All aforementioned probes combined in two-to multi-color-FISH approaches are applied in molecular cytogenetic routine diagnostics. While in metaphase-FISH there is no restriction in number and types of probes, in interphase-FISH preferentially less than six LSPs and/or centromeric probes are applied. Especially during last decade, probe sets were developed not only to detect loss or gain of copy numbers, reciprocal translocations/gene fusions or gene splitting, but also such to distinguish different fusion partners and/or detect even inversions in interphase nuclei (Liehr, 2017).

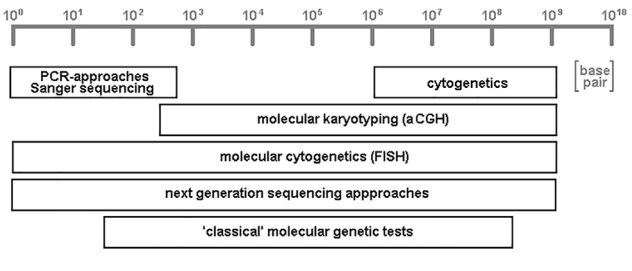


FIGURE 2 | Schematic depiction of resolutions achievable by variants of different cytogenomic approaches. Cytogenetics provides low resolution, but accesses the single cell level; PCR- and Sanger sequencing enable unique high resolution of genomes, while "classical" molecular genetic tests (like Southern-blot and restriction fragment length polymorphism analyses) aCGH (array-comparative genomic hybridization) have low to intermediate resolution. Molecular cytogenetics, together with its variant molecular combing, and all variants of next-generation sequencing approaches are the only two general techniques which can access the whole genome, from low to high resolution.

Diagnostic applications of molecular cytogenetics are already summarized elsewhere, and thus not further treated here; corresponding literature is listed in **Table 1**.

MOLECULAR CYTOGENETIC APPLICATIONS FOR CHROMOSOMIC RESEARCH IN THE CONCERT OF CYTOGENOMICS APPROACHES

Some of the approaches listed below are able to help in characterization of DNA-stretches of several to hundreds of base pair in length, while others are directed towards chromosomal subregions, bands, or whole chromosomes; some even give information on whole genome level. It must be admitted that here a subjective selection of research and diagnostic fields has been put together. This is necessary due to the sheer amount of the possible FISH-applications, and influenced by the focus of the author in human genetics field. Completely not covered are, e.g., molecular cytogenetic applications in plant-research (Lavania, 1998; Liehr, 2017) or microorganisms (Bensimon et al., 1994; Liehr, 2017). In the given examples, it will be highlighted that molecular cytogenetics (like next-generation sequencing approaches; **Figure 2**) is one of two cytogenomic approaches being able to analyze whole genomes from base pair to chromosomal levels.

For limitations of FISH in general, there are to mention, that 1) the resolution in standard FISH is limited to kilo-to megabasepair (except for molecular combing approach), and thus exact mapping of chromosomal breakpoints needs combined approaches like, e.g., applied in Jancuskova et al. (2013) or Moysés-Oliveira et al. (2019), 2) disease causing gene mutations at base pair level normally cannot be accessed by FISH—one exception was recently published (*Molecular Combing* Section) (Nguyen et al., 2019), 3) as in cytogenetics and next-generation sequencing (NGS) for qualified evaluation and interpretation of FISH-results experienced specialists are needed, 4) corresponding to question to be studied costs of consumables may be relatively high, and 5) number of probes being applied simultaneously is limited by number of fluorochromes and software; however, recent developments proved also solutions for this (Su et al., 2020).

Molecular Combing

The approach molecular combing refers to the physical combing of high molecular weight DNA on a glass surface. This approach was already suggested in 1994 and was deduced from what others published as fiber-FISH (Heng et al., 1992). However, molecular combing got a boost during last few years, as then molecular combing became also commercially available (Florian et al., 2019). This approach enables research on most stretched DNA-fibers; FISH probes can be hybridized and basic studies on DNA-replication, replication kinetics, but also for copy number variations of satellite sequences down to single nucleotide polymorphisms (SNPs) are possible and can be visualized (Florian et al., 2019). Besides, diagnostics for facio scapulohumeral muscular dystrophy (FSHD) became much more feasible, as by molecular combing the D4Z4 sequence in 4q35 and 10q26 can be clearly distinguished from each other (Nguyen et al., 2019). Much more breakthroughs from this high resolution FISH-approach are to be expected.

Chromosome Orientation-FISH (CO-FISH)

Chromosome orientation-FISH (CO-FISH) enables to selectively mark exactly one of the two homologous DNA strands of a chromosome. This is done by incorporation of 5-bromodeoxyuridine (BrdU) in one DNA strand and destroying it by UV-light and *EXOIII* enzyme treatment (the latter detects UV-induced gaps and starts degradation of DNA strand there) (Goodwin and Meyne, 1993). CO-FISH has been successfully applied to study orientation of repeated sequences or long unique DNA sequences by now (Liehr, 2017). The unique possibilities and advantages of this specialized FISH-approach have not been explored in full by scientific community, yet.

Telomeres Accessed by Q-FISH

Telomeres are important objects of research, as they are on the one hand known to be important for chromosome stability and also suggested to play a role in aging, cancer development but also apoptosis and senescence (Smith et al., 2019). Telomeres are low-copy repetitive elements, which are hard to access by molecular genetic approaches like sequencing. Telomere length can only be measured by few approaches. Available assays include 1) quantitative polymerase chain reaction, 2) terminal restriction fragment analysis, 3) telomere dysfunctional induced foci analysis, 4) single telomere length analysis, 5) telomere shortest length assay, and 6) quantitative FISH (Q-FISH). The latter is the only available *in situ* approach (Lai et al., 2018). It is even principally possible to do chromosome-specific telomere length studies by that technique.

Parental Origin Determination FISH (POD-FISH)

In 2001, the now well-known copy number variations (CNVs) were reported first for the human genome (Redon et al., 2006). Their detection was due to the, during that time in large scale studies applied approach aCGH—nowadays mostly referred to as CMA (chromosomal micro-array). CNVs, these previously undetectable structural variations of the human genome comprise losses, gains, insertions, and inversions in kilo-to mega-base-pair-range. CNVs of that size are accessible by FISH. Thus, it was logical to develop the following idea: these CNVs have an individual pattern along each chromosome and it is possible to use them, as before microsatellite markers, as markers to distinguish individual homologous chromosomes. When studying a trio (father, mother, and child) by microsatellite analyses, it is possible to follow up inheritance of chromosomes. Taking advantage of CNVs, the same can be done in trio-analyses of chromosome-preparations. Apart from uniparental disomy testing, by this approach (called parental origin determination FISH - POD-FISH) also the inheritance of individual chromosomes can be visualized. In microsatellite analyses, this is not possible as no individual chromosomes can be distinguished (Liehr, 2017; Weise et al., 2015).

Inter- and Intrachromosomal Interactions

The spatial organization of chromosomes in interphase nuclei, as well as the organization of metaphase chromosomes—which turned out to be not that different—is, as we know now, key to understand gene regulation (Daban, 2020). Inter- and intrachromosomal interactions can be studied in two ways: On the one hand, there are the high-throughput chromosome conformation capture (also abbreviated as high 3C or Hi-C) approaches, used for genome architecture mapping providing a multi-cell based genomic view. Such high-throughput, sequencing-based approaches have provided tremendously to our knowledge of genomic architecture, by giving contact information chromatin loci pairs. However, real 3D position information of individual alleles and/or loci cannot be deduced from this kind of data (Su et al., 2020). On the other hand, the three-dimensional genome organization can be studied on single cell level either by single-cell Hi-C or by imaging-based approaches. The latter enable spatial positioning of several chromatin loci at a time in single cells. Specifically, it is the FISH approach, which provides such characterizations in fixed cells. Living cells can since recently accessed by the clustered regularly interspaced short palindromic repeats

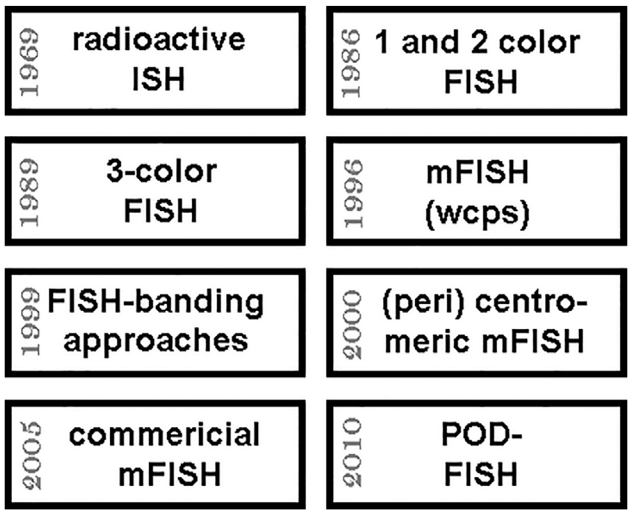


FIGURE 3 | Approximate timeline of important steps towards multicolor FISH (mFISH) is shown in the upper 4 boxes; the lower boxes give four further years when important mFISH-applications were kicked-off and/or became more and more available to research and diagnostics. Abbreviations: ISH = *in situ* hybridization; POD-FISH = parental origin determination FISH (see 4.4.); wcps = whole chromosome paints.

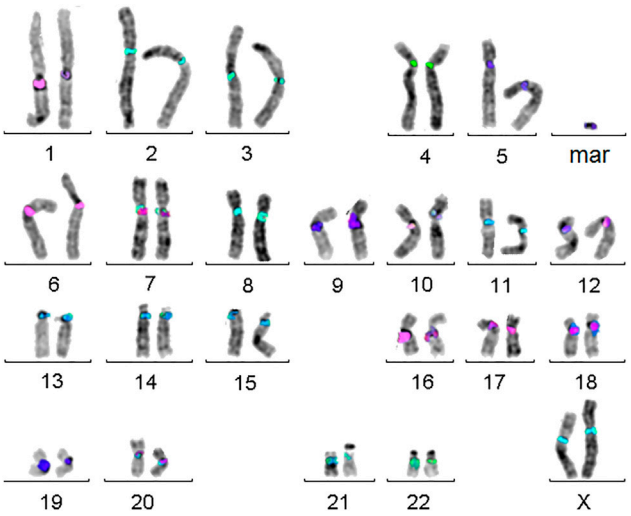


FIGURE 4 | Centromere-specific multicolor-FISH (cenM-FISH) was established to characterize small supernumerary marker chromosomes (sSMCs) for their chromosomal origin. As majority of sSMCs is derived from pericentric regions, an mFISH probe set consisting of probes specific for centromeres of chromosomes 1, 1/5/19, 2, 3, 4, 5/6, 7, 8, 9, 10, 11, 12, 13/21, 14/22, 15, 16, 17, 18, 20, 22, X, and Y is suited to characterize the here shown sSMC as a being derived from a chromosome 5.

(CRISPR) system (Su et al., 2020). Topologically associating domains (TADs) (Dixon et al., 2015) and related intra- and inter-chromosomal interactions (Maas et al., 2019) were recently identified by combining both complementing approaches. Especially here molecular cytogenetics is an indispensable research tool.

Multicolor-FISH in Research

Multicolor-FISH (mFISH) approaches and probe sets are applied—if reimbursed by the health systems somehow (Liehr, 2020) —in routine diagnostics and often independent of such issues in research; for review, see (Liehr, 2017; 2021). As mentioned under point 3, mFISH routine application in human genetics was initiated in 1996 by simultaneous painting of all 24 human chromosomes by whole chromosome probes applied in multiplex-FISH (M-FISH) and spectral karyotyping (SKY). Afterwards, countless mFISH assays have been established (Liehr, 2017; 2021). While most mFISH-probe sets for the characterization of the human genome were implemented primarily to study acquired or inherited chromosomal aberrations in diagnostics (see also point 4.7. below; **Figure 3**), others are pure research oriented. Specifically of interest are here the FISH-based chromosome-banding approaches (FISH-banding), like multicolor banding. Murine multicolor banding (mcb), for example, is used in studies in murine chromosome evolution or to characterize murine tumor cell lines (Liehr, 2021).

Research on Small Supernumerary Marker Chromosomes

Small supernumerary marker chromosomes (sSMCs) are a rare condition in human, resembling B-chromosomes in many other species. They can be found in ~3.3 million carriers worldwide, with ~2.2 million of them being asymptomatic. The remainders constitute a pool of patients with dozens of rare diseases. As also clinically normal sSMC carriers can have partial tri- or tetrasomies of euchromatic centromere–near regions they are as well in focus of research [for review (Liehr, 2021)]. As recently shown, molecular cytogenetics is the most straightforward approach to characterize sSMCs for their origin and genetic content, as sSMCs tend to be missed by molecular karyotyping or sequencing approaches due to their (low) mosaic and/or heterochromatic state (Liehr, 2021). The best suited approach to characterize sSMC's origin is the so-called centromere-specific multicolor-FISH (cenM-FISH) (Liehr, 2021)—an example for an sSMC derived from chromosome 5 characterized by cenM-FISH is shown in **Figure 4**.

Chromoanagenesis—Research

Complex chromosomal rearrangements and how they form has been studied for decades, applying cytogenetics and molecular cytogenetics (Heng et al., 2020). Besides, (molecular) cytogenetic studies already reported phenomena like single cells with extremely rearranged chromosomes and/or chromosome pulverization (Stephens et al., 2011). However, those results were widely ignored until they were "newly discovered" as chromothripsis in 2011 based on NGS studies (Stephens et al., 2011). Since then phenomena like chromothripsis, chromoanasynthesis and chromoplexy were subsumed under the term chromoanagenesis (for review see (Hattori and Fukami, 2020)). Meanwhile, there are more and more chromoanagenesis studies combining advantages of NGS and molecular cytogenetics (e.g., (Gu et al., 2016)).

Chromosomal Heteromorphisms and Repetitive DNA-Elements

Chromosomal heteromorphisms (CHMs), like length variants of acrocentric's short arms, are still exclusively accessible by cytogenetics and can be characterized in more detail only by molecular cytogenetics. These CHMs, consisting mainly of repetitive DNA-elements, like satellite DNAs, are definitely understudied. These genomic regions are widely ignored, and this is maybe best underlined by the fact that all in the 1980s characterized satellite DNAs known to be localized in the pericentric and/or heterochromatic regions of the human chromosomes are yet not included in any genomic browser. Their localizations and sizes are published, the probes like DXZ1 and DYZ3 are commercially available centromere-specific probes for chromosomes X and Y, and however, they remain unmentioned in the human genome browsers (Liehr, 2021).

FISH and Microdissection

Another underrated cytogenomic possibility is the application of glass-needle based chromosome microdissection (midi) for research (Maslova et al., 2015). Here, picogram of DNA can be taken directly from chromosomes and studied in multiple ways afterwards, including NGS approaches and others. Also prior FISH-labelled metaphases can be applied in midi, which can help to extract the correct (part of a) chromosome (Kosyakova et al., 2013).

Comet-FISH

Comet-assay is also a longstanding approach, leading to a bunch of new research possibilities if combined with molecular cytogenetics. "The comet assay is a rapid and very sensitive fluorescent microscopy-based method for measuring DNA damage, protection, and repair at the level of individual cells. In this assay, cells are embedded in agarose, lysed, and then electrophoresed. Negatively charged broken DNA strands exit from the lysed cell under the electric field and form a comet with "head" and "tail". The amount of DNA in the tail, relative to the head, is proportional to the amount of strand breaks. Results from the comet assay alone reflect only the level of overall DNA damage in single cells. The introduction of FISH in comet has allowed adding new abilities and to enhance resolution and validity of these two methods. FISH permitted to supplement potential of the comet assay with an opportunity to recognize genome regions of interest on comet images. The use of Comet-FISH will enable to achieve a higher sensitivity for the adequate hazard assessment of mutagens and will lead to a better understanding of the biological mechanisms involved" (Hovhannisyan, 201).

CRISPR-Mediated FISH-Applications

As already seen, molecular cytogenetics can be combined with multiple other approaches, which can lead to new possibilities to decipher multiple biological questions. The most recent advance is to combine FISH with the CRISPR/Cas9 system (Carroll, 2012); this can be done to get FISH-results in dead cells (Němečková et al., 2019), as well as to perform CRISPR-mediated live imaging, the latter allowing insights into living cells (Anton et al., 2014). Which new chromosomic research will become possible by these approaches has to be waited for.

CONCLUSION

Overall, it is still valid what Serakinci and Koelvraa, 2009 stated in 2009: "FISH techniques were originally developed as extra tools in attempts to map genes and a number of advances were achieved with this new technique. However, it soon became apparent that the FISH concept offered promising possibilities also in a number of other areas in biology and its use spread into new areas of research and also into the area of clinical diagnosis. In very general terms the virtues of FISH are in two areas of biology, namely genome characterization and cellular organization, function and diversity. To what extend FISH technology will be further developed and applied in new areas of research in the future remains to be seen" (Serakinci and Koelvraa, 2009).

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Al-Rikabi, A. B. H., Cioffi, M. B., and Liehr, T. (2019). Chromosome Microdissection on Semi-archived Material. *Cytometry A* 95, 1285–1288. doi:10.1002/cyto.a.23896

Anton, T., Bultmann, S., Leonhardt, H., and Markaki, Y. (2014). Visualization of Specific DNA Sequences in Living Mouse Embryonic Stem Cells with a Programmable Fluorescent CRISPR/Cas System. *Nucleus* 5, 163–172. doi:10.4161/nucl.28488

Arrighi, F. E., and Hsu, T. C. (1971). Localization of Heterochromatin in Human Chromosomes. *Cytogenet. Genome Res.* 10, 81–86. doi:10.1159/000130130

Bensimon, A., Simon, A., Chiffaudel, A., Croquette, V., Heslot, F., and Bensimon, D. (1994). Alignment and Sensitive Detection of DNA by a Moving Interface. *Science* 265, 2096–2098. doi:10.1126/science.7522347

Bernheim, A. (1999). Exploration du génome dans les proliférations malignes: de la cytogénétique à la cytogénomique. *Ann. Pathol.* 19, S1–S3.

Bottari, B., Ercolini, D., Gatti, M., and Neviani, E. (2006). Application of FISH Technology for Microbiological Analysis: Current State and Prospects. *Appl. Microbiol. Biotechnol.* 73, 485–494. doi:10.1007/s00253-006-0615-z

Boveri, T. (1902). Über mehrpolige Mitosen als Mittel zur Analys des Zellkerns. *Verhandlungen der physikalisch-medizinischen Gesellschaft Würzburg* 35, 67–90.

Carroll, D. (2012). A CRISPR Approach to Gene Targeting. *Mol. Ther.* 20, 1658–1660. doi:10.1038/mt.2012.171

Cheng, L., Zhang, S., Wang, L., MacLennan, G. T., and Davidson, D. D. (2017). Fluorescencein Situhybridization in Surgical Pathology: Principles and Applications. *J. Path: Clin. Res.* 3, 73–99. doi:10.1002/cjp2.64

Claussen, U. (2005). Chromosomics. *Cytogenet. Genome Res.* 111, 101–106. doi:10.1159/000086377

Cui, C., Shu, W., and Li, P. (2016). Fluorescence *In Situ* Hybridization: Cell-Based Genetic Diagnostic and Research Applications. *Front. Cel Dev. Biol.* 4, 89. doi:10.3389/fcell.2016.00089

Daban, J. R. (2020). Supramolecular Multilayer Organization of Chromosomes: Possible Functional Roles of Planar Chromatin in Gene Expression and DNA Replication and Repair. *FEBS Lett.* 594, 395–411. doi:10.1002/1873-3468.13724

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., et al. (2015). Chromatin Architecture Reorganization during Stem Cell Differentiation. *Nature* 518, 331–336. doi:10.1038/nature14222

Ferguson-Smith, M. A., Yang, F., Rens, W., and O'Brien, P. C. M. (2005). The Impact of Chromosome Sorting and Painting on the Comparative Analysis of Primate Genomes. *Cytogenet. Genome Res.* 108, 112–121. doi:10.1159/000080809

Flemming, W. (1878). Beiträge zur Kenntniss der Zelle und ihrer Lebenserscheinungen. *Arch. Mikroskop. Anat.* 16, 302–436.

Florian, R. T., Kraft, F., Kraft, F., Leitão, E., Kaya, S., Klebe, S., et al. (2019). Unstable TTTTA/TTTCA Expansions in MARCH6 Are Associated with Familial Adult Myoclonic Epilepsy Type 3. *Nat. Commun.* 10, 4919. doi:10.1038/s41467-019-12763-9

Goodpasture, C., and Bloom, S. E. (1975). Visualization of Nucleolar Organizer Regions in Mammalian Chromosomes Using Silver Staining. *Chromosoma* 53, 37–50. doi:10.1007/bf00329389

Goodwin, E., and Meyne, J. (1993). Strand-specific FISH Reveals Orientation of Chromosome 18 Alphoid DNA. *Cytogenet. Cel Genet.* 63, 126–127. doi:10.1159/000133516

Gu, S., Szafranski, P., Akdemir, Z. C., Yuan, B., Cooper, M. L., Magriñá, M. A., et al. (2016). Mechanisms for Complex Chromosomal Insertions. *Plos Genet.* 12, e1006446. doi:10.1371/journal.pgen.1006446

Hattori, A., and Fukami, M. (2020). Established and Novel Mechanisms Leading to De Novo Genomic Rearrangements in the Human Germline. *Cytogenet. Genome Res.* 160, 167–176. doi:10.1159/000507837

Heng, H. H. Q., Regan, S. M., Liu, G., and Ye, C. J. (2016). Why it Is Crucial to Analyze Non Clonal Chromosome Aberrations or NCCAs?. *Mol. Cytogenet.* 9, 15. doi:10.1186/s13039-016-0223-2

Heng, H. H., Squire, J., and Tsui, L. C. (1992). High-resolution Mapping of Mammalian Genes by *In Situ* Hybridization to Free Chromatin. *Proc. Natl. Acad. Sci.* 89, 9509–9513. doi:10.1073/pnas.89.20.9509

Hochstenbach, R., van Binsbergen, E., Schuring-Blom, H., Buijs, A., and Ploos van Amstel, H. K. (2019). A Survey of Undetected, Clinically Relevant Chromosome Abnormalities when Replacing Postnatal Karyotyping by Whole Genome Sequencing. *Eur. J. Med. Genet.* 62, 103543. doi:10.1016/j.ejmg.2018.09.010

Hopman, A. H. N., Wiegant, J., Raap, A. K., Landegent, J. E., van der Ploeg, M., and van Duijn, P. (1986). Bi-color Detection of Two Target DNAs by Non-radioactive *In Situ* Hybridization. *Histochemistry* 85, 1–4. doi:10.1007/bf00508646

Hovhannisyan, G. G. (2010). Fluorescence *In Situ* Hybridization in Combination with the Comet Assay and Micronucleus Test in Genetic Toxicology. *Mol. Cytogenet.* 3, 17. doi:10.1186/1755-8166-3-17

Jancuskova, T., Plachy, R., Stika, J., Zemankova, L., Hardekopf, D. W., Liehr, T., et al. (2013). A Method to Identify New Molecular Markers for Assessing Minimal Residual Disease in Acute Leukemia Patients. *Leuk. Res.* 37, 1363–1373. doi:10.1016/j.leukres.2013.06.009

Kallioniemi, A., Kallioniemi, O., Sudar, D., Rutovitz, D., Gray, J., Waldman, F., et al. (1992). Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors. *Science* 258, 818–821. doi:10.1126/science.1359641

Koch, J., Hindkjær, J., Kølvraa, S., and Bolund, L. (1995). Construction of a Panel of Chromosome-specific Oligonucleotide Probes (PRINS-Primers) Useful for the Identification of Individual Human Chromosomes *In Situ*. *Cytogenet. Cel Genet* 71, 142–147. doi:10.1159/000134094

Kosyakova, N., Hamid, A. B., Chaveerach, A., Pinthong, K., Siripiyasing, P., Supiwong, W., et al. (2013). Generation of Multicolor Banding Probes for Chromosomes of Different Species. *Mol. Cytogenet.* 6, 6. doi:10.1186/1755-8166-6-6

Lai, T.-P., Wright, W. E., and Shay, J. W. (2018). Comparison of Telomere Length Measurement Methods. *Phil. Trans. R. Soc. B* 373, 20160451. doi:10.1098/rstb.2016.0451

Lavania, U. C. (1998). Fluorescence *In Situ* Hybridization in genome., Chromosome and Gene Identification in Plants. *Curr. Sci. India.* 74, 126–133.

Liehr, T. (2021)Chromosomic Databases. Available at: http://cs-tl.de/DB.html (Accessed 05 20, 2021).

Liehr, T. (2017). *Fluorescence* in Situ *Hybridization (FISH) – Application Guide*. 2nd ed. Berlin., Germany: Springer.

Liehr, T. (2019). From Human Cytogenetics to Human Chromosomics. *Ijms* 20, 826. doi:10.3390/ijms20040826

Liehr, T., and Hamid Al-Rikabi, A. B. (2018). Impaired Spermatogenesis Due to Small Supernumerary Marker Chromosomes: The Reason for Infertility Is Only Reliably Ascertainable by Cytogenetics. *Sex. Dev.* 12, 281–287. doi:10.1159/000491870

Liehr, T. (2020). Is Molecular Cytogenetic Diagnostics of Rare Diseases in Europe Close to Extinction. *J. Genet. Genomics* 4, 2.

Liehr, T., Othman, M. A., Rittscher, K., and Alhourani, E. (2015). The Current State of Molecular Cytogenetics in Cancer Diagnosis. *Expert Rev. Mol. Diagn.* 15, 517–526. doi:10.1586/14737159.2015.1013032

Maass, P. G., Barutcu, A. R., and Rinn, J. L. (2019). Interchromosomal Interactions: A Genomic Love story of Kissing Chromosomes. *J. Cel. Biol.* 218, 27–38. doi:10.1083/jcb.201806052

Majka, J., Majka, M., Kwiatek, M., and Wiśniewska, H. (2017). Similarities and Differences in the Nuclear Genome Organization within Pooideae Species Revealed by Comparative Genomic *In Situ* Hybridization (GISH). *J. Appl. Genet.* 58, 151–161. doi:10.1007/s13353-016-0369-y

Maslova, A., Zlotina, A., Kosyakova, N., Sidorova, M., and Krasikova, A. (2015). Three-dimensional Architecture of Tandem Repeats in Chicken Interphase Nucleus. *Chromosome Res.* 23, 625–639. doi:10.1007/s10577-015-9485-5

McGowan-Jordan, J., Hastings, R., and Moore, S. (2020). *ISCN 2020 - an International System for Human Cytogenomic Nomenclature*. Basel., Switzerland: Karger, 6.

Mendel, G. (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn* 4, 3–47. doi:10.5962/bhl.title.61004

Moysés-Oliveira, M., Di-Battista, A., Zamariolli, M., Meloni, V. A., Bragagnolo, S., Christofolini, D. M., et al. (2019). Breakpoint Mapping at Nucleotide Resolution in X-Autosome Balanced Translocations Associated with Clinical Phenotypes. *Eur. J. Hum. Genet.* 27, 760–771. doi:10.1038/s41431-019-0341-5

Nederlof, P. M., Robinson, D., Abuknesha, R., Wiegant, J., Hopman, A. H., Tanke, H. J., et al. (1998). Three-color Fluorescence *In Situ* Hybridization for the Simultaneous Detection of Multiple Nucleic Acid Sequences. *Cytometry* 10, 20–27. doi:10.1002/cyto.990100105

Nguyen, K., Broucqsault, N., Chaix, C., Roche, S., Robin, J. D., Vovan, C., et al. (2019). Deciphering the Complexity of the 4q and 10q Subtelomeres by Molecular Combing in Healthy Individuals and Patients with Facioscapulohumeral Dystrophy. *J. Med. Genet.* 56, 590–601. doi:10.1136/jmedgenet-2018-105949

Nguyen, K., Walrafen, P., Bernard, R., Attarian, S., Chaix, C., Vovan, C., et al. (2011). Molecular Combing Reveals Allelic Combinations in Facioscapulohumeral Dystrophy. *Ann. Neurol.* 70, 627–633. doi:10.1002/ana.22513

Němečková, A., Wäsch, C., Schubert, V., Ishii, T., Hřibová, E., and Houben, A. (2019). CRISPR/Cas9-based RGEN-ISL Allows the Simultaneous and Specific Visualization of Proteins, DNA Repeats, and Sites of DNA Replication. *Cytogenet. Genome Res.* 159, 48–53.

Pellestor, F., Anahory, T., Lefort, G., Puechberty, J., Liehr, T., Hédon, B., et al. (2011). Complex Chromosomal Rearrangements: Origin and Meiotic Behavior. *Hum. Reprod. Update* 17, 476–494. doi:10.1093/humupd/dmr010

Pinkel, D., Gray, J. W., Trask, B., van den Engh, G., Fuscoe, J., and van Dekken, H. (1986) Cytogenetic Analysis by *In Situ* Hybridization with Fluorescently Labeled Nucleic Acid Probes. *Cold Spring Harb. Symp. Quant. Biol.*, 51 Pt 1 Pt 1, 151–157. doi:10.1101/sqb.1986.051.01.018

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., et al. (1998). High Resolution Analysis of DNA Copy Number Variation Using Comparative Genomic Hybridization to Microarrays. *Nat. Genet.* 20, 207–211. doi:10.1038/2524

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global Variation in Copy Number in the Human Genome. *Nature* 444, 444–454. doi:10.1038/nature05329

Sabile, A., Poras, I., Cherif, D., Goodfellow, P., and Avner, P. (1997) Isolation of monochromosomal hybrids for mouse chromosomes 3, 6, 10, 12, 14, and 18. *Mamm. Genome.*, 8, 81–85. doi:10.1007/s003359900362

Sala, E., Conconi, D., Crosti, F., Villa, N., Redaelli, S., and Roversi, G. (2019). Interphase FISH: A Helpful Assay in Prenatal Cytogenetics Diagnosis. *OBM Genet.* 3, 1901063. doi:10.21926/obm.genet.1901063

Salman, M., Jhanwar, S., and Ostrer, H. (2004). Will the New Cytogenetics Replace the Old Cytogenetics?. *Clin. Genet.* 66, 265–275. doi:10.1111/j.1399-0004.2004.00316.x

Schlegelberger, B. (2013), In Memoriam: Prof. Dr. Rer. Nat. Dr. Med. h.C. Lore Zech; 24.9.1923 - 13.3.2013: Honorary Member of the European Society of Human Genetics, Honorary Member of the German Society of Human Genetics, Doctor Laureate, the University of Kiel, Germany. *Mol. Cytogenet.*, 6, 20. doi:10.1186/1755-8166-6-20

Schröck, E., du Manoir, S., Veldman, T., Schoell, B., Wienberg, J., Ferguson-Smith, M. A., et al. (1996). Multicolor Spectral Karyotyping of Human Chromosomes. *Science* 273, 494–497. doi:10.1126/science.273.5274.494

Seabright, M. (1971). A Rapid Banding Technique for Human Chromosomes. *The Lancet* 298, 971–972. doi:10.1016/s0140-6736(71)90287-x

Serakinci, N., and Koelvraa, S. (2009). "Molecular Cytogenetic Applications in Diagnostics and Research – an Overview," in *Fluorescence* in Situ *Hybridization (FISH) – Application Guide*. Editor T. Liehr (Berlin: Springer), 3–21.

Silva, M., de Leeuw, N., Mann, K., Schuring-Blom, H., Morgan, S., Giardino, D., et al. (2919). European Guidelines for Constitutional Cytogenomic Analysis. *Eur. J. Hum. Genet.* 27, 1–16. doi:10.1038/s41431-018-0244-x

Smith, C. M., Ma, N. S., Nowak, N. J., Shows, T. B., and Gerhard, D. S. (1997). A 3-Mb Contig fromD11S987toMLK3,a Gene-Rich Region in 11q13: Figure 1. *Genome Res.* 7, 835–842. doi:10.1101/gr.7.8.835

Smith, L., Luchini, C., Demurtas, J., Soysal, P., Stubbs, B., Hamer, M., et al. (2019). Telomere Length and Health Outcomes: An Umbrella Review of Systematic Reviews and Meta-Analyses of Observational Studies. *Ageing Res. Rev.* 51, 1–10. doi:10.1016/j.arr.2019.02.003

Speicher, M. R., Ballard, S. G., and Ward, D. C. (1996). Karyotyping Human Chromosomes by Combinatorial Multi-Fluor FISH. *Nat. Genet.* 12, 368–375. doi:10.1038/ng0496-368

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., et al. (2011). Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* 144, 27–40. doi:10.1016/j.cell.2010.11.055

Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B., and Zhuang, X. (2020). Genome-scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* 182, 1641–1659. e26. doi:10.1016/j.cell.2020.07.032

Sutton, W. S. (1903). The Chromosomes in Heredity. *Biol. Bull.* 4, 231–250. doi:10.2307/1535741

Waldeyer, W. (1888). Ueber Karyokinese und ihre Beziehungen zu den Befruchtungsvorgängen. *Archiv F. Mikrosk. Anatomie* 32, 1–122. doi:10.1007/bf02956988

Weise, A., and Liehr, T. (2008). Fluorescencein Situhybridization for Prenatal Screening of Chromosomal Aneuploidies. *Expert Rev. Mol. Diagn.* 8, 355–357. doi:10.1586/14737159.8.4.355

Weise, A., Othman, M. A. K., Bhatt, S., Löhmer, S., and Liehr, T. (2015). Application of BAC-Probes to Visualize Copy Number Variants (CNVs). *Methods Mol. Biol.* 1227, 299–307. doi:10.1007/978-1-4939-1652-8_16

Zhang, C., Cerveira, E., Rens, W., Yang, F., and Lee, C. (2018). Multicolor Fluorescence *In Situ* Hybridization (FISH) Approaches for Simultaneous Analysis of the Entire Human Genome. *Curr. Protoc. Hum. Genet.* 99, e70. doi:10.1002/cphg.70

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Liehr. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

MINI REVIEW
published: 22 November 2021
doi: 10.3389/fgene.2021.766485

Genetic Analysis for Coronary Artery Disease Toward Diverse Populations

*Kazuo Miyazawa and Kaoru Ito**

Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

Coronary artery disease is one of the leading causes of death in the world, and as such, it is one of the diseases for which genetic analyses have been actively conducted. In the early days, analyses of families with the aggregation of early-onset myocardial infarction, such as those with familial hypercholesterolemia, was the main focus, but since the practical application of genome-wide association study, the analysis of coronary artery disease as a common disease has progressed, and many disease-susceptibility loci have been identified. In addition, with the advancement of technologies, it has become possible to identify relatively rare genetic variants in a population-based analysis. These advances have not only revealed the detailed disease mechanisms but have also enabled the quantification of individual genetic risk and the development of new therapeutic agents. In this paper, some of those items, which are important to know in the current genetic analyses for coronary artery disease, are discussed.

Keywords: coronary artery disease, genome-wide association study, population diversity, biobank, precision medicine

OPEN ACCESS

Edited by:
Jordi Pérez-Tur, Consejo Superior de Investigaciones Científicas (CSIC), Spain

Reviewed by:
Nirmala Dushyanthi Sirisena, University of Colombo, Sri Lanka
Jaideep Mahendra, India

****Correspondence:***
Kaoru Ito
kaoru.ito@riken.jp

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: *29 August 2021*
Accepted: *08 November 2021*
Published: *22 November 2021*

Citation:
Miyazawa K and Ito K (2021) Genetic Analysis for Coronary Artery Disease Toward Diverse Populations. Front. Genet. 12:766485. doi: 10.3389/fgene.2021.766485

INTRODUCTION

Coronary artery disease (CAD) causes myocardial ischemia due to narrowing or blockage of the coronary arteries that feed the heart based on atherosclerotic predisposition, leading to myocardial infarction (MI). The disease can also lead to the development of arrhythmia, heart failure, and death. Epidemiologically, CAD is one of the leading causes of death in the world, affecting approximately 1.72% of the world's population, with nine million deaths reported in 2017 (Khan et al., 2020). Such high morbidity and mortality rates have a significant impact in terms of medical economic and social burden, which urgently needs to be addressed. Therefore, genetic studies for CAD and attempts to apply the results to clinical practice are being spearheaded around the world.

Advances in the Genetic Analyses for Coronary Artery Disease

The history of genetic research on CAD began with the analysis of families that suffered early-onset MI. In the familial analysis, the method called linkage analysis was mainly used, which assumes that the disease is caused by a single, high-penetrance genetic mutation and is transmitted from generation to generation according to Mendelian laws. Diseases with this characteristic are also called monogenic disorders, but this form of inheritance has not been observed in many patients with CAD and only explains some of the genetic factors. However, family-based analysis of the disease was a very effective tool not only for elucidating the pathogenesis of the disease but also for drug discovery, because it allowed us to observe a clear phenotype leading to the development of the disease and to identify the causative genes. One of the earliest reports of such family-based analysis was of a patient with familial hypercholesterolemia (FH) who had early-onset MI as well as abnormally high cholesterol levels and xanthoma. The discovery of such families with a *LDLR*

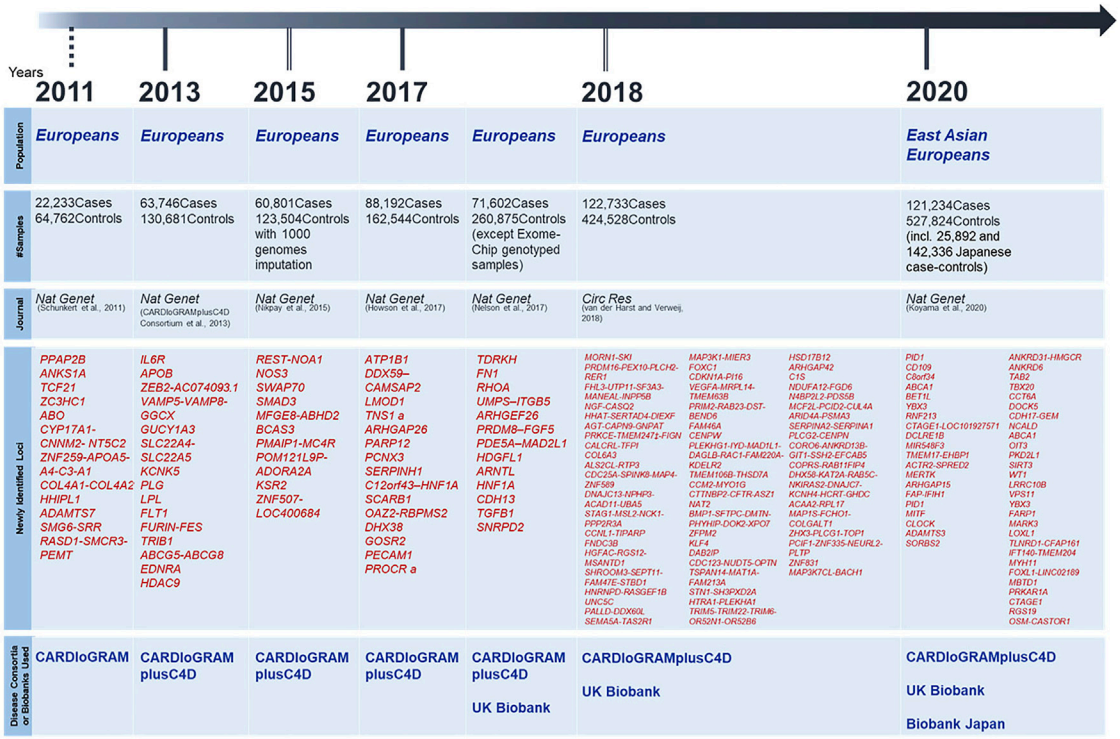


FIGURE 1 | Landmark papers of CAD-GWAS and Newly Identified Disease Susceptibility Loci Since 2010.

gene mutation led to the subsequent identification of genes associated with FH (Vrablik et al., 2020), including *APOB*, *PCSK9*, *LDLRAP1*, *ABCG5*, and *ABCG8*. In addition, *LRP6* gene mutations associated with LDL cholesterol, triglycerides, hypertension, diabetes mellitus, and osteoporosis (Mani et al., 2007) and *DYRK1B* gene mutations associated with obesity, severe hypertension, and diabetes mellitus (Keramati et al., 2014) have been reported as familial-onset MI genes that are different from FH and cause MI at a young age.

On the other hand, since the majority of patients with CAD do not show clear Mendelian heredity, it was necessary to conduct a genetic analysis of these patients using another method. Although the linkage analysis did not work well for such samples, Tanaka, Ozaki and their colleagues were the first in the world to conduct a genome-wide association study (GWAS) using the information on approximately 90,000 single nucleotide polymorphisms (SNPs), followed by identifying a disease-susceptibility locus (Ozaki et al., 2002). The SNP identified in the study was on the lymphotoxin-alpha (*LTA*) gene located in the HLA region of chromosome 6q21, which encodes a pro-inflammatory cytokine. Subsequently, several SNPs related to inflammation were identified, strongly suggesting that the inflammatory cascade was very important in the pathogenesis of CAD. In the early days, GWAS were performed on a scale of a few hundred individuals or less, but with the emergence of disease consortia and huge biobanks, the sample size has rapidly increased. Looking at landmark papers from the last decade, as shown in **Figure 1**, the study conducted by CARDIoGRAM consortium in 2011 comprised approximately 90,000 individuals (Schunkert et al., 2011), the study conducted by CARDIoGRAMplusC4D consortium in 2013 comprised approximately 190,000 individuals (Consortium et al., 2013), and Nikpay et al. performed a GWAS in 2015 using the whole genome sequencing (WGS) data from the 1,000 Genomes Project as a reference panel for imputation to test about 10 million variants (Nikpay et al., 2015). In 2017, Howson et al. (2017) encompassed about 250,000 samples while Nelson et al.'s study included approximately 320,000 samples (excluding exome-chip analysis samples) (Nelson et al., 2017), and van et al.'s study in 2018 increased the sample size to about 400,000 (van der Harst and Verweij, 2018). However, most of the samples in those studies so far have been collected from European populations. The ethnic specificity of genetic architecture has already been discussed, and considering the clinical application of genetic analysis results, each ethnic group needs to have its own evidence of genetic research to rely on. In 2017, Lu et al. (2017) analyzed 47,532 East Asians for lipid levels, which are heritable risk factors for CAD, using exome arrays and identified three chip-wide disease susceptibility loci. They further compared these results with a GWAS of 28,899 Chinese subjects with CAD and a GWAS of approximately 190,000 subjects in the CARDIoGRAMplusC4D consortium, and found rs7901016, a non-coding variant near *MCU* gene, lower LDL cholesterol level and reduce the risk of CAD. In 2020, a Japanese group conducted a GWAS using approximately 180,000 Japanese and identified eight disease-susceptibility loci that had not been identified in larger

European studies, which demonstrated the importance of genetic analysis in non-European populations. In addition, they conducted a meta-analysis with European GWAS with a total of 650,000 subjects (Koyama et al., 2020), which is the largest scale analysis in the world so far, and identified 35 novel disease-susceptibility loci. As such, like the situation of GWAS for other diseases, genetic studies for CAD have shown a rapid increase in sample size and a concomitant rapid increase in the number of disease-susceptibility loci.

Chasing Rare Variants in the Population-Level Analysis

The existence of inexpensive genotyping arrays for GWAS of CAD makes it possible to increase the sample size, and so far such attempts have been very successful. However, genotyping arrays have a limited number of probes and the genomic information that can be obtained at a time is limited. On the other hand, family-based analysis identifies rare genetic variants with strong effects, which are difficult to capture with genotyping arrays targeting common variants. Genetic variants detected in family-based analysis are very rare, and sequencing was necessary to detect such rare variants at the population level. However, the identification of rare variants continued to be difficult due to the high cost and difficulty in increasing the number of samples compared to genotyping arrays. Thus, there is a large gap between the search for common and rare variants, and the identification of rare variants at the population level has been considered to be difficult.

One of the solutions is to perform WGS on all samples, although the cost is much higher. Currently, the only large-scale WGS project is the Million Veterans Program's 100,000 (as of February 2021: https://www.darkdaily.com/vas-million-veterans-program-research-study-receives-its-100000th-human-genome-sequence/), and the TOPMed project led by the U.S. National Heart, Lung, and Blood Institute (160,000 people as of Freeze 8, 2021: https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8), etc.

On the other hand, there is a method called imputation that uses haplotype information obtained from WGS as a means of obtaining genomic information that is not as sparse in terms of variant information as genotyping arrays, although it does not have the near-complete coverage of WGS. In this method, haplotype information from WGS is used as a reference panel, and variant information typed by genotyping arrays is used as a scaffold to predict untyped regions. For a long time, the WGS information of 2,504 individuals (Phase 3) of the 1,000 Genomes Project was used as the golden standard as the reference panel for imputation. However, since the panel covered Europeans, Americans, Africans, and East Asians equally, it did not have enough information on haplotypes specific to each ethnic group (just several hundred individuals for each group) to impute variants with low allele frequencies. In Europe, the Haplotype Reference Consortium (HRC) was launched, whose imputation reference panel using the WGS information of about 65,000 people has been established since the initial goal. Currently, more precise imputation is possible by incorporating the WGS data of the TOPMed project mentioned above.

However, because of the ethnic specificity of haplotype information, there is little advantage for other ethnic groups to use the HRC, which is mainly used by Europeans. In a recent landmark paper on GWAS for CAD (Koyama et al., 2020), they not only used the WGS data of 4,930 Japanese but also created a reference panel containing disease-specific haplotype information for 1,782 patients with CAD for imputation. Due to the advantageous conditions such as the unity of the Japanese population and disease specificity, their imputation was able to include variants with allele frequencies of about 0.02% in the analysis with less than one-tenth of the WGS data samples of HRC. As a result, they identified a missense mutation in the *RNF213* gene, which has been reported as a causative gene of Moyamoya disease, for the first time in the population analysis of CAD. In addition, they identified a group of genetic variants that cause FH, including protein-coding variants in the *LDLR*, *PCSK9*, and *APOB* genes. The magnitude of the effect on the onset of the disease and the effect on the age of disease onset were also shown at the population level. This identification and characterization of variants previously identified in family-based analyses at the population-level using genotyping arrays prove that population-level analysis is now able to approach rare variants that were previously within the scope of genealogical analysis. Thus, imputation is a cost-effective method that is still frequently used to reuse genotyped array samples but will be replaced by WGS in the not-too-distant future.

From Genetic Variants to Functional Analysis

In contrast to monogenic disorders, disease-associated SNPs detected by GWAS are often located in non-protein-coding regions, and the presence of linkage disequilibrium (LD) blocks means that lead SNPs are not always the cause of the disease. It is not easy to infer the exact mechanism leading to the disease phenotype from such SNPs. For example, the chromosome 9p21 region, which is a locus particularly strongly associated with CAD, does not contain protein-coding genes, making it difficult to elucidate the mechanism leading to the disease development. Although candidate genes such as *ANRIL* (Lo Sardo et al., 2018), a long non-coding RNA, and *CDKN2A* and *CDKN2B* (Kojima et al., 2014), which are located in the vicinity of *ANRIL*, were identified in several wet experiments, no conclusion has been reached as to which gene or mechanism is responsible. Chromosome 6p24 is also a disease susceptibility locus strongly associated with CAD, carotid artery dissection, and hypertension, and recent omics analyses have reported several disease pathogenic mechanisms. 6p24 has two candidate genes, phosphatase and action regulator 1 gene (*PHACTR1*) and endothelin-1 gene (*EDN1*). Although rs9349379, the lead SNP of this region, is located in the intron region of *PHACTR1*, the effect of PHACTR1 protein on the pathogenesis of CAD was unclear since the past molecular biological studies. On the other hand, EDN1 protein was well known as a protein involved in atherosclerosis. Expression quantitative trait loci analysis (eQTL

analysis) using vascular tissue showed a strong correlation between rs9349379 and *PHACTR1* gene expression (Beaudoin et al., 2015). On the other hand, when rs9349379 was introduced into induced Pluripotent Stem Cells (iPSC)-derived vascular-like cells using genome editing technology, a strong correlation was observed between rs9349379 and the expression of the *EDN1* gene, suggesting that *EDN1* is the causative gene (Gupta et al., 2017). Thus, conflicting results were obtained depending on the experimental system, and it is necessary to accumulate more evidence to determine which of these genes is involved in the disease development, or both. Thus, the number of causative genes and their independence in a single disease-susceptible locus are often unknown, and even if immortalized human cultured cells or iPSC are used as a human model, and their detection is difficult due to their small effect. In addition, repeating such experiments for each individual cell would require a large amount of monetary and time resources. Therefore, there is an urgent need to establish a fast and multiplex experimental system that can reproduce the human environment and detect the effects of many genetic variants at once.

Realization of Precision Medicine From Genomic Analysis

Genetic risk identified by genomic analysis is a risk factor to which we are constantly exposed from birth to death, unlike lifestyle risk that becomes apparent after we reach adulthood. If this information can be used to predict the disease onset, assess the disease severity, and intervene in treatment, it will be a great help in achieving precision medicine. Various studies/researches for this purpose are currently being actively conducted around the world. Unlike monogenic disorders, in which a single genetic variant can be a definitive marker for the diagnosis, CAD is mainly developed by the accumulation of many common variants with weak effects and high frequency. To predict the genetic risk of such a disease, it was necessary to integrate information from multiple genetic markers, and the genetic risk score (GRS) was devised for this purpose. The GRS is calculated as the sum of (the number of an alternative allele multiplied by the amount of effect of a variant on the disease onset as estimated by GWAS), using the lead variants that meet the genome-wide significance level in GWAS as markers. Recently, given the fact that including more variants improves the performance (to some extent), a polygenic risk score (PRS) has been developed, in which thousands to tens of millions of variants are included. Since GRS is a concept that includes PRS, PRS will also be referred to simply as GRS in this paper.

A landmark paper demonstrating the usefulness of CAD-GRS was reported in 2016, which examined the contrast with lifestyle risk (Khera et al., 2016). This paper showed that the hazard ratio of developing coronary events in the top 20% of the GRS group was similar to that in the group of patients with unfavorable lifestyle habits. The paper also showed that even with a high GRS, a healthy lifestyle was able to offset the genetic risk (45% relative risk reduction), indicating not only the usefulness of the GRS but also that aggressive intervention was effective in reducing coronary events in patients with a high GRS. The next important paper demonstrating the usefulness of GRS was published in 2018 (Inouye et al., 2018). The authors proposed a new framework to improve the performance of GRS, where they showed that the derived GRS was better than existing clinical risk factors in predicting the CAD onset and that the integration of GRS and clinical risk scores further improved the performance. Finally, the therapeutic utility of the GRS for CAD was demonstrated in a paper (Kullo et al., 2016), where the authors followed LDL cholesterol levels, considered one of the strongest risk factors for the disease, in 203 patients at intermediate risk for CAD who were not taking lipid-lowering drugs such as statins, in two groups: one with the clinical risk score and the other with both the clinical risk score and the GRS. After 6 months, LDL cholesterol levels were significantly lower in the group that received both scores, and there was a particularly strong reduction in LDL cholesterol levels in the group with the high GRS. Although this is not a study of a protocol that mandates therapeutic intervention according to the GRS, it is an example of how the presentation of this information to patients and their physicians could have a strong influence on subsequent treatment decisions.

Although the CAD-GRS is expected to be helpful in clinical practice, there are concerns about its applications. The study design of the GWAS from which the GRS is derived influences the characteristics of the GRS. This could be the definition of cases and controls, or it could be the issue of population specificity. Regarding the former, a position paper published in Nature 2021 calls for transparency and disclosure of information in the derivation, performance evaluation, and validation of the GRS (Wand et al., 2021). In fact, regarding the additive effect of CAD-GRS, Mosley et al. published a paper demonstrating no improvement in the performance by adding CAD-GRS to a clinical risk score (Mosley et al., 2020), contrary to the previous report (Inouye et al., 2018). Therefore, the use of GRS should be carefully evaluated to ensure that an appropriate GRS is being used and that the evaluation is fair. Concerning the latter issue of ethnic specificity, the current situation that European studies dominate the majority of academic journals gives us few choices other than using European GWAS for the GRS development. Thus, the importance of genomic analysis for other populations is now being emphasized. On the other hand, a Japanese group reported that utilizing a trans-ancestry meta-analysis with European GWAS as a GRS-derivation GWAS improved the performance of CAD-GRS, which implied that non-European populations could make use of the abundant European GWAS results for their GRS development utilizing trans-ancestry meta-analysis. As such, GRS is awaiting not only the improvement of its performance but also the emergence of new methods to deal with various populations.

From Genome Analysis to Therapy

Genomic analysis can identify disease-related genes, and the results are expected to lead to therapeutic applications; causative genes are inferred from disease-susceptibility loci in GWAS, and drug repositioning has been performed using gene-drug interaction databases (Okada et al., 2014). However, genes

derived from single-gene diseases, which provide clear information on the pathogenesis of the disease and its accompanying symptoms, are more likely to reach the clinical stage because their efficacy and adverse reactions are easier to predict when targeted for therapy than genes identified by GWAS. A recent representative example is a human monoclonal antibody against the *PCSK9* gene identified in the genomic analysis of FH, which shows a strong lipid-lowering effect. In addition, based on a report of a family lineage with mutations in the *ANGPTL3* gene (Musunuru et al., 2010) that showed low LDL cholesterol, low HDL cholesterol, and low triglycerides, antibodies (Dewey et al., 2017) or antisense oligonucleotides (Graham et al., 2017) targeting the gene are being developed (the similar trials are being performed for the *ANGPTL4* gene). In addition, although not limited to CAD, CRISPR/Cas9, which enables easier gene modification, or methods that avoid DNA double-strand break and reduce off-target effects, and improve safety, have recently emerged.

DISCUSSION

Genomic research on CAD is steadily expanding the horizon, with the help of the emergence of huge biobanks as well as the activities of powerful disease consortia. Furthermore, CAD-PRS is beginning to pave the way for clinical applications. Because of the various risk factors like so-called lifestyle-related diseases such as hypertension, diabetes, and dyslipidemia, and because of the several sub-forms such as stable angina, acute coronary syndrome, and vasospastic angina, the mechanisms of the disease development are expected to be very diverse, and there are still many issues to be addressed for further research. However, as mentioned, populations other than Europeans are underrepresented, and it will be necessary to pay more attention to population diversity as well as to the scale and technology of the research.

In the near future, WGS will play a leading role in the genetic analysis of CAD, and as the corresponding information in other omics layers becomes more complete, CAD genome research will continue to develop. It is hoped that the results will lead to the clinical implementation of precision medicine for CAD.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

KM is funded by the Japan Society for the Promotion of Science. KI is funded by Japan Agency for Medical Research and Development under Grant Numbers JP20km0405209, and JP20ek0109487.

REFERENCES

Beaudoin, M., Gupta, R. M., Won, H.-H., Lo, K. S., Do, R., Henderson, C. A., et al. (2015). Myocardial Infarction-Associated SNP at 6p24 Interferes with MEF2 Binding and Associates with PHACTR1 Expression Levels in Human Coronary Arteries. *Atvb* 35, 1472–1479. doi:10.1161/atvbaha.115.305534

Consortium, C. A. D., Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., et al. (2013). Large-scale Association Analysis Identifies New Risk Loci for Coronary Artery Disease. *Nat. Genet.* 45, 25–33. doi:10.1038/ng.2480

Dewey, F. E., Gusarova, V., Dunbar, R. L., O'Dushlaine, C., Schurmann, C., Gottesman, O., et al. (2017). Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. *N. Engl. J. Med.* 377, 211–221. doi:10.1056/nejmoa1612790

Graham, M. J., Lee, R. G., Brandt, T. A., Tai, L.-J., Fu, W., Peralta, R., et al. (2017). Cardiovascular and Metabolic Effects of ANGPTL3 Antisense Oligonucleotides. *N. Engl. J. Med.* 377, 222–232. doi:10.1056/nejmoa1701329

Gupta, R. M., Hadaya, J., Trehan, A., Zekavat, S. M., Roselli, C., Klarin, D., et al. (2017). A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell* 170, 522–533. doi:10.1016/j.cell.2017.06.049

Howson, J. M. M., Zhao, W., Zhao, W., Barnes, D. R., Ho, W.-K., Young, R., et al. (2017). Fifteen New Risk Loci for Coronary Artery Disease Highlight arterial-wall-specific Mechanisms. *Nat. Genet.* 49, 1113–1119. doi:10.1038/ng.3874

Inouye, M., Abraham, G., Nelson, C. P., Wood, A. M., Sweeting, M. J., Dudbridge, F., et al. (2018). Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults. *J. Am. Coll. Cardiol.* 72, 1883–1893. doi:10.1016/j.jacc.2018.07.079

Keramati, A. R., Fathzadeh, M., Go, G.-W., Singh, R., Choi, M., Faramarzi, S., et al. (2014). A Form of the Metabolic Syndrome Associated with Mutations inDYRK1B. *N. Engl. J. Med.* 370, 1909–1919. doi:10.1056/nejmoa1301824

Khan, M. A., Hashim, M. J., Mustafa, H., Baniyas, M. Y., Al Suwaidi, S. K. B. M., AlKatheeri, R., et al. (2020). Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study. *Cureus* 12, e9349. doi:10.7759/cureus.9349

Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P., Bick, A. G., Cook, N. R., et al. (2016). Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *N. Engl. J. Med.* 375, 2349–2358. doi:10.1056/nejmoa1605086

Kojima, Y., Downing, K., Kundu, R., Miller, C., Dewey, F., Lancero, H., et al. (2014). Cyclin-dependent Kinase Inhibitor 2B Regulates Efferocytosis and Atherosclerosis. *J. Clin. Invest.* 124, 1083–1097. doi:10.1172/jci70391

Koyama, S., Ito, K., Terao, C., Akiyama, M., Horikoshi, M., Momozawa, Y., et al. (2020). Population-specific and Trans-ancestry Genome-wide Analyses Identify Distinct and Shared Genetic Risk Loci for Coronary Artery Disease. *Nat. Genet.* 52, 1169–1177. doi:10.1038/s41588-020-0705-3

Kullo, I. J., Jouni, H., Austin, E. E., Brown, S.-A., Kruisselbrink, T. M., Isseh, I. N., et al. (2016). Incorporating a Genetic Risk Score into Coronary Heart Disease Risk Estimates. *Circulation* 133, 1181–1188. doi:10.1161/circulationaha.115.020109

Lo Sardo, V., Chubukov, P., Ferguson, W., Kumar, A., Teng, E. L., Duran, M., et al. (2018). Unveiling the Role of the Most Impactful Cardiovascular Risk Locus through Haplotype Editing. *Cell* 175, 1796–1810. doi:10.1016/j.cell.2018.11.014

Lu, X., Peloso, G. M., Peloso, G. M., Liu, D. J., Wu, Y., Zhang, H., et al. (2017). Exome Chip Meta-Analysis Identifies Novel Loci and East Asian-specific Coding Variants that Contribute to Lipid Levels and Coronary Artery Disease. *Nat. Genet.* 49, 1722–1730. doi:10.1038/ng.3978

Mani, A., Radhakrishnan, J., Wang, H., Mani, A., Mani, M.-A., Nelson-Williams, C., et al. (2007). LRP6 Mutation in a Family with Early Coronary Disease and Metabolic Risk Factors. *Science* 315, 1278–1282. doi:10.1126/science.1136370

Mosley, J. D., Gupta, D. K., Tan, J., Yao, J., Wells, Q. S., Shaffer, C. M., et al. (2020). Predictive Accuracy of a Polygenic Risk Score Compared with a Clinical Risk Score for Incident Coronary Heart Disease. *JAMA* 323, 627–635. doi:10.1001/jama.2019.21782

Musunuru, K., Pirruccello, J. P., Do, R., Peloso, G. M., Guiducci, C., Sougnez, C., et al. (2010). Exome Sequencing,ANGPTL3Mutations, and Familial Combined Hypolipidemia. *N. Engl. J. Med.* 363, 2220–2227. doi:10.1056/nejmoa1002926

Nelson, C. P., Goel, A., Goel, A., Butterworth, A. S., Kanoni, S., Webb, T. R., et al. (2017). Association Analyses Based on False Discovery Rate Implicate New Loci for Coronary Artery Disease. *Nat. Genet.* 49, 1385–1391. doi:10.1038/ng.3913

Nikpay, M., Goel, A., Won, H. H., Hall, L. M., Willenborg, C., Kanoni, S., et al. (2015). A Comprehensive 1,000 Genomes-Based Genome-wide Association Meta-Analysis of Coronary Artery Disease. *Nat. Genet.* 47, 1121–1130. doi:10.1038/ng.3396

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., et al. (2014). Genetics of Rheumatoid Arthritis Contributes to Biology and Drug Discovery. *Nature* 506, 376–381. doi:10.1038/nature12873

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., et al. (2002). Functional SNPs in the Lymphotoxin-α Gene that Are Associated with Susceptibility to Myocardial Infarction. *Nat. Genet.* 32, 650–654. doi:10.1038/ng1047

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., et al. (2011). Large-scale Association Analysis Identifies 13 New Susceptibility Loci for Coronary Artery Disease. *Nat. Genet.* 43, 333–338. doi:10.1038/ng.784

van der Harst, P., and Verweij, N. (2018). Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* 122, 433–443. doi:10.1161/circresaha.117.312086

Vrablik, M., Tichý, L., Freiberger, T., Blaha, V., Satny, M., and Hubacek, J. A. (2020). Genetics of Familial Hypercholesterolemia: New Insights. *Front. Genet.* 11, 574474. doi:10.3389/fgene.2020.574474

Wand, H., Lambert, S. A., Tamburro, C., Iacocca, M. A., O'Sullivan, J. W., Sillari, C., et al. (2021). Improving Reporting Standards for Polygenic Scores in Risk Prediction Studies. *Nature* 591, 211–219. doi:10.1038/s41586-021-03243-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Miyazawa and Ito. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 November 2021
doi: 10.3389/fgene.2021.747576

Identification and Validation of Prognostic Factors of Lipid Metabolism in Obstructive Sleep Apnea

Lu Peng[1,2†], *Xiaodi Wang*[1†] *and Dan Bing*[1*]

[1]*Department of Otorhinolaryngology Head and Neck Surgery, Tongji Hospital Affiliated to Tongji Medical College of Huazhong University of Science and Technology, Wuhan, China,* [2]*Department of Otorhinolaryngology Head and Neck Surgery, Second Affiliated Hospital of Guangxi Medical University, Nanning, China*

Background: Obstructive sleep apnea (OSA) is considered to be an independent factor affecting lipid metabolism. This study explored the relationship between immune genes and lipid metabolism in OSA.

Methods: Immune-related Differentially Expressed Genes (DEGs) were identified by analyzing microarray data sets from the Gene Expression Omnibus (GEO) database. Subsequently, we conducted protein-protein interaction (PPI) network analysis and calculated their Gene Ontology (GO) semantic similarity. The GO, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Disease Ontology (DO), gene set enrichment analysis (GSEA), and gene set variation analysis (GSVA) were employed for functional enrichment analyses and to determine the most significant functional terms. Combined with the results of boruta and random forest, we selected predictors to build a prognostic model, along with seeking out the potential TFs and target drugs for the predictive genes.

Results: Immune-related DEGs included 64 genes upregulated and 98 genes downregulated. The enrichment analysis might closely associate with cell adhesion and T cell-mediated immunity pathways and there were many DEGs involved in lipid and atherosclerosis signaling pathways. The highest-ranking hub gene in PPI network have been reported lowly expressed in OSA. In line with the enrichment analysis, DO analysis reveal that respiratory diseases may be associated with OSA besides immune system disorders. Consistent with the result of the KEGG pathway, the analysis of GSVA revealed that the pro-inflammation pathways are associated with OSA. Monocytes and CD8 T cells were the predominant immune cells in adipose tissue. We built a prognostic model with the top six genes, and the prognostic genes were involved in the polarization of macrophage and differentiation of T lymphocyte subsets. In vivo experimental verification revealed that EPGN, LGR5, NCK1 and VIP were significantly down-regulated while PGRMC2 was significantly up-regulated in mouse model of OSA.

Conclusions: Our study demonstrated strong associations between immune genes and the development of dyslipidemia in OSA. This work promoted the molecular mechanisms and potential targets for the regulation of lipid metabolism in OSA.

Keywords: lipid metabolism, obstructive sleep apnea, immunologic factors, macrophage activation, microarray analysis

OPEN ACCESS

Edited by:
Jordi Pérez-Tur, Instituto de Biomedicina de Valencia (CSIC), Spain

Reviewed by:
Sompop Bencharit, Virginia Commonwealth University, United States
Premkumar Kumpati, Bharathidasan University, India

****Correspondence:***
Dan Bing
didibing1981@aliyun.com

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: *26 July 2021*
Accepted: *27 October 2021*
Published: *22 November 2021*

Citation:
Peng L, Wang X and Bing D (2021) Identification and Validation of Prognostic Factors of Lipid Metabolism in Obstructive Sleep Apnea. Front. Genet. 12:747576. doi: 10.3389/fgene.2021.747576

INTRODUCTION

Obstructive Sleep Apnea (OSA) is a common disease characterized by repeated episodes of upper airway closure during sleep. The symptoms of OSA, such as snoring, nocturnal awakening, nocturia, and daytime sleepiness has low specificity to distinguish (Patel, 2019). The apnea-hypopnea index (AHI) and hypopnea per hour of sleep is the key metric to measure OSA (Jordan et al., 2014). Overnight polysomnography is the best test of OSA, as the procedure is dedicated and expensive. Atypical symptoms and inconvenient detection methods lead to a low diagnostic rate of OSA. OSA affects 9–38% of the adult population, from 6 to 19% in women and 13–33% in men (Senaratna et al., 2017).

OSA is associated with increased risk for hypertension, coronary artery disease, heart failure, stroke, type 2 diabetes, and fatty liver diseases (Tan et al., 2018; Chung et al., 2021; Yeghiazarians et al., 2021). Nowadays, more and more evidence shows that in the treatment of those diseases, OSA should be paid more attention to (Tan et al., 2018; Yeghiazarians et al., 2021). There has been a great interest in the interaction between OSA and metabolic dysfunction. Patients with OSA usually have abnormal metabolism of glucose and lipids. Though obesity is one of the main risk factors of OSA, many investigations have shown that OSA can have an independent effect on dyslipidemia as well as obesity (Karkinski et al., 2017; Silva et al., 2018; Alterki et al., 2020). In non-obese patients, OSA could aggravate abnormal lipid metabolism (Karkinski et al., 2017). Dysregulation lipid profiles are related to sleep hypoxemia even in mild OSA (Silva et al., 2018). But in obese patients, the role of OSA in the changes of dyslipidemia is not as important as in non-obese patients (Karkinski et al., 2017). However, it has also been reported that after eliminating interference factors, only severe OSA had an independent association with dyslipidemia (Martínez-Cerón et al., 2021). Treatment with OSA, either multilevel sleep surgery or continuous positive airway pressure (CPAP) therapy, has a positive impact on the metabolic status (Alterki et al., 2020; Simon et al., 2020). So, the question is, which key factors through what signaling pathways contribute to abnormal lipid metabolism in patients with OSA.

Lots of studies using high throughput microarray to analyze the differential expression genes and functional pathways related to the mechanisms and consequences of OSA. A previous study took two systems biology approaches to detect hub proteins associated with OSA in subcutaneous and visceral fat tissues. The hub genes were different between using biased methods and unbiased methods, because of the nature of the two approaches (Liu et al., 2011). Two studies were obtained the same result that the olfactory transduction pathway plays an important role in OSA using visceral adipose tissues (Gu et al., 2019) and subcutaneous adipose tissues (Cao et al., 2021) respectively. Yet, the different biomarkers and enriched pathways were using the same microarray data from visceral adipose tissues (GSE38792) (Chen et al., 2015; Gu et al., 2019), for the small sample size increased the false-positive of the results. Moreover, although a series of bioinformatics analyses has thoroughly investigated the potential biomarkers and functional pathways of OSA in adipose tissues, they remain to use cross analysis in various datasets to explore the possible mechanisms. There is a high false-positive rate using a single dataset or single method that may contribute to discordant results across these studies. Previous studies have shown that inflammatory is involved in the development of OSA (Gharib et al., 2020; Liu et al., 2020), the landscape of immune infiltration in OSA has not been entirely revealed. Accordingly, we conducted cross-analysis in immune-related biomarkers and the predicted target drugs of treatment in OSA dyslipidemia.

In the present study, we discover the key immune molecules and signaling pathways involved in lipid metabolism and identify immune molecule-related transcription factors (TFs) and drug targets. To obtain more accurate results, we downloaded two microarray datasets from the Gene Expression Omnibus (GEO) database, and then analyzed and verified them. We obtained immune-related differentially expressed genes (DEGs) between normal and OSA groups. With these DEGs, we conducted protein-protein interaction (PPI) network analysis and calculated their Gene Ontology (GO) semantic similarity. Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Disease Ontology (DO), gene set enrichment analysis (GSEA), and gene set variation analysis (GSVA) were employed for functional enrichment analyses and to determine the most significant functional terms. Combined with the results of boruta and random forest, we selected predictors to build a prognostic model, along with seeking out the potential TFs and target drugs for the predictive genes. The aim of this study was to provide a theoretical basis for immune genes that affected lipid metabolism in OSA.

MATERIAL AND METHODS

Data Collection and Processing

The datasets were obtained from the Gene Expression Omnibus database (GEO) (http://www.ncbi.nlm.nih.gov/geo/). GSE135917 (Gharib et al., 2020) and GSE38792 (Gharib et al., 2013) microarray datasets were performed on the same platform GPL6244 (HuGene-1_0-st; Affymetrix Human Gene 1.0 ST). GSE135917 contained fifty subcutaneous adipose tissue samples including normal controls (n = 8) and OSA patients without treatment (n = 34). GSE38792 contained eighteen visceral adipose tissue samples including normal controls (n = 8) and OSA patients without treatment (n = 10). We used GSE135917 as the training set and GSE38792 as the testing set. The clinical and demographic characteristics of the study patients in GSE135917 was shown in **Table 1**.

Raw data were downloaded using the GEOquery package (Davis and Meltzer, 2007) and analyzed using the oligo package (Carvalho and Irizarry, 2010) of Bioconductor in R version 4.1.0. The data were normalized with the RMA

TABLE 1 | Clinical and demographic characteristics of patients in two datasets.

	GSE135917			GSE38792	
	Group1 (individuals undergoing ventral hernia repair surgery)		Group2 (individuals diagnosed with OSA getting initiated on CPAP therapy)	Control (n = 8)	OSA (n = 10)
	Control (n = 8)	OSA (n = 10)	OSA (n = 24)		
Age (years)	54.5 ± 11.6	56.1 ± 10.8	49.4 ± 10.9	54.5 ± 11.6	56.1 ± 10.8
Gender					
Male	1	3	13	1	3
Female	7	7	11	7	7
Body mass index (kg/m^2)	35.2 ± 5.8	36.1 ± 9.3	42.6 ± 9.4	35.2 ± 5.8	36.1 ± 9.3
RDI or AHI (events/hour)	0.6 ± 0.5	19.2 ± 25.9	41.5 ± 23.8	0.6 ± 0.5	19.2 ± 25.9
Diabetes	0	3	5	0	3
Hypertension	3	3	9	3	3
Heart disease	0	0	2	0	0

method and probe IDs were converted into gene names according to the platform annotation information.

Differentially Expressed Genes Analysis

Limma package (Ritchie et al., 2015) in R was used to identify DEGs between OSA and normal adipose tissues. Genes with adjusted p-value < 0.01 were considered to be statistically significant.

The immune-related gene list was obtained from the ImmPort database (http://www.immport.org) (Bhattacharya et al., 2014). Then the gene set with immune-related genes was identified in the GSE135917 and the immune-related genes with adjusted p-value < 0.01 were considered to be statistically significant.

Protein-Protein Interaction Network Construction With Immune-Related Differentially Expressed Genes

The STRING database (https://string-db.org/) (von Mering et al., 2003) is a biological database and web resource of known and predicted protein-protein interactions. We uploaded the immune-related DEGs to the STRING database. The species was set as Homo sapiens and the minimum interaction score was 0.4 to build a protein interaction network. The PPI network of the immune-related DEGs was visualized with Cytoscape 3.8.2 software (Shannon et al., 2003).

Calculation of Gene Ontology Semantic Similarity

GO terms include biological process (BP), molecular function (MF), and cellular component (CC). The GO semantic similarity score can be applied to quantify the functional similarity between genes. To assess immune-related gene functional similarity, we calculated semantic similarity scores of GO terms using the R GOSemSim package (Yu et al., 2010; Yu, 2020). MgeneSim automatically removes genes without annotations and computed the semantic similarity among GO terms. The functional similarity score of the target gene is calculated as follows:

$$\mathrm{Fsim} = \sqrt[3]{simbp * simcc * simmf}$$

Gene Ontology and Pathway Analysis

To reveal the functions and pathways of immune-related DEGs, GO and KEGG pathway analyses were performed using the R clusterProfiler package (Yu et al., 2012). Significant KEGG pathways and participating genes were visualized with the R pathview package (Luo and Brouwer, 2013). In all enrichment analyses, Benjamini-Hochberg (BH) adjustment to calculate the false discovery rate (FDR) was applied. A q-value < 0.05 was set as the cutoff criterion.

DOSE (Yu et al., 2015) is an R package for disease ontology semantic and enrichment analysis. We used the DOSE R package to analyze the enrichment of immune-related DEGs with Disease Ontology (DO) terms.

Gene Set Enrichment Analysis and Gene Set Variation Analysis

Gene set enrichment analysis was performed using the R clusterProfiler package (Yu et al., 2012). p values were adjusted by the BH method. We used FDR (false discovery rate) < 0.1, and p-value < 0.01 as the threshold to determine significant enrichment of the gene sets. Then Gene Set Variation Analysis (GSVA) (Hänzelmann et al., 2013), a nonparametric unsupervised method, was used to display differential enrichment pathways between normal controls and OSA patients. A p-value < 0.01 was set as the cutoff criterion. In this study, we used the R package "GSVA" to explore KEGG pathways of immune-related genes. Gene terms were considered statistically significant, and a heatmap was generated using R.

Prognostic Model Building and Validation

Boruta feature selection ("Boruta" package in R) (Miron B. Kursa and Rudnicki, 2010) was used on the training dataset

(GSE135917) to identify the immune-related genes that contribute significantly to OSA. Random forest was implemented in R using the randomForest package (Liaw and Wiener, 2001), and important immune-related genes were selected as features that could construct a prognostic model.

Based on the results of boruta and random forest, the selected set of predictors were used to construct a prognostic model. The receiver operating characteristic (ROC) curve was used to confirm the performance of the model and the area under the curve (AUC) was estimated in the training dataset (GSE135917), and the GSE38792 dataset was applied to verify the established prognostic model. At last, we analyzed the expression of predictors between normal and OSA groups in the two datasets.

Immune Cell Infiltrate Analysis

To understand the immune microenvironment of adipose tissue, we analyzed the differential expression of different types of immune cells. The CIBERSORTx algorithm (Newman et al., 2015) was used to calculate and analyze the immune microenvironment of adipose tissue involved in OSA and normal controls.

Correlation Analysis Between Predictors and Immune Cells

To further understand the relationship between predictors and immune cells, Pearson's correlation was performed to analyze the correlation between the expression value of predictors and the different types of immune cells.

Transcription Factors and Target Drugs Analysis

To further investigate the transcription factor binding motifs of predictors, the iRegulon (Janky et al., 2014) software was used. The set of predictors was submitted to iRegulon and analyzed using the following options: minimum NEScore = 5.0. The results were visualized using Cytoscape software.

Furthermore, we identified drug-gene interactions using Drug-Gene Interaction Database (DGIdb) (Cotto et al., 2018). The list of predictors was uploaded to DGIdb and matched with drugs that could be the potential therapeutic targets of OSA.

Animal Model and Chronic Intermittent Hypoxia Protocols

C57BL/6J adult male mice (8 weeks old) were purchased from the Model Animal Research Center of Tongji Medical College of Huazhong University of Science and Technology (Wuhan, China). Animals were randomly assigned to control and OSA groups (n = 6 animals/group). The OSA group mice were exposed to 4 weeks of CIH (8 daylight hours per day, 10:00 am to 6:00 pm), whereas the control group was maintained under normal oxygenation conditions. The animal study was approved by the Institutional Ethics Committee for Animal Research of Tongji Medical College, Huazhong University of Science and Technology. All procedures conformed to the Guide for the Care and Use of Laboratory Animals.

Gas-control delivery equipment was installed to regulate nitrogen and oxygen flow into the customized chamber. The equipment was composed of sensors for O2 and gas injectors. During each 510-s cycle there included 4 stages. In stage 1, with N2 infused into the chamber, the concentration of O2 lowered from 21 to 5% in 150 s, and then maintained at 5% for 120 s in stage 2. In stage 3, the chamber was infused with O2 for 120 s to restore O2 to an ambient concentration of 21%, and it was maintained in stage 4 until the beginning of the next CIH cycle.

Reverse Transcription-Polymerase Chain Reaction

Visceral adipose tissue was collected and total mRNA and was extracted using TRIzol Reagent following the manufacturer's protocol. The extracted mRNA (1 μg) was reverse transcribed into cDNA using. Real-time PCR was performed on a LightCycler System 2.0 (Roche, Mannheim, Germany) using SYBR Premix EX Taq kit (Takara, Dalian, China). RT-PCR was performed at 95°C for 5 min, then 95°C (45 s), 56°C (30 s), and 72°C (45 s) followed by a 10 min extension at 72°C for 40 cycles. Each sample was run in triplicate and averaged. The relative gene expression was calculated by the 2-△△Ct method.

The primer sequence is as follows: EPGN forward primer: 5′-GGGGGTTCTGATAGCAGTCTG-3′, reverse primer: 5′-TCG GTGTTGTTAAATGTCCAGTT-3'. LGR5 forward primer: 5′-CCTACTCGAAGACTTACCCAGT-3′, reverse primer: 5′- GCA TTGGGGTGAATGATAGCA-3'. NCK1 forward primer: 5′-TCCTGCTGATGATAGCTTTGTTG-3′, reverse primer: 5′-ACGATCACCTTGGTCCCTTTTAT-3'. PGRMC2 forward primer: 5′- TGGGAAAGTCTTCGACGTGAC-3′, reverse primer: 5′- GTGCATCCTTATCCAGGCAGA-3'. VIP forward primer: 5′- AGTGTGCTGTTCTCTCAGTCG-3′, reverse primer: 5′- GCCATTTTCTGCTAAGGGATTCT-3'. β-actin forward primer: 5′- GCGCAAGTACTCTGTGTGGA-3′, reverse primer: 5′-GAAAGGGTGTAAAAC GCAGC-3'.

RESULTS

Data Pre-processing

Our workflows are shown in **Figure 1**. To eliminate batch expression difference, each data was normalized by the RMA method in R (**Figure 2**). The original expression values varied significantly between the samples and the mean values of gene expression for each sample were fundamentally the same after normalization.

Identification of Differentially Expressed Genes in Obstructive Sleep Apnea

The GSE135917 after normalization was utilized to obtain the differentially expressed genes between OSA and normal adipose tissues. With adj p-value < 0.01 as the screening threshold, DEGs were obtained with 249 genes upregulated and 133 genes downregulated. The volcano plot and heatmap of the DEGs are shown in **Figures 3B,C**. With adj p-value < 0.01 as the

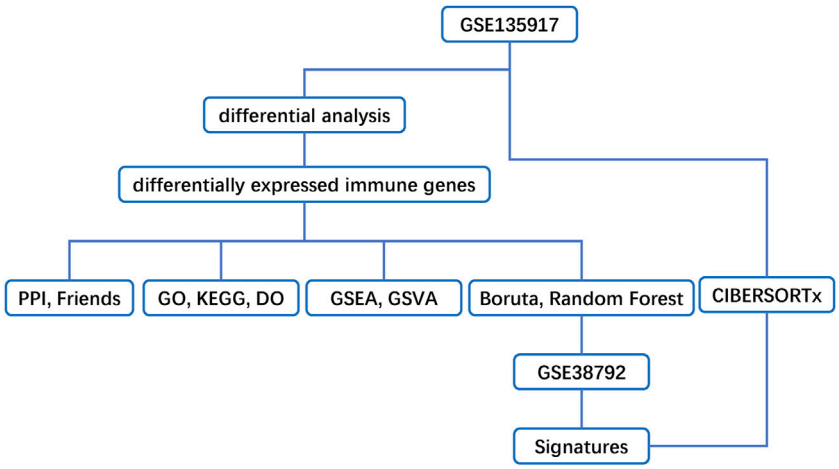


FIGURE 1 | Flow chart of methodologies applied in the current study.

screening threshold, immune-related DEGs were obtained with 64 genes upregulated and 98 genes downregulated. The volcano plot and heatmap of the DEGs are shown in **Figures 3E,F**. In vocano plot, red dots represent significant different expression genes, and green dots represent no significant different expression genes. In heatmap, each row represents one gene, and each column represents one sample. Red indicates that the expression of genes is relatively upregulated, and blue indicates that the expression of genes is relatively downregulated.

We applied principle component analysis (PCA) on the DEGs (**Figure 3A**) and immune-related DEGs (**Figure 3D**). Unsupervised clustering of the two set DEGs showed that samples from normal and OSA could be divided into two main clusters. The result revealed that the features of OSA adipose tissue could be explained only by immune-related DEGs, then we used 162 immune-related DEGs to perform subsequent analysis.

Protein-Protein Interaction Network Construction and Semantic Similarity Analysis

To analyze the interaction among 162 immune-related DEGs, the STRING database was used. A total of 122 nodes and 496 edges were obtained with a combined score >0.7, as shown in **Figure 4A**. The top3 hub genes with the highest ranking were found: interleukin 6 (*IL6*), proopiomelanocortin (*POMC*), mitogen-activated protein kinase 3 (*MAPK3*).

We also calculated the average semantic similarity for immune-related DEGs in GO terms, including biological process (BP), molecular function (MF), and cellular component (CC) categories. Based on the average functional similarity, we ranked the top 10 genes among the immune genes (**Figure 4B**). Defensin beta 129 (*DEFB129*), fibroblast growth factor 16 (*FGF16*), and proteasome 26S subunit, non-ATPase 8 (*PSMD8*) were the top three genes potentially playing key roles in OSA.

Gene Ontology, Kyoto Encyclopedia of Genes and Genomes Pathway, and Disease Ontology Terms Enrichment Analysis

The results of GO functional, KEGG pathway and Do terms enrichment analysis are shown in **Figure 5**. In the BP category of the GO enrichment analysis, immune-related DEGs were mainly enriched in items such as "positive regulation of cell-cell adhesion", "positive regulation of cell adhesion", and "regulation of cell-cell adhesion" (**Figure 5A**). In the CC category of the GO enrichment analysis, these genes were mainly enriched in items such as "proteasome accessory complex", "proteasome regulatory particle", and "secretory granule lumen" (**Figure 5B**). In the MF category of the GO enrichment analysis, these genes were mainly enriched in items such as "receptor ligand activity", "signaling receptor activator activity", and "hormone activity" (**Figure 5C**). Meanwhile, based on the results of the KEGG pathway enrichment analysis (**Figure 5D**), most of the immune-related DEGs were significantly enriched for the terms: "cytokine-cytokine receptor interaction" (**Figure 5H**), "T cell receptor signaling pathway" (**Figure 5F**), and "lipid and atherosclerosis" (**Figure 5G**). Therefore, the GO terms from BP and KEGG signaling pathways might closely associate with cell adhesion and T cell-mediated immunity pathways. Interestingly, among the differentially expressed genes, there were many involved in lipid and atherosclerosis signaling pathways.

We found DO terms mainly enriched in "bacterial infectious disease", "osteoarthritis", and "primary bacterial infectious disease" (**Figure 5E**). In line with the above GO-term and KEGG pathway analysis, disease ontology (DO) revealed that respiratory diseases may be associated with OSA besides immune system disorders.

Pathway Enrichment Analysis of the Immune-Related Differentially Expressed Genes

To further explore the signaling pathway associated with immune genes involved in OSA, we identified the pathways significantly

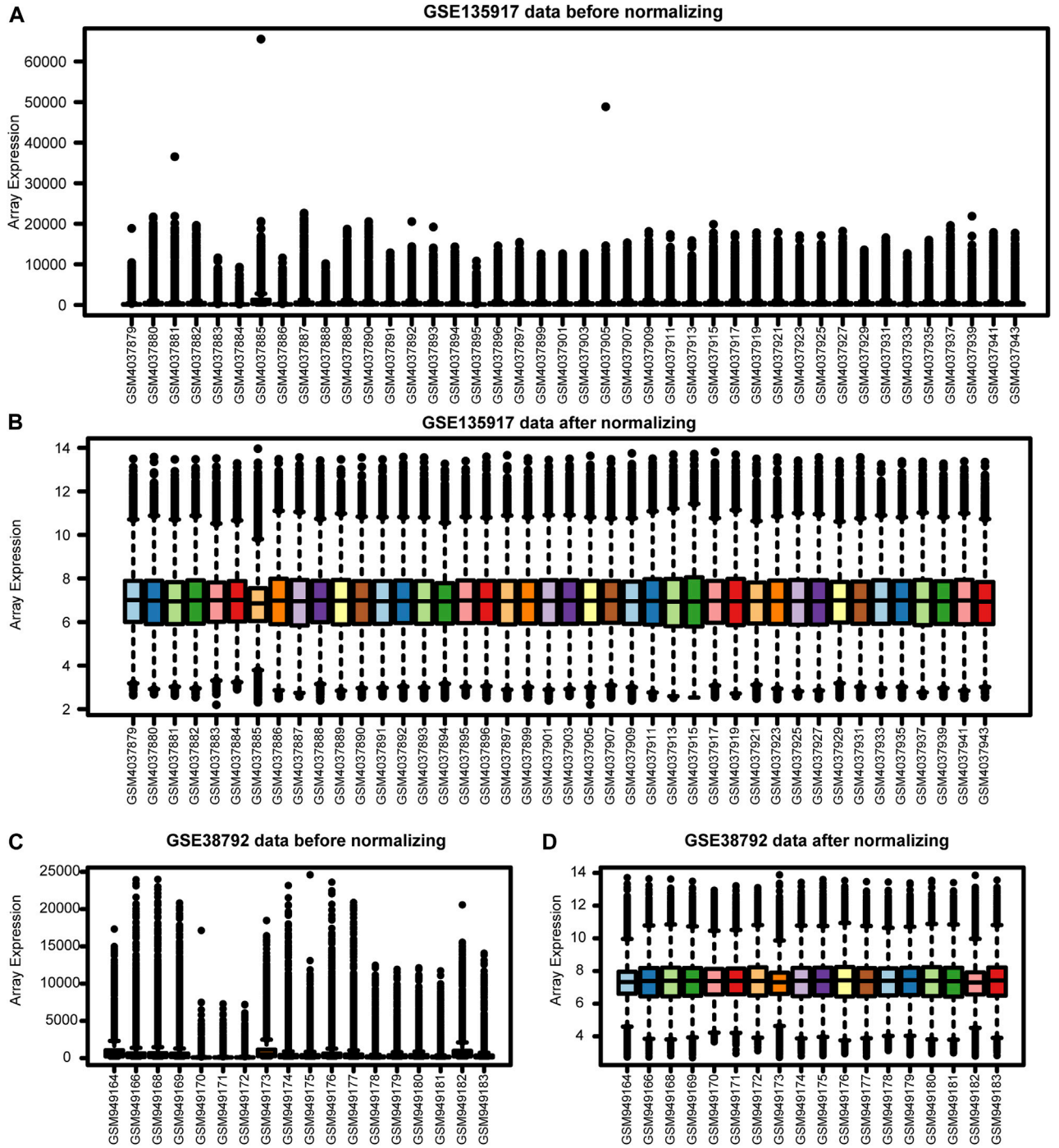


FIGURE 2 | The Expression profiles before and after normalization. **(A)** GSE135917 data before normalization. **(B)** GSE135917 data after normalization. **(C)** GSE38792 data before normalization. **(D)** GSE38792 data after normalization.

enriched through GSEA and GSVA analysis. GSEA analysis indicated that OSA is predominantly associated with an IL-17 signaling pathway, pertussis, rheumatoid arthritis, parathyroid hormone synthesis, secretion, and action, TNF signaling pathway, amyotrophic lateral sclerosis, and non-alcoholic fatty liver disease (**Figures 6A,B**). This result was consistent with the analysis of KEGG pathways that the IL-17 signaling pathway may play an important role in OSA. It is noteworthy that the non-alcoholic fatty liver disease signaling pathway is involved in the genesis of OSA disease.

The GSVA result is presented in the heat map (**Figure 6C**), and further uncovered differences of normal and OSA samples.

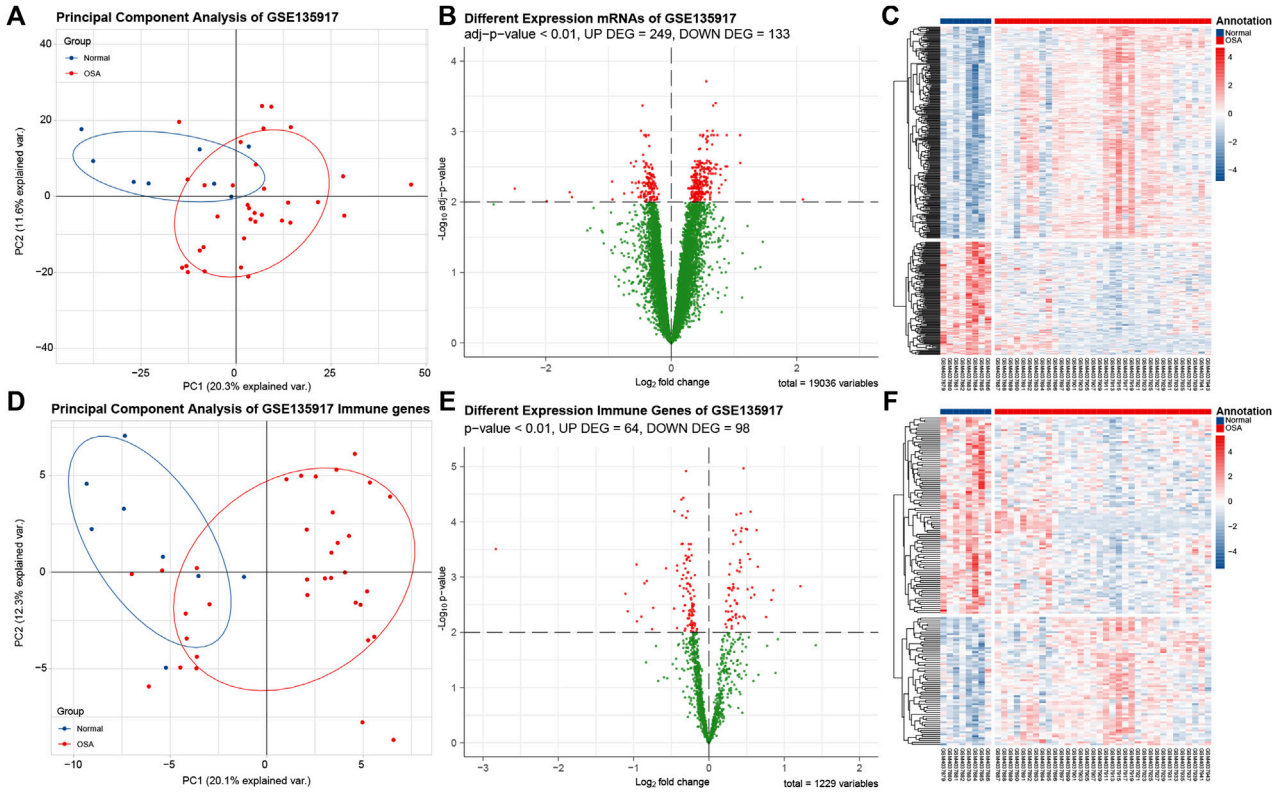


FIGURE 3 | Analysis of DEGs and immune-related DEGs in data set GSE135917. **(A)** Principle component analysis (PCA) plot of 382 DEGs shows that samples be divided into two clusters. Blue dots indicate normal samples, red dots indicates obstructive sleep apnea (OSA) samples. **(B)** The volcano plot of differentially expressed genes (DEGs). Red dots represent significant different expression genes, and green dots represent no significant different expression genes. **(C)** The heatmap of DEGs. Each row represents one gene, and each column represents one sample. Red indicates that the expression of genes is relatively upregulated, and blue indicates that the expression of genes is relatively downregulated. **(D)**. PCA plot of 162 immune-related DEGs. **(E)** The volcano plot of immune-related DEGs. **(F)** The heatmap of immune-related DEGs.

Consistent with the result of the KEGG pathway, toll-like receptor signaling pathway, MAPK signaling pathway, and T cell receptor signaling pathway are associated with OSA.

Prognostic Model Building and Validation

Boruta feature selection method was used to select significant predictors to improve prediction in immune-related DEGs. Along with Boruta running, the z score evolution is shown in **Figure 7A**. A total of 27 genes were selected by the Boruta algorithm (**Figure 7B**). Random forest analysis provided further support for the predictors' selection (**Figures 7C,D**). The average error rate was minimum with five sample trees from **Figure 7C**. Then we analyzed the variable importance of random forest by using accuracy and the Gini index of a mean decrease (**Figure 7D**). Therefore, we built the final prognostic predictors with the top six genes, that were vasoactive intestinal peptide (*VIP*), progesterone receptor membrane component 2 (*PGRMC2*), NCK adaptor protein 1 (*NCK1*), leucine rich repeat containing G protein-coupled receptor 5 (*LGR5*), epithelial mitogen (*EPGN*), and defensin beta 135 (*DEFB135*). ROC curves were also applied to compare the efficiency of the predictive model and those genes.

Training and testing sets were used for each evaluation to confirm the performance and reliability of the prognostic model (**Figures 6E–H**). The expression trend of the predictive genes in the training set (GSE135917) (**Figure 7F**) was consistent with the testing set (GSE38792) (**Figure 7G**). The AUC of the ROC for this prognostic model was 1 in both of the two sets (**Figures 7E,H**), indicating that these predictors showed good performance in distinguishing persons who will easily lead to OSA.

Immune Cell Infiltrate Analysis

We investigated whether distinct patterns of immune infiltration could be discerned based on the 10 kinds and 22 types of the immune cell by the CIBERSORTx method. First, we evaluated the composition of the immune cell infiltrate in OSA (**Figure 8A**). In adipose tissue, the predominant immune cell type was monocytes, followed by CD8 T cells. The main cell types of monocytes kind were monocytes, anti-inflammatory macrophages (M2), and inactive macrophages (M0), with almost no inflammatory macrophages (M1).

The differential proportions of immune infiltration cells in normal and OSA groups are shown in **Figure 8B**. Within the 10 kinds of immune cells, there were no significant differences in

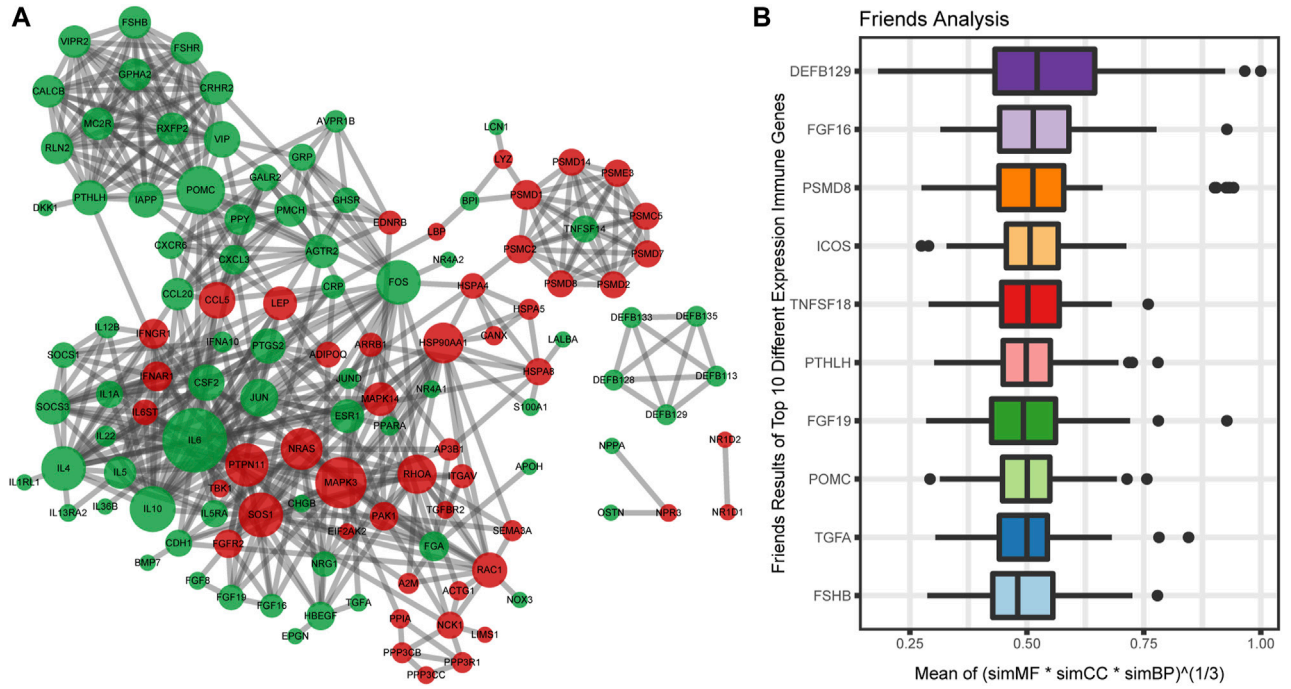


FIGURE 4 | Protein-protein interactions (PPI) network construction and semantic similarity analysis of immune-related DEGs. **(A)** The PPI network of immune-related DEGs. Each circle represents a gene. The upregulated genes (red) and downregulated genes (green) are represented by circles. Different sizes indicate the core degree of genes in the PPI network, whereas bigger size indicates more important in the network. **(B)** Summary of functional similarities of the top 10 immune-related DEGs. The aggregate score is between 0 and 1. The higher the score is, the more similarity genes are.

immune cell composition between OSA patients and normal, whereas when divided into the 22 types, macrophages M0 was statistically significantly different between OSA patients and normal. Macrophages M0 was significantly higher in the adipose tissue of OSA patients compared with normal.

Furthermore, Pearson correlation analysis was used to investigate the correlations of immune cells in the training set (**Figure 8C**). We observed that monocytes had a significant positive correlation with CD8 T cells, and a significantly negative with macrophages M0, macrophages M2. Meanwhile, T cells CD8 was negatively correlated with macrophages M0 and macrophages m2. The relationship between CD8 T cells and monocytes, as well as the polarization of monocytes, requires further study. In addition, there was a positive correlation between NK cell activated and mast cell resting.

Correlation Analysis Between Predictors and Immune Cells

Significant correlations between six predictive genes and immune cells are shown in **Figures 9A–G**. The abundance of macrophages M0 in normal and OSA groups is illustrated using violin plots (**Figure 9H**). *NCK1* had a significant correlation with more immune cells than the other five genes. There was a significant negative correlation between *NCK1* and monocytes, which was following the results in front. That indicated that *NCK1* may be associated with monocytes and subsequent polarization in the adipose tissue of OSA patients.

Transcription Factors and Target Drugs Analysis

Candidate transcription factors, being hypothetically able to control the expression of the six predictors, were predicted (**Figure 10A**). We also used DGIdb to identify essential genes that are potentially "druggable". **Figure 10B** shows a drug-gene network visualization using gene-centric fashions.

Analysis of the Expression Level of the Predictive Genes *In Vivo*.

Mice exposed to chronic intermittent hypoxia (CIH), which mimic hypoxia condition during OSA, are most frequently used as an animal model for OSA. The flow chart of CIH exposure procedure in our study was shown in **Figure 11A**. As *DEFB135* was not expressed in mice, the expression of the other five predictive genes after chronic intermittent hypoxia for 4 weeks is shown (**Figures 11B–F**). The levels of *EPGN*, *LGR5*, *NCK1*, and *VIP* in visceral adipose tissue of CIH group mice were significantly down-regulated compared to the control group while the level of *PGRMC2* was significantly up-regulated.

DISCUSSION

Obesity is a clear risk factor in the development of OSA, the incidence rate of OSA is increasing year by year with the prevalence of obesity. The synergistic effect of obesity and OSA

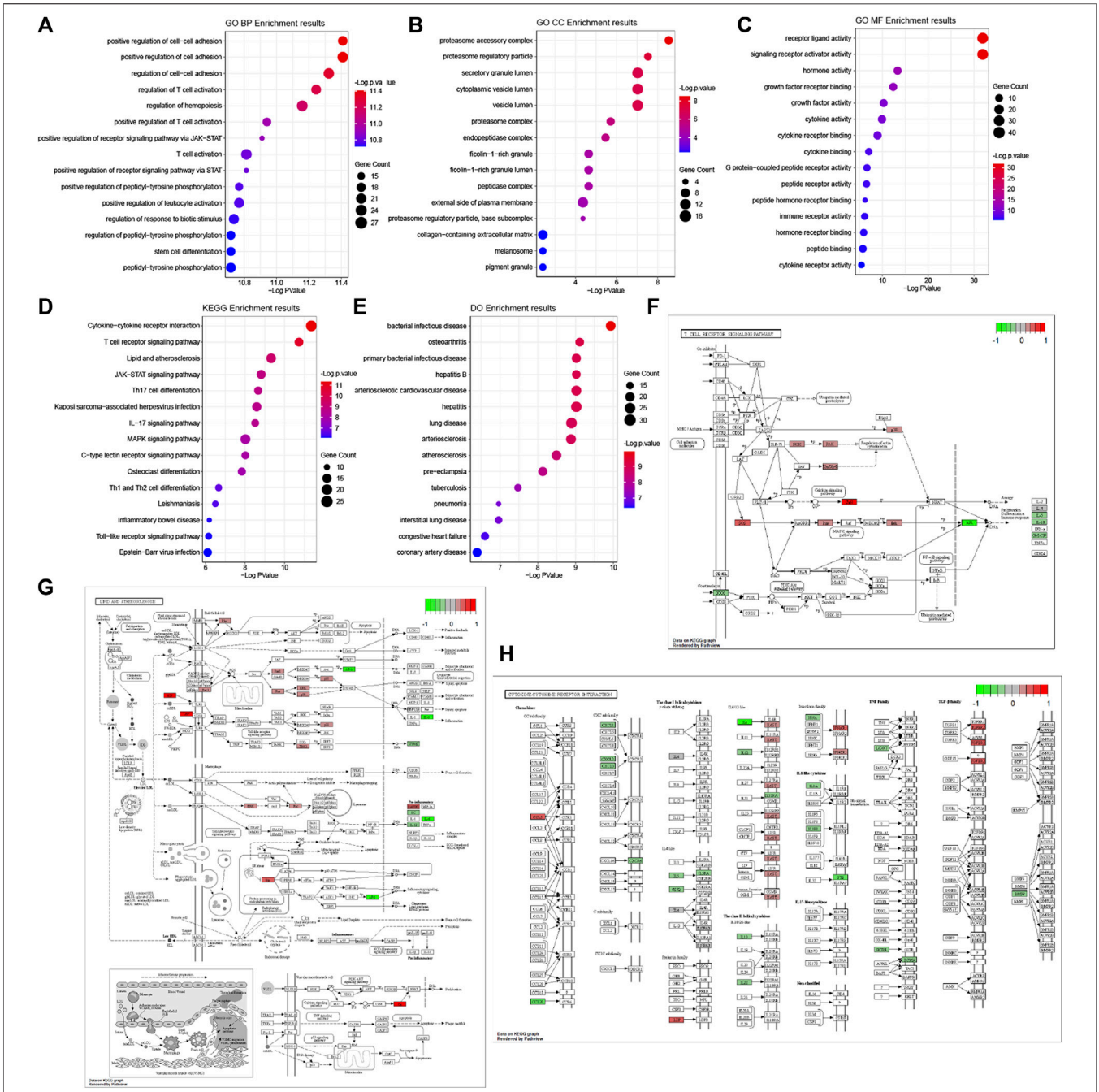


FIGURE 5 | GO, KEGG pathway and Do terms enrichment analysis of immune-related DEGs. GO terms enrichment analysis of the DEGs, including BP **(A)**, CC **(B)**, and MF **(C)** categories. **(D)** KEGG pathway enrichment analysis of the DEGs. **(E)** Do terms enrichment analysis of DEGs. The size of the symbol represents the gene counts enriched in the signaling pathway. The color indicates the degree of significance. Signaling pathways of the T cell receptor signaling pathway **(F)**, lipid and atherosclerosis **(G)**, and cytokine-cytokine receptor interaction **(H)**. The genes significantly up-regulated filled in red color and down-regulated filled in green color.

increased the incidence of metabolic diseases, such as dyslipidemia, hypertension, insulin resistance, cardiovascular diseases, and non-alcoholic fatty liver disease. (Jordan et al., 2014). Recently, OSA is considered to be an independent factor affecting lipid metabolism and there has been a great interest in the interaction between OSA and lipid metabolic dysfunction. Rodents are used to study the occurrence and development mechanism of dyslipidemia in OSA but they do not naturally exhibit OSA. Animal models to research the physiological mechanisms underlying OSA always achieve its hallmarks outcomes as intrathoracic pressure swings, sleep fragmentation, hypercapnia, and

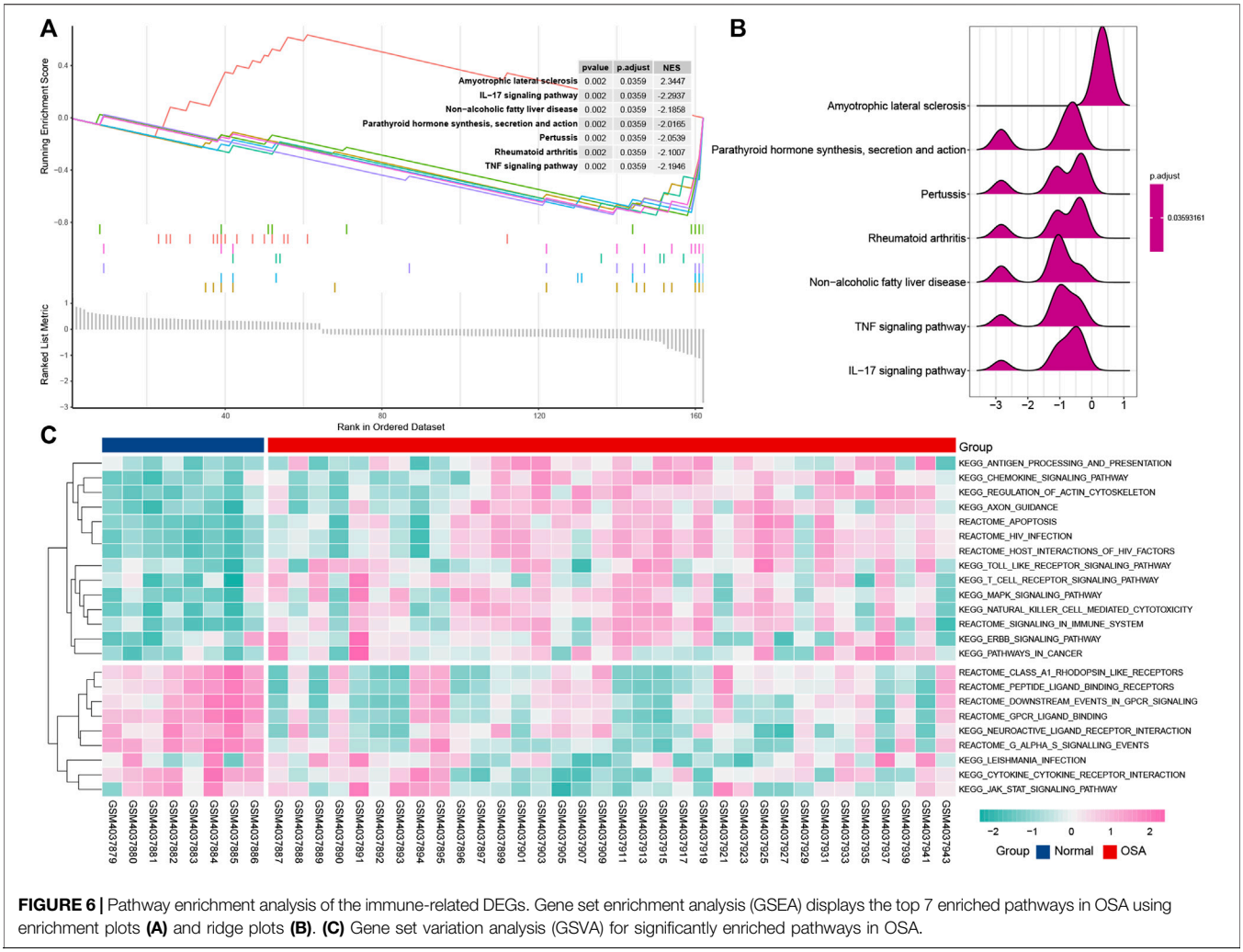


FIGURE 6 | Pathway enrichment analysis of the immune-related DEGs. Gene set enrichment analysis (GSEA) displays the top 7 enriched pathways in OSA using enrichment plots **(A)** and ridge plots **(B)**. **(C)** Gene set variation analysis (GSVA) for significantly enriched pathways in OSA.

intermittent hypoxia (Barros and García-Río, 2019; Mesarwi et al., 2019).

Inflammation and lipid signaling are synergism to maintain the stability of the internal environment and immunity (Shimobayashi et al., 2018), so the dyslipidemia of OSA may be the result of the immune response. Therefore, we performed an integrated analysis to identify the effect of immune-related genes on dyslipidemia in OSA. In the first part of our study, the results of the PCA analysis showed that normal and OSA adipose tissue samples could be clearly distinguished by the immune-related DEGs as well as DEGs. It laterally proved that immune responses contribute to the pathogenesis and progress in adipose tissue of OSA.

Next, we used the immune-related DEGs to build a PPI network and find the functional similarity between them. GO, KEGG, DO, GSEA, and GSVA analyses were performed to explore the biological functions, enriched signaling pathways, and related diseases. IL-6, the highest ranked gene in the PPI network, was significantly downregulated in cytokine-cytokine receptor interaction and lipid and atherosclerosis signaling pathways. MAPK3, one of top3 hub genes in PPI network, is a member of the MAP kinase family. MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act in a signaling cascade that regulates various cellular processes such as proliferation, differentiation, and cell cycle progression in response to a variety of extracellular signals. Protein kinase C and ERK activation are required for TFF-peptide-stimulated bronchial epithelial cell migration and tumor necrosis factor-alpha-induced interleukin-6 (IL-6) and IL-8 secretion (Graness et al., 2002). The IL6-202 and IL6-205 transcripts that confer drug resistance to Vemurafenib by reactivating the MAPK pathway while IL6-201 is not responsible for the resistance in A375 melanoma cells. Neutralizing IL-6 significantly increased the sensitivity of drug-resistant cells to Vemurafenib (Zhao et al., 2020). From our results, those two pathways were the top3 pathways negatively correlated with the OSA group. It is the same as a previous meta-analysis that either children or adults with obstructive sleep apnea syndrome had higher serum/plasma IL-6 levels compared to healthy controls (Imani et al., 2020). Results of GO BP terms, KEGG pathway, and GSVA revealed that immune-related DEGs were mainly enriched in cell adhesion and T cell-mediated immunity pathways, according to well with the

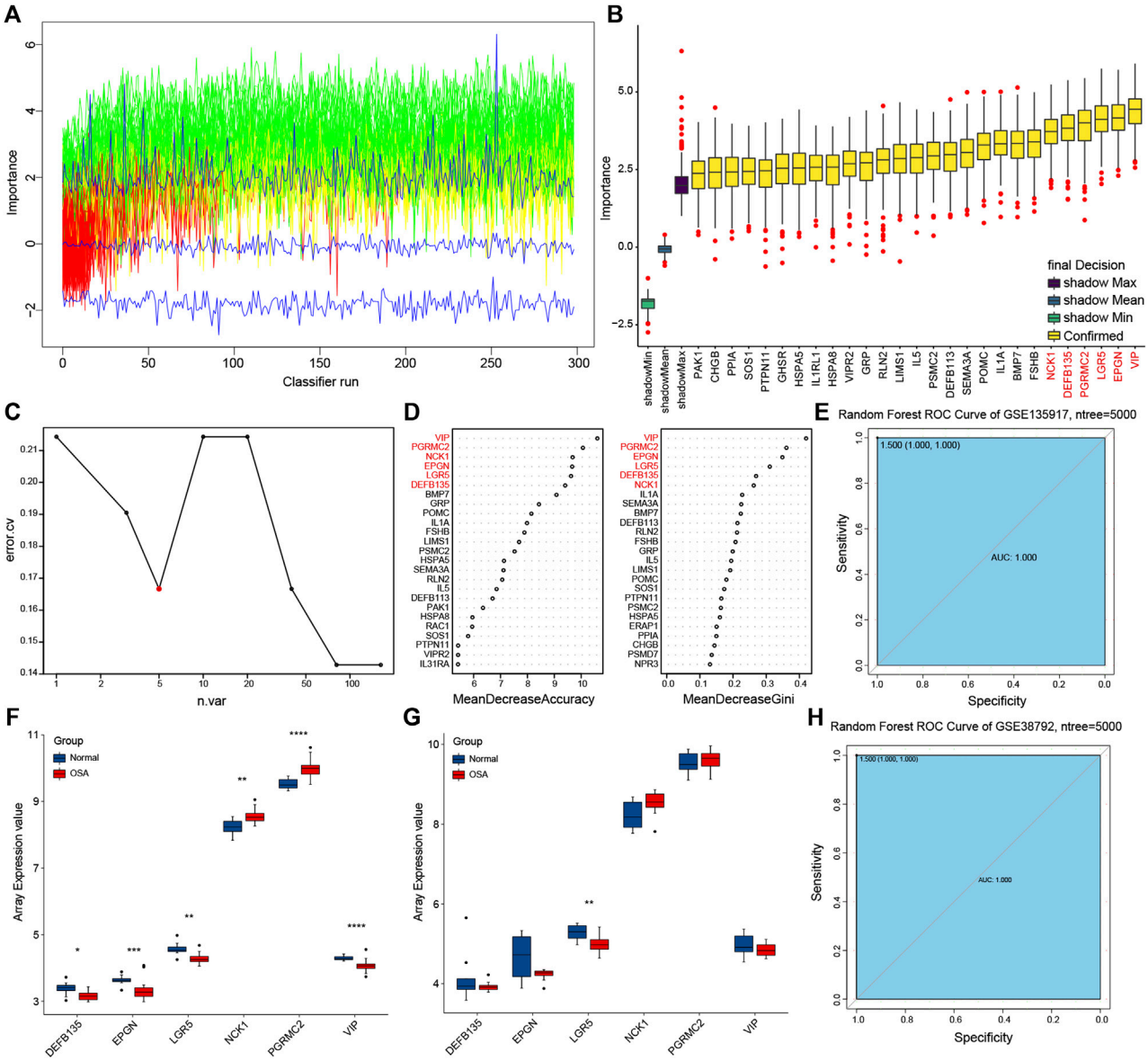


FIGURE 7 | Prognostic model building and validation. **(A)** The z score evolution with Boruta run. **(B)** Selected genes by Boruta algorithm. **(C)** The average error rate of random forest model. **(D)** Variable importance ordered by accuracy and the gini index of a mean decrease in random forest. **(E)** Receiver operating characteristic (ROC) curve with area under the curve (AUC) values for GSE135917. **(F)** The expression of predictors in GSE135917. **(G)** The expression of predictors in GSE38792. **(H)** ROC curve with AUC values for GSE38792.

pathophysiological mechanism of DO term enriched diseases. Notably, the IL-17 signaling pathway, the result enriched both in KEGG analysis and GSEA, is IL-17 family mediated immune response in both acute and chronic inflammatory responses. Clinical evidence indicated that the pathogenesis of OSAS may be associated with increased IL-17A broken the balance of peripheral Th17/Treg (Ye et al., 2012; Ying et al., 2014). Following clinical observation (Jordan et al., 2014; Chung et al., 2021; Yeghiazarians et al., 2021), disease ontology (DO) queries revealed multiple DO terms not only associated with inflammation, but also with metabolic diseases such as cardiovascular diseases and liver diseases. GSEA analysis showed the immune-related DEGs significantly enriched in pathways related to non-alcoholic fatty liver disease, maybe it was the combined action of obesity and OSA in lipid metabolism.

In the following part of our study, we selected significant diagnostic genes (*VIP, PGRMC2, NCK1, LGR5, EPGN, and DEFB135*) and constructed a diagnostic model using the genes. Concerning diagnostic value, the AUC of the diagnostic model and the expression of the six diagnostic genes were analyzed using cross-validation. The results showed that those six genes may be promising targets for the diagnosis of dyslipidemia of OSA. Although none of the genes has been reported to be associated with OSA, it is also reasonable for them to influence the

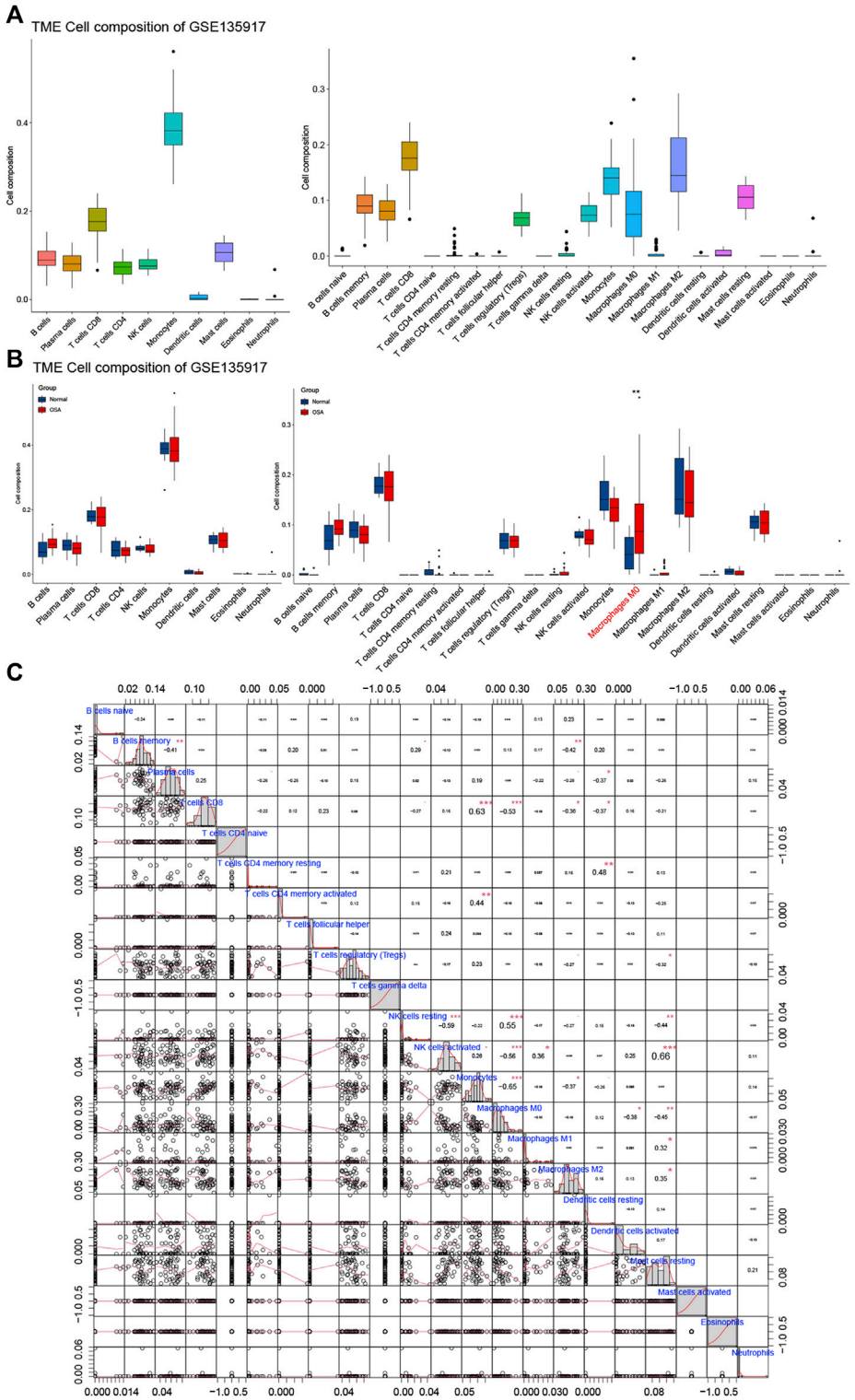


FIGURE 8 | Immune cell infiltrate analysis. **(A)** The composition of the immune cell infiltrate in OSA. **(B)** The differential expression of different types of immune cells between normal and OSA tissues. **(C)** Correlation matrix of 22 types of immune cell proportions.

development of dyslipidemia in OSA. *PGRMC2*, ubiquitous expression in fat, is involved in adipose tissue development and steroid hormone mediated signaling pathway (Galmozzi et al., 2019). As metabolic molecules, it is not surprising that they could be involved in the lipid metabolism of OSA. Vasoactive intestinal peptide (*VIP*) has immune regulatory

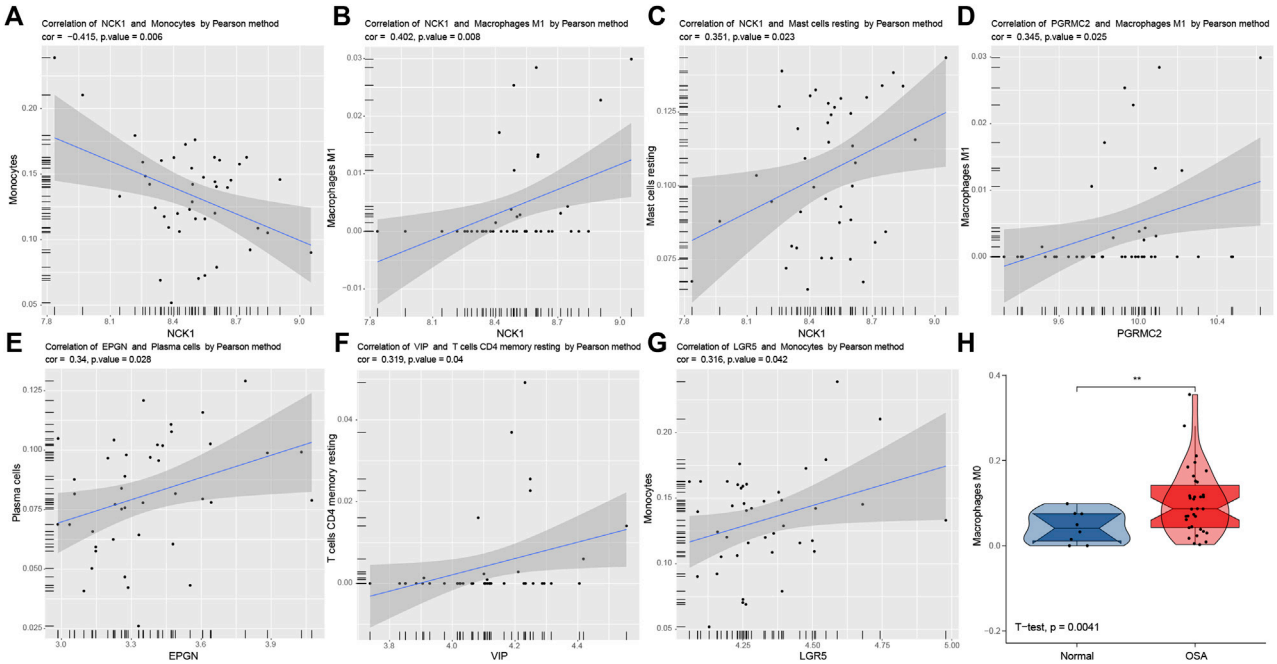


FIGURE 9 | Correlation analysis between predictors and immune cells. Significantly correlations between predictors and immune cells: NCK1 and monocytes **(A)**, NCK1 and macrophages M1 **(B)**, NCK1 and mast cells resting **(C)**, PGRMC2 and macrophages M1 **(D)**, EPGN and plasma cells **(E)**, VIP and T cells CD4 memory resting **(F)**, and LGR5 and monocytes **(G,H)**. Violin plots of the abundance of macrophages M0. The box plots in the violin indicate the median and interquartile range of the data distribution.

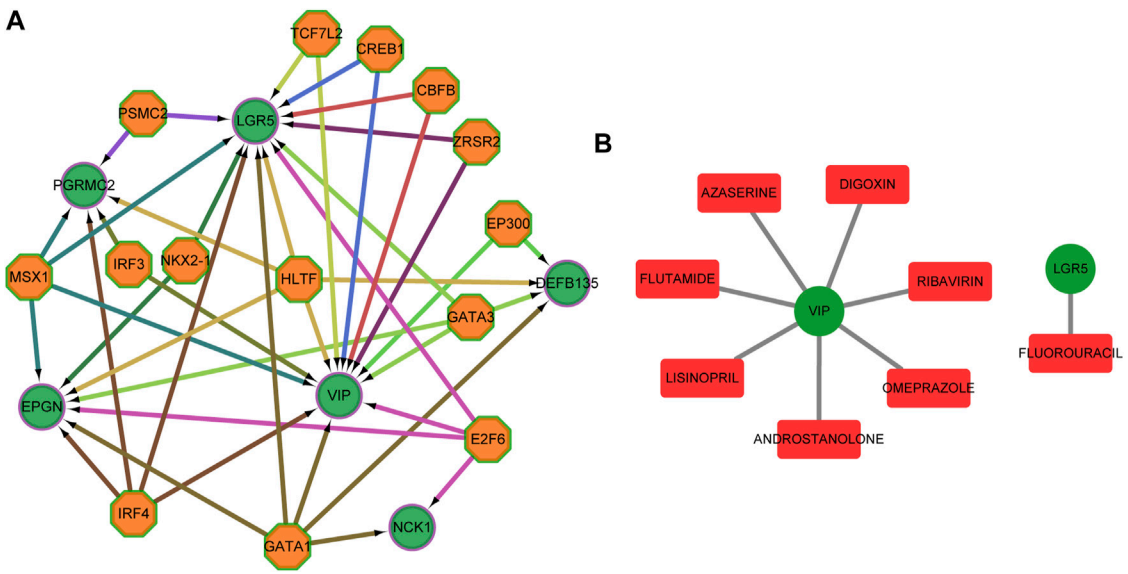


FIGURE 10 | Transcription factors and target drugs analysis. **(A)** Regulatory network of the predicted transcription factors and the target genes. **(B)** Drug-gene network using gene-centric fashions. Green circles indicate target genes, orange octagons indicate predictive transcription factors, and red quadrilateral indicate predictive drug.

functions, and administration of *VIP* can inhibit experimental colitis (Sun et al., 2019). *EPGN* is a member of the epidermal growth factor family, and EPGN SNP (single nucleotide polymorphism) is significantly associated with variations in cytokine secretion to vaccinia virus stimulation in smallpox vaccine recipients (Kennedy et al., 2012). *LGR5*, a receptor for

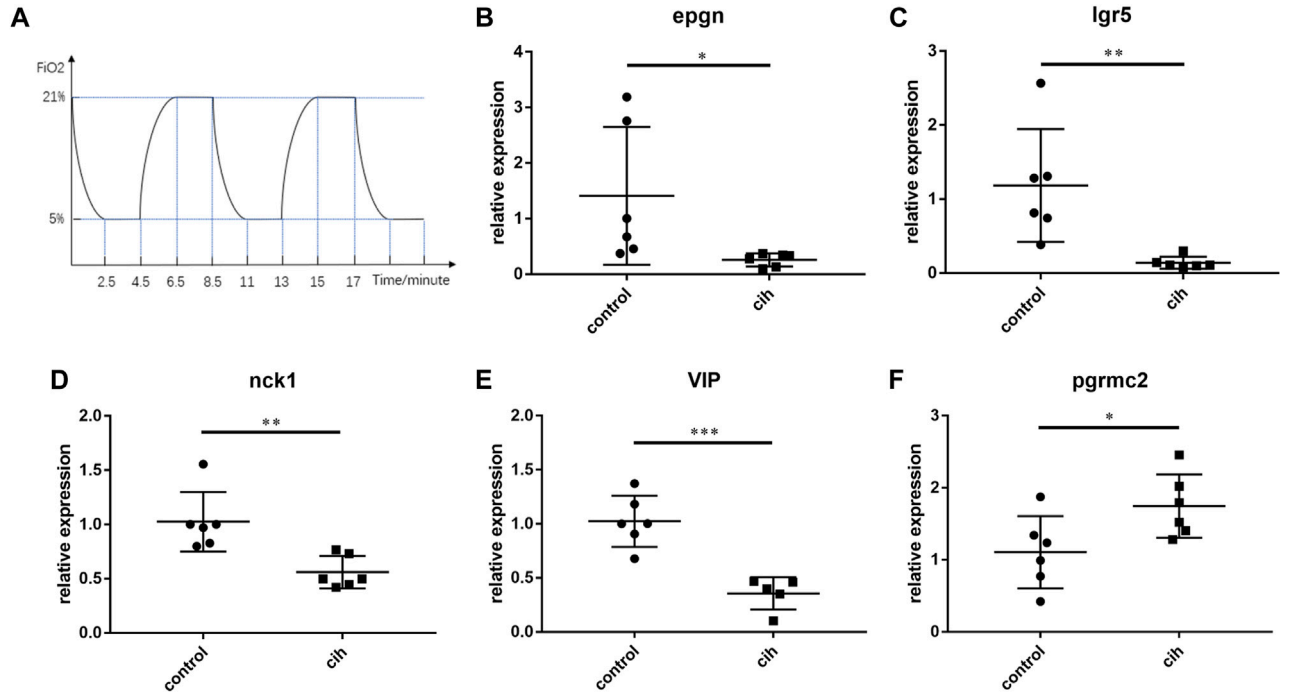


FIGURE 11 | The expression trend of prognostic factors after 4 weeks of chronic intermittent hypoxia (CIH). **(A)** The flow chart of CIH exposure procedure. The expression of EPGN **(B)**, IGR5 **(C)**, NCK1 **(D)**, VIP **(E)**, and PGRMC2 **(F)**. Data are presented as means ± SEM, n = 6 for each group. $^{*}p < 0.05$, $^{**}p < 0.01$ and $^{***}p < 0.001$ compared with control animals.

R-spondins, promotes epithelial-mesenchymal transition by activating the Wnt/β-catenin pathway in glioma (Zhang et al., 2018). *NCK1* is involved in enhancing downstream T cell activation signaling (Wipa et al., 2020). *DEFB135* also was one of the diagnostic genes, while *DEFB129* got the highest scores in functional similarity analysis. Both of those two genes are a member of the beta defensin protein family, and defensins are the only group of antimicrobial peptides found in animals, involved in the first line of defense in their innate immune response against pathogens (Xu and Lu, 2020). By now, the researches of beta-defensins mainly focus on infection (Xu and Lu, 2020) and reproductive (Batra et al., 2019), and there is no report on the relationship between beta-defensins and lipid metabolism. Although not directly be regulated in lipid metabolism, the latter four diagnostic genes have been reported to be involved in inflammation. Infection and inflammation are associated with marked changes in lipid and lipoprotein metabolism (Khovidhunkit et al., 2004). Those changes may through effecting on liver lipid synthesis, adipose tissue lipolysis, and postprandial lipid clearance lead to dyslipidemia in OSA (Barros and García-Río, 2019).

To further explore the reason for the changes of immune molecules, we analyzed the immune microenvironment of adipose tissue in the OSA group and then seek out the relationship of diagnostic immune genes and immune cells. In our research, monocytes were mainly enriched kinds of immune cells. Among the kind of monocytes, the proportion of macrophages M0 in the OSA group was significantly higher than that in the normal group, while the ratios of macrophages M1 were almost no expression. This suggested that the major immune cell involved in dyslipidemia of OSA is the macrophage. Macrophages, belonging to the monocyte-macrophage system, modulate inflammatory responses and microbial killing. Macrophages need to display function plasticity to respond to different microenvironmental. Inflammatory stimuli such as lipopolysaccharide (LPS) and interferon-γ (IFN-γ) induce classically activated (M1) macrophages, and anti-inflammatory cytokines such as interleukin-4 (IL-4) or IL-13 induce an alternatively activated (M2) macrophages (Motwani and Gilroy, 2015; Saha et al., 2017). The lipid metabolism signaling pathway and its products play a key role in regulating macrophage polarization (Saha et al., 2017). In *in vitro* experiment, M2 macrophages depend on fatty acid oxidation whereas M1 macrophages depend on an increase in glycolysis (Tabas and Bornfeldt, 2016; Saha et al., 2017; Shimobayashi et al., 2018). The relationship between macrophage phenotypic states and pathological conditions of metabolism disease have been demonstrated in numerous studies. Macrophages play important roles in all stages of atherosclerosis, and pure M1 and M2 macrophages almost certainly do not occur in atherosclerotic lesions. Early in the disease, macrophages accumulate in susceptible regions of arteries. When macrophages are exposed to a plethora of stimuli, they differentiate into different types and play different roles (Tabas and Bornfeldt, 2016). M1 macrophages dominated the rupture-prone shoulder regions of the plaque while increasing M2

activation was displayed in vascular adventitial tissue (Motwani and Gilroy, 2015). Obesity is considered chronic tissue information and causes insulin resistance. In obesity, the balance is tilted toward the M1-like macrophage polarization state (Shimobayashi et al., 2018). In our study, extremely no expression of macrophages M1 further confirmed that dyslipidemia in OSA is not simply caused by obesity. We found that the cells of the monocyte-macrophage system were mainly composed of monocytes, macrophages M0, and macrophages M0 in adipose tissue of OSA. Only macrophages M0 was significantly different between normal and OSA group, and it can onset of polarization adopting variable states of activation.

In adipose tissue, the second immune cell type was CD8 T cells. CD8 T cells are key members of adaptive immunity and immunological memory. The control of lipid metabolism is central to the appropriate differentiation and functions of T lymphocytes (Howie et al., 2017), and it was according to our result that T cell pathways mainly enriched in adipose tissue of OSA. In early atherosclerosis, CD8 T cells control monopoiesis and macrophage accumulation and contribute to macrophage cell death in atherosclerotic plaques (Schäfer and Zernecke, 2020). While in our study, CD8 T cells were positively correlated with Macrophages M0, Macrophages M2, and negatively correlated with monocytes. Regulatory pathways between the macrophage subsets and other immune cells need to be further studied to help us better understand the mechanism of dyslipidemia in OSA.

In the animal model experiment, the expression trend of *LGR5, EPGN, PGRMC2,* and *VIP* were consistent with the conclusion of the testing set (GSE38792) and the training set (GSE135917). Notably, *NCK1* had the opposite trend with the conclusion of the two sets. Perhaps *NCK1* is a good molecule of penetration. As a prognostic gene associated with various immune cells, *NCK1* is highly expressed in adipose tissue. *NCK1* covers aspects of tissue development and homeostasis, invasiveness of tumor cells, and immune cell function. When T-cell antigen receptor (TCR) is triggered, *NCK1* is recruited to the CD3ε subunit of the TCR then switches on downstream T-cell activation pathway (Wipa et al., 2020). In our study, *NCK1* was significantly negatively correlated with monocytes and positively correlated with macrophages M1. Perhaps because the activated T cell response of *NCK1* initiates activated inflammatory macrophages (M1), leading to the corresponding decrease of monocytes.

Moreover, we analyzed candidate transcription factors and target drugs for the prognostic genes. We discovered that *LGR5* might be the likely target gene for the treatment of dyslipidemia in OSA. *LGR5* was the only predictive gene with a significant difference in our study and the two datasets. It also showed a positive correlation with monocytes. Further analyses were necessary to analyze the role of *LGR5* in the polarization of macrophages in the lipid metabolism of OSA.

The current work is the first to investigate the role of immune-related genes in the pathogenesis of dyslipidemia in OSA patients through bioinformatics methods. However, there were still several limitations to our research. First, to clarify the role of lipid metabolism in OSA, all kinds of clinical factors should be considered, such as whether obesity, the severity of intermittent hypoxia, daily physical activity, and fasting-fed state. Second, we have not verified the expression of key immune genes in clinical samples. At the last, the sample size of our data was relatively small, which may affect the gene expression in OSA. More clinical characteristics of OSA are needed to be included in the study for further analysis. The public datasets in our study had small sample size of control group and there was gender bisa. Therefore, we aim to use more samples and perform further experiments to confirm the potential mechanism and clinical utility. First, it would be interesting to examine basic expression of these predicted genes with western blot, IHC, IF assays and so on. Second, to clarify the function of the genes in OSA, a clean loss-of-function and gain-on-function study with tissue-type specificity and cell-type specificity remains warranted. A recent series of molecular experiments may prove strong evidence for the possible phenotype and pathway regulation of these predicted genes. Third, the co-expression and interaction among predicted genes is a new exciting Frontier that awaits further investigation.

CONCLUSION

The clinical evidence confirms the link between OSA and dyslipidemia. However, because of a bias for clinicians who may not consider routine screening for OSA in lean individuals, some dyslipidemia caused by OSA cannot be identified early. Our study demonstrated strong associations between immune genes and the development of dyslipidemia in OSA. Six prognostic genes were found and showed great testing efficacy. The analysis between immune filtration landscape and prognostic genes revealed those genes may affect the polarization of macrophage and differentiation of T lymphocyte subsets brought about abnormal lipid metabolism in OSA. This work will contribute to explore the relationship between inflammation and dyslipidemia, thus promoting our understanding of the molecular mechanisms of lipid metabolism in OSA and may offer potential targets for the regulation of lipid metabolism and treatment of adipose dysfunction. Additionally, the expression of *EPGN*, *LGR5*, *NCK1, PGRMC2*, and *VIP* were also comfired at transcriptional using qPCR, indicating that those genes are closely linked to dyslipidemia in OSA. However, the limitations of our study are lacking in the function validation of the genes such as loss-of-function and gain-on-function study. Therefore, further functional investigations of these targets are essential. The roles of the prognostic genes in different macrophage subsets and other immune cell metabolic properties require further experimental validation.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: http://www.ncbi.nlm.nih.gov/geo/GSE135917 and GSE38792

ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional Ethics Committee for Animal Research of Tongji Medical College, Huazhong University of Science and Technology.

AUTHOR CONTRIBUTIONS

DB conceived and designed the study and performed data analysis; LP contributed analysis tools and wrote the paper; XW conducted the experiment of animal model and qPCR. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Fundation of China (grant numbers 81500794); and the China Postdoctoral Science Fundation (grant numbers 2017M613326).

ACKNOWLEDGMENTS

The authors would like to express their sincere thanks to snowball who has lent her hands in our paper; and we are also grateful for all those who have contributed to improving the value of our paper.

REFERENCES

Alterki, A., Joseph, S., Thanaraj, T. A., Al-Khairi, I., Cherian, P., Channanath, A., et al. (2020). Targeted Metabolomics Analysis on Obstructive Sleep Apnea Patients after Multilevel Sleep Surgery. *Metabolites* 10 (9), 358. doi:10.3390/metabo10090358

Barros, D., and García-Río, F. (2019). Obstructive Sleep Apnea and Dyslipidemia: from Animal Models to Clinical Evidence. *Sleep* 42 (3). doi:10.1093/sleep/zsy236

Batra, V., Maheshwarappa, A., Dagar, K., Kumar, S., Soni, A., Kumaresan, A., et al. (2019). Unusual Interplay of Contrasting Selective Pressures on β-defensin Genes Implicated in Male Fertility of the Buffalo (Bubalus Bubalis). *BMC Evol. Biol.* 19 (1), 214. doi:10.1186/s12862-019-1535-8

Bhattacharya, S., Andorf, S., Gomes, L., Dunn, P., Schaefer, H., Pontius, J., et al. (2014). ImmPort: Disseminating Data to the Public for the Future of Immunology. *Immunol. Res.* 58 (2-3), 234–239. doi:10.1007/s12026-014-8516-1

Cao, Y., Cai, X., Zhu, Q., and Li, N. (2021). Screening and Identification of Potential Biomarkers for Obstructive Sleep Apnea via Microarray Analysis. *Medicine (Baltimore)* 100 (4), e24435. doi:10.1097/md.0000000000024435

Carvalho, B. S., and Irizarry, R. A. (2010). A Framework for Oligonucleotide Microarray Preprocessing. *Bioinformatics* 26, 2363–2367. doi:10.1093/bioinformatics/btq431

Chen, J. H., Yang, R., Wang, Y. P., and Zhang, W. (2015). Expression Data Analysis to Identify Key Target Genes in Visceral Fat Tissue Associated with Obstructive Sleep Apnea. *Eur. Rev. Med. Pharmacol. Sci.* 19 (22), 4293–4299.

Chung, G. E., Cho, E. J., Yoo, J.-J., Chang, Y., Cho, Y., Park, S.-H., et al. (2021). Nonalcoholic Fatty Liver Disease Is Associated with the Development of Obstructive Sleep Apnea. *Sci. Rep.* 11 (1), 13473. doi:10.1038/s41598-021-92703-0

Cotto, K. C., Wagner, A. H., Feng, Y.-Y., Kiwala, S., Coffman, A. C., Spies, G., et al. (2018). DGIdb 3.0: a Redesign and Expansion of the Drug-Gene Interaction Database. *Nucleic Acids Res.* 46 (D1), D1068–d1073. doi:10.1093/nar/gkx1143

Davis, S., and Meltzer, P. S. (2007). GEOquery: a Bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi:10.1093/bioinformatics/btm254

Galmozzi, A., Kok, B. P., Kim, A. S., Montenegro-Burke, J. R., Lee, J. Y., Spreafico, R., et al. (2019). PGRMC2 Is an Intracellular Haem Chaperone Critical for Adipocyte Function. *Nature* 576 (7785), 138–142. doi:10.1038/s41586-019-1774-2

Gharib, S. A., Hayes, A. L., Rosen, M. J., and Patel, S. R. (2013). A Pathway-Based Analysis on the Effects of Obstructive Sleep Apnea in Modulating Visceral Fat Transcriptome. *Sleep* 36 (1), 23–30. doi:10.5665/sleep.2294

Gharib, S. A., Hurley, A. L., Rosen, M. J., Spilsbury, J. C., Schell, A. E., Mehra, R., et al. (2020). Obstructive Sleep Apnea and CPAP Therapy Alter Distinct Transcriptional Programs in Subcutaneous Fat Tissue. *Sleep* 43 (6). doi:10.1093/sleep/zsz314

Graness, A., Chwieralski, C. E., Reinhold, D., Thim, L., and Hoffmann, W. (2002). Protein Kinase C and ERK Activation Are Required for TFF- Peptide-Stimulated Bronchial Epithelial Cell Migration and Tumor Necrosis Factor-α-Induced Interleukin-6 (IL-6) and IL-8 Secretion. *J. Biol. Chem.* 277 (21), 18440–18446. doi:10.1074/jbc.M200468200

Gu, X., Yang, W., Luo, X., Wang, X., Tang, J., and Cai, Z. (2019). Bioinformatics Analysis to Reveal the Key Genes Related to Obstructive Sleep Apnea. *Sleep Breath* 23 (1), 259–267. doi:10.1007/s11325-018-1694-7

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinformatics* 14, 7. doi:10.1186/1471-2105-14-7

Howie, D., Ten Bokum, A., Necula, A. S., Cobbold, S. P., and Waldmann, H. (2017). The Role of Lipid Metabolism in T Lymphocyte Differentiation and Survival. *Front. Immunol.* 8, 1949. doi:10.3389/fimmu.2017.01949

Imani, M. M., Sadeghi, M., Khazaie, H., Emami, M., Sadeghi Bahmani, D., and Brand, S. (2020). Evaluation of Serum and Plasma Interleukin-6 Levels in Obstructive Sleep Apnea Syndrome: A Meta-Analysis and Meta-Regression. *Front. Immunol.* 11, 1343. doi:10.3389/fimmu.2020.01343

Janky, R. s., Verfaillie, A., Imrichová, H., Van de Sande, B., Standaert, L., Christiaens, V., et al. (2014). iRegulon: from a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *Plos Comput. Biol.* 10 (7), e1003731. doi:10.1371/journal.pcbi.1003731

Jordan, A. S., McSharry, D. G., and Malhotra, A. (2014). Adult Obstructive Sleep Apnoea. *The Lancet* 383 (9918), 736–747. doi:10.1016/s0140-6736(13)60734-5

Karkinski, D., Georgievski, O., Dzekova-Vidimliski, P., Milenkovic, T., and Dokic, D. (2017). Obstructive Sleep Apnea and Lipid Abnormalities. *Open Access Maced J. Med. Sci.* 5 (1), 19–22. doi:10.3889/oamjms.2017.011

Kennedy, R. B., Ovsyannikova, I. G., Shane Pankratz, V., Haralambieva, I. H., Vierkant, R. A., and Poland, G. A. (2012). Genome-wide Analysis of Polymorphisms Associated with Cytokine Responses in Smallpox Vaccine Recipients. *Hum. Genet.* 131 (9), 1403–1421. doi:10.1007/s00439-012-1174-2

Khovidhunkit, W., Kim, M.-S., Memon, R. A., Shigenaga, J. K., Moser, A. H., Feingold, K. R., et al. (2004). Thematic Review Series: The Pathogenesis of Atherosclerosis. Effects of Infection and Inflammation on Lipid and Lipoprotein Metabolism Mechanisms and Consequences to the Host. *J. Lipid Res.* 45 (7), 1169–1196. doi:10.1194/jlr.R300019-JLR200

Kursa, Miron. B., and Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *J. Stat. Softw. Found. Open Access Stat.* 36 (11), 1–13. doi:10.18637/jss.v036.i11

Liaw, A., and Wiener, M. (2001). Classification and Regression by randomForest. *R. News* 2, 18–22.

Liu, X., Ma, Y., Ouyang, R., Zeng, Z., Zhan, Z., Lu, H., et al. (2020). The Relationship between Inflammation and Neurocognitive Dysfunction in Obstructive Sleep Apnea Syndrome. *J. Neuroinflammation* 17 (1), 229. doi:10.1186/s12974-020-01905-2

Liu, Y., Patel, S., Nibbe, R., Maxwell, S., Chowdhury, S. A., Koyuturk, M., et al. (2010). Systems Biology Analyses of Gene Expression and Genome Wide Association Study Data in Obstructive Sleep Apnea. *Pac. Symp. Biocomput*, 14–25. doi:10.1142/9789814335058_0003

Luo, W., and Brouwer, C. (2013). Pathview: an R/Bioconductor Package for Pathway-Based Data Integration and Visualization. *Bioinformatics* 29 (14), 1830–1831. doi:10.1093/bioinformatics/btt285

Martínez-Cerón, E., Casitas, R., Galera, R., Sánchez-Sánchez, B., Zamarrón, E., Garcia-Sanchez, A., et al. (2021). Contribution of Sleep Characteristics to the

Association between Obstructive Sleep Apnea and Dyslipidemia. *Sleep Med.* 84, 63–72. doi:10.1016/j.sleep.2021.05.012

Mering, C. v., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a Database of Predicted Functional Associations between Proteins. *Nucleic Acids Res.* 31 (1), 258–261. doi:10.1093/nar/gkg034

Mesarwi, O. A., Loomba, R., and Malhotra, A. (2019). Obstructive Sleep Apnea, Hypoxia, and Nonalcoholic Fatty Liver Disease. *Am. J. Respir. Crit. Care Med.* 199 (7), 830–841. doi:10.1164/rccm.201806-1109TR

Motwani, M. P., and Gilroy, D. W. (2015). Macrophage Development and Polarization in Chronic Inflammation. *Semin. Immunol.* 27 (4), 257–266. doi:10.1016/j.smim.2015.07.002

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Patel, S. R. (2019). Obstructive Sleep Apnea. *Ann. Intern. Med.* 171 (11), Itc81–itc96. doi:10.7326/aitc201912030

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007

Saha, S., Shalova, I. N., and Biswas, S. K. (2017). Metabolic Regulation of Macrophage Phenotype and Function. *Immunol. Rev.* 280 (1), 102–111. doi:10.1111/imr.12603

Schäfer, S., and Zernecke, A. (2020). CD8+ T Cells in Atherosclerosis. *Cells* 10 (1), 37. doi:10.3390/cells10010037

Senaratna, C. V., Perret, J. L., Lodge, C. J., Lowe, A. J., Campbell, B. E., Matheson, M. C., et al. (2017). Prevalence of Obstructive Sleep Apnea in the General Population: A Systematic Review. *Sleep Med. Rev.* 34, 70–81. doi:10.1016/j.smrv.2016.07.002

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Shimobayashi, M., Albert, V., Woelnerhanssen, B., Frei, I. C., Weissenberger, D., Meyer-Gerspach, A. C., et al. (2018). Insulin Resistance Causes Inflammation in Adipose Tissue. *J. Clin. Invest.* 128 (4), 1538–1550. doi:10.1172/jci96139

Silva, L. O. e., Guimarães, T. M., Luz, G. P., Coelho, G., Badke, L., Almeida, I. R., et al. (2018). Metabolic Profile in Patients with Mild Obstructive Sleep Apnea. *Metab. Syndr. Relat. Disord.* 16 (1), 6–12. doi:10.1089/met.2017.0075

Simon, B., Gabor, B., Barta, I., Paska, C., Boszormenyi Nagy, G., Vizi, E., et al. (2020). Effect of 5-year Continuous Positive Airway Pressure Treatment on the Lipid Profile of Patients with Obstructive Sleep Apnea: A Pilot Study. *J. Sleep Res.* 29 (2), e12874. doi:10.1111/jsr.12874

Sun, X., Guo, C., Zhao, F., Zhu, J., Xu, Y., Liu, Z.-Q., et al. (2019). Vasoactive Intestinal Peptide Stabilizes Intestinal Immune Homeostasis through Maintaining Interleukin-10 Expression in Regulatory B Cells. *Theranostics* 9 (10), 2800–2811. doi:10.7150/thno.34414

Tabas, I., and Bornfeldt, K. E. (2016). Macrophage Phenotype and Function in Different Stages of Atherosclerosis. *Circ. Res.* 118 (4), 653–667. doi:10.1161/circresaha.115.306256

Tan, X., van Egmond, L., Chapman, C. D., Cedernaes, J., and Benedict, C. (2018). Aiding Sleep in Type 2 Diabetes: Therapeutic Considerations. *Lancet Diabetes Endocrinol.* 6 (1), 60–68. doi:10.1016/s2213-8587(17)30233-4

Wipa, P., Paensuwan, P., Ngoenkam, J., Woessner, N. M., Minguet, S., Schamel, W. W., et al. (2020). Actin Polymerization Regulates Recruitment of Nck to CD3 ε upon T-cell Receptor Triggering. *Immunology* 159 (3), 298–308. doi:10.1111/imm.13146

Xu, D., and Lu, W. (2020). Defensins: A Double-Edged Sword in Host Immunity. *Front. Immunol.* 11, 764. doi:10.3389/fimmu.2020.00764

Ye, J., Liu, H., Zhang, G., Li, P., Wang, Z., Huang, S., et al. (2012). The Treg/th17 Imbalance in Patients with Obstructive Sleep Apnoea Syndrome. *Mediators Inflamm.* 2012, 1–11. doi:10.1155/2012/815308

Yeghiazarians, Y., Jneid, H., Tietjens, J. R., Redline, S., Brown, D. L., El-Sherif, N., et al. (2021). Obstructive Sleep Apnea and Cardiovascular Disease: A Scientific Statement from the American Heart Association. *Circulation* 144. doi:10.1161/cir.0000000000000988

Ying, L., Li, H., Pan, Z., Ma, S., Zhang, P., Wang, Q., et al. (2014). Relationship of Redundant Th17 Cells and IL-17A, but Not IL-17 F, with the Severity of Obstructive Sleep Apnoea/hypopnoea Syndrome (OSAHS). *BMC Pulm. Med.* 14, 84. doi:10.1186/1471-2466-14-84

Yu, G. (2020). Gene Ontology Semantic Similarity Analysis Using GOSemSim. *Methods Mol. Biol.* 2117, 207–215. doi:10.1007/978-1-0716-0301-7_11

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R Package for Measuring Semantic Similarity Among GO Terms and Gene Products. *Bioinformatics* 26 (7), 976–978. doi:10.1093/bioinformatics/btq064

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Yu, G., Wang, L.-G., Yan, G.-R., and He, Q.-Y. (2015). DOSE: an R/Bioconductor Package for Disease Ontology Semantic and Enrichment Analysis. *Bioinformatics* 31 (4), 608–609. doi:10.1093/bioinformatics/btu684

Zhang, J., Cai, H., Sun, L., Zhan, P., Chen, M., Zhang, F., et al. (2018). LGR5, a Novel Functional Glioma Stem Cell Marker, Promotes EMT by Activating the Wnt/β-Catenin Pathway and Predicts Poor Survival of Glioma Patients. *J. Exp. Clin. Cancer Res.* 37 (1), 225. doi:10.1186/s13046-018-0864-6

Zhao, K., Lu, Y., Chen, Y., Cheng, J., and Zhang, W. (2020). Transcripts 202 and 205 of IL-6 Confer Resistance to Vemurafenib by Reactivating the MAPK Pathway in BRAF(V600E) Mutant Melanoma Cells. *Exp. Cel Res.* 390 (2), 111942. doi:10.1016/j.yexcr.2020.111942

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Peng, Wang and Bing. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

OPINION
published: 07 December 2021
doi: 10.3389/fgene.2021.801653

Benefits of Cohort Studies in a Consortia-Dominated Landscape

*Selam Zenebe-Gete, Rebecca Salowe and Joan M. O'Brien**

Department of Ophthalmology, Scheie Eye Institute, University of Pennsylvania, Philadelphia, PA, United States

Keywords: GWAS - genome-wide association study, consortia, cohort, study design, genetics, big data

INTRODUCTION

Since their inception in 2005, Genome-Wide Association Studies (GWAS) have risen in popularity as a key method for gene discovery. Efforts to make use of existing genetic data, identify the causes of complex disease, and improve study power by increasing sample sizes have led to the formation of many GWAS consortia (Lutz et al., 2013). Consortia can bypass major logistical and financial challenges that accompany the recruitment of large study populations (Benjamin et al., 2018). Additionally, data sharing via consortia can help researchers diversify their study populations and can encourage research collaborations. While consortia do circumvent several financial, logistical, statistical, and demographic challenges to conducting GWAS, the collective swing towards their formation leaves several gaps in the field of genetics.

The majority of GWAS, including both small cohort studies and large consortia, often fail to identify actionable genetic determinants of complex disease. Only 2.2% of GWAS conducted between 2005 and 2016 had follow-up functional studies (Gallagher and Chen-Plotkin, 2018). Additionally, over 90% of phenotype-associated single nucleotide polymorphisms (SNPs) discovered in GWAS are in non-coding regions of the genome (Qu and Fang, 2013), with their effects on clinical outcomes either unknown or under investigation. Additionally, of the variants identified as statistically significant, a subset may not be clinically actionable because of their essential roles in cellular function. For example, p53 plays a significant role in many disease pathways (i.e. cancer), but its regulatory functions in all cells makes it a poor target for drug therapy.

Though these limitations apply to most GWAS, there are several unique advantages to cohort studies, including: 1) discovery of clinically actionable targets through subtyping; 2) the necessity of post-GWAS follow-up on variants; and 3) identification of population-specific findings.

SUBSECTIONS

Clinically Actionable Variants

Stratifying study populations by disease subtype may facilitate identification of clinically actionable variants. The discovery of variants through subtype-stratified GWAS has been well documented in the literature. In breast cancer research, stratified GWAS have identified subtype-defining SNPs, such as variants in the *HER2* gene (O'Brien et al., 2014). The development of therapies targeting HER2 pathways improved outcomes for this subtype of breast cancer (Arteaga et al., 2012; Figueroa-Magalhães et al., 2014). Similarly, in an ischemic stroke study, all statistically significant variants were subtype-specific (Traylor et al., 2012). Subtype-stratified GWAS of bipolar disorder (BD) (Charney et al., 2017), autism spectrum disorder (ASD) (Hu et al., 2011), and Parkinson's disease (von Coelln and Shulman, 2016) patients also led to identification of novel disease subtype SNPs.

While subtyping can be implemented in both cohort and consortia studies, cohort studies may more easily accommodate subtyping methods, particularly when investigating complex diseases. Since some complex diseases do not have clearly defined subtypes, individual studies may define

OPEN ACCESS

Edited by:
Dov Greenbaum, Yale University, United States

Reviewed by:
Chun Chieh Fan, University of California, San Diego, United States

****Correspondence:***
Joan M. O'Brien
joan.o'brien@pennmedicine.upenn.edu

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *25 October 2021*
Accepted: *15 November 2021*
Published: *07 December 2021*

Citation:
Zenebe-Gete S, Salowe R and O'Brien JM (2021) Benefits of Cohort Studies in a Consortia-Dominated Landscape. Front. Genet. 12:801653. doi: 10.3389/fgene.2021.801653

subtypes using different criteria. For example, a lack of methodological standardization in neuroimaging of Alzheimer's patients has hindered efforts towards consistent subtyping (Mohanty et al., 2020). Data harmonization methods are then needed to address differences in subtype classification between study cohorts within a consortium, increasing risk of introducing bias. A cohort study avoids this potential source of bias since it operates under a single set of subtyping criteria. Additionally, many cohort studies investigate a single disease type (i.e., open-angle glaucoma) rather than larger combinations (i.e., all forms of glaucoma), allowing for further subtyping of one form of disease. Since different forms of this disease may have different biological causes, putting them together to increase power may yield results that do not have biologic or physiologic relevance. Well-defined subtype-stratified GWAS in cohort studies may improve researchers' attempts to identify clinically actionable disease targets compared to consortia-based GWAS.

Post-Genome-Wide Association Study Follow-Up on Variants

Cohort studies tend to be more amenable to collecting longitudinal data and conducting follow-up on variants of interest (Wijmenga and Zhernakova, 2018). As more variants have been identified by GWAS, interest in investigating them further through functional studies, "multi-omic" research, and other "post-GWAS" methods has peaked. Cohort studies' focus on a single, typically local study population allows researchers to re-contact patients with interesting variants for follow-up. Cohort studies also allow for collection of longitudinal phenotypic data, which is key to the study of disease progression. Since consortia patients originally enrolled in studies at a variety of sites, patient re-contact and collection of longitudinal data are complicated and may not be possible in studies of such a wide scale. Furthermore, information collected at each site may lack consistency in data type and collection method, requiring additional harmonization efforts to make them comparable. For example, individual studies within the Psychiatric GWAS Consortium used different genotyping platforms, requiring imputation against existing gene expression data to standardize their genotypes (Sullivan, 2010). Even basic phenotypic parameters such as age and diet can affect an individual's gene expression (Wijmenga and Zhernakova, 2018), highlighting another difference between populations that can complicate combining multiple populations in consortia.

Population-Specific Variants

Cohort studies can also aid in identifying variants that are specific to minority populations, who remain dramatically under-represented in genetic studies. Population-specific GWAS are important to both understanding the genetics of complex traits and to elucidating the role of specific variants in minority populations (Sirugo et al., 2019). Studies have shown that differences in ancestry contribute to variations in disease prevalence, severity, and resistance across populations (Haga, 2010). However, while extensive GWAS testing has been done among individuals of European-descent, investigations of similar scale have not been conducted in African ancestry populations (Campbell and Tishkoff, 2008), with only 3% of GWAS participants being of African descent as of 2016 (Popejoy and Fullerton, 2016).

Although some consortia can increase study population diversity through collaborative efforts to catalog human genetic variation, such as the 1000 Genomes Project (Jankovic et al., 2010), extensive data compilation may prevent identification of variants that are specific to underrepresented groups when they are un-stratified within the larger population. For example, an investigation of 3,899 SNPs in 313 genes in self-identified Caucasians, African Americans, Asians, and Hispanics found distinct and non-overlapping clustering of the Caucasian, African American and Asian samples (Stephens et al., 2001). This finding suggests differing genetic architecture between these groups, supporting the need for ancestry-specific genetic studies. Additionally, recruiting patients from a single city, as many cohorts do, minimizes the differences within a demographic group. Self-identified African Americans differ in their genetic admixture across different geographic locations (Bryc et al., 2015), suggesting that single city studies have less population heterogeneity than consortia.

DISCUSSION

While consortia do play an important role in genetic research, cohort-based studies may be better suited to identifying clinically actionable disease pathways and studying underrepresented minority populations. Although the recent publication climate emphasizes large consortia, GWAS of specified cohorts may produce more precise results that can be used in studies aiming to link genetics to endophenotypic data over time. Returning to the example of breast cancer, underlying pathology and genetics are now used to subtype patients in order to personalize treatment. One consortium including over 120,000 breast cancer patients identified 65 novel loci associated with overall breast cancer risk via GWAS (Michailidou et al., 2017). However, advancements in breast cancer technology have demonstrated the importance of molecular subtyping in patient prognosis and treatment (Yang and Polley, 2019). When multiple diseases or subtypes with distinct molecular pathways are categorized together in GWAS, as in the above study, the resulting genetic findings may not be clinically actionable. Putting this into context, future genetic studies may benefit from refocusing on the end goal of all GWAS—not only to find statistical significance, but to identify variants with the potential to improve health outcomes. Improving the accuracy of GWAS findings and translating these results to the clinic may be facilitated through a greater balance between both consortia- and cohort-based methods.

AUTHOR CONTRIBUTIONS

JO'B conceptualized the presented idea. SZ-G investigated and wrote the original draft. All authors reviewed and edited later drafts. All authors reviewed and approved the final version of the manuscript.

FUNDING

This work was supported by the National Eye Institute, Bethesda, Maryland (grant #1RO1EY023557-01) and Vision Research Core Grant (P30 EY001583). Funds also come from the F. M. Kirby Foundation, Research to Prevent Blindness, The UPenn Hospital Board of Women Visitors, The Paul and Evanina Bell Mackall Foundation Trust, and Regeneron Pharmaceuticals. The Ophthalmology Department at the Perelman School of Medicine and the VA Hospital in Philadelphia, PA also provided support.

REFERENCES

Arteaga, C. L., Sliwkowski, M. X., Osborne, C. K., Perez, E. A., Puglisi, F., and Gianni, L. (2012). Treatment of HER2-Positive Breast Cancer: Current Status and Future Perspectives. *Nat. Rev. Clin. Oncol.* 9 (1), 16–32. doi:10.1038/nrclinonc.2011.177

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., et al. (2018). Redefine Statistical Significance. *Nat. Hum. Behav.* 2 (1), 6–10. doi:10.1038/s41562-017-0189-z

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., and Mountain, J. L. (2015). The Genetic Ancestry of African Americans, Latinos, and European Americans Across the United States. *Am. J. Hum. Genet.* 96 (1), 37–53. doi:10.1016/j.ajhg.2014.11.010

Campbell, M. C., and Tishkoff, S. A. (2008). African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annu. Rev. Genom. Hum. Genet.* 9, 403–433. doi:10.1146/annurev.genom.9.081307.164258

Charney, A. W., Ruderfer, D. M., Stahl, E. A., Moran, J. L., Chambert, K., Belliveau, R. A., et al. (2017). Evidence for Genetic Heterogeneity Between Clinical Subtypes of Bipolar Disorder. *Transl. Psychiatry* 7 (1), e993. doi:10.1038/tp.2016.242

Figueroa-Magalhães, M. C., Jelovac, D., Connolly, R. M., and Wolff, A. C. (2014). Treatment of HER2-Positive Breast Cancer. *Breast* 23 (2), 128–136. doi:10.1016/j.breast.2013.11.011

Gallagher, M. D., and Chen-Plotkin, A. S. (2018). The post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* 102 (5), 717–730. doi:10.1016/j.ajhg.2018.04.002

Haga, S. B. (2010). Impact of Limited Population Diversity of Genome-Wide Association Studies. *Genet. Med.* 12 (2), 81–84. doi:10.1097/gim.0b013e3181ca2bbf

Hu, V. W., Addington, A., and Hyman, A. (2011). Novel Autism Subtype-Dependent Genetic Variants are Revealed by Quantitative Trait and Subphenotype Association Analyses of Published GWAS Data. *PLoS One* 6 (4), e19067. doi:10.1371/journal.pone.0019067

Lutz, S. M., Fingerlin, T., and Fardo, D. W. (2013). Statistical Approaches to Combine Genetic Association Data. *J. Biomet. Biostat.* 04 (3), 1000166. doi:10.4172/2155-6180.1000166

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Jui, S., and Kar, S. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* 551 (7678), 92–94. doi:10.1038/nature24284

Mohanty, R., Mårtensson, G., Poulakis, K., Muehlboeck, J.-S., Rodriguez-Vieitez, E., Chiotis, K., et al. (2020). Comparison of Subtyping Methods for Neuroimaging Studies in Alzheimer's Disease: A Call for Harmonization. *Brain Commun.* 2 (2), fcaa192. doi:10.1093/braincomms/fcaa192

O'Brien, K. M., Cole, S. R., Engel, L. S., Bensen, J. T., Poole, C., Herring, A. H., et al. (2014). Breast Cancer Subtypes and Previously Established Genetic Risk Factors: A Bayesian Approach. *Cancer Epidemiol. Biomarkers Prev.* 23 (1), 84–97. doi:10.1158/1055-9965.EPI-13-0463

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics is Failing on Diversity. *Nature* 538 (7624), 161–164. doi:10.1038/538161a

Qu, H., and Fang, X. (2013). A Brief Review on the Human Encyclopedia of DNA Elements (ENCODE) Project. *Genomics Proteomics Bioinformatics* 11 (3), 135–141. doi:10.1016/j.gpb.2013.05.001

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. (2010). Genome-Wide Association Studies in Diverse Populations. *Nat. Rev. Genet.* 11 (5), 356–366. doi:10.1038/nrg2760

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell* 177 (1), 26–31. doi:10.1016/j.cell.2019.02.048

Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., et al. (2001). Haplotype Variation and Linkage Disequilibrium in 313 Human Genes. *Science* 293, 489–493. doi:10.1126/science.1059431

Sullivan, P. F. (2010). The Psychiatric GWAS Consortium: Big Science Comes to Psychiatry. *Neuron* 68 (2), 182–186. doi:10.1016/j.neuron.2010.10.003

Traylor, M., Farrall, M., Holliday, E. G., Sudlow, C., Hopewell, J. C., Cheng, Y.-C., et al. (2012). Genetic Risk Factors for Ischaemic Stroke and its Subtypes (The METASTROKE Collaboration): A Meta-Analysis of Genome-Wide Association Studies. *Lancet Neurol.* 11 (11), 951–962. doi:10.1016/S1474-4422(12)70234-X

von Coelln, R., and Shulman, L. M. (2016). Clinical Subtypes and Genetic Heterogeneity: Of Lumping and Splitting in Parkinson Disease. *Curr. Opin. Neurol.* 29 (6), 727–734. doi:10.1097/WCO.0000000000000384

Wijmenga, C., and Zhernakova, A. (2018). The Importance of Cohort Studies in the Post-GWAS Era. *Nat. Genet.* 50 (3), 322–328. doi:10.1038/s41588-018-0066-3

Yang, S. X., and Polley, E. C. (2019). Systemic Treatment and Radiotherapy, Breast Cancer Subtypes, and Survival After Long-Term Clinical Follow-Up. *Breast Cancer Res. Treat.* 175 (2), 287–295. doi:10.1007/s10549-019-05142-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Zenebe-Gete, Salowe and O'Brien. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 December 2021
doi: 10.3389/fgene.2021.736441

Identification and Functional Verification Reveals that miR-195 Inhibiting *THRSP* to Affect Fat Deposition in Xinyang Buffalo

*Shuzhe Wang[1,2], Cuili Pan[1], Xiaojie Ma[2], Chaoyun Yang[1], Lin Tang[1], Jieping Huang[3], Xuefeng Wei[2], Hui Li[3] and Yun Ma[1,2]**

[1]Key Laboratory of Ruminant Molecular and Cellular Breeding of Ningxia Hui Autonomous Region, School of Agriculture, Ningxia University, Yinchuan, China, [2]College of Life Sciences, Xinyang Normal University, Xinyang, China, [3]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Animal Science and Technology, Guangxi University, Nanning, China

The buffalo population is extensive in China, but its meat quality is relatively inferior. Therefore, improving meat quality should be one of the breeding goals. microRNAs (miRNAs) play an essential regulatory role in the post-transcriptional expression of genes. Some studies have reported their function regulating genes related to fat deposition and adipocyte differentiation in cattle, but there is limited reports in buffalo. We performed small RNA transcriptome sequencing of Xinyang buffalo adipose tissue between calves and adults in this study. As a result, 282 mature miRNAs were significantly differentially expressed, and co-expression analysis showed that 454 miRNAs were significantly associated with developmental stages. Target gene identification, GO (gene ontology) annotation, and KEGG analysis of miRNAs showed that miR-195, miR-192, and miR-24-3p could target key genes for lipogenesis and thus regulate adipose deposition and differentiation. Among them, miR-195 was significantly upregulated in adipose tissue and induced adipocytes of adult buffaloes, and its overexpression significantly inhibited lipid accumulation in primary adipocytes. Dual-luciferase reporter gene analysis showed that miR-195 reduced the expression of thyroid hormone response protein (THRSP) by targeting its 3′ untranslated terminal region, suggesting that miR-195 may inhibit lipid accumulation in adipocytes by regulating *THRSP*. The results confirmed the reliability of predictive screening of miRNAs and provided theoretical support for buffalo fattening.

Keywords: fat deposition, microRNA sequencing, bta-miR-195, Xinyang buffalo, THRSP

OPEN ACCESS

Edited by:
William C. Cho, QEH, Hong Kong SAR, China

Reviewed by:
Guoze Wang, Guizhou Medical University, China
Linsen Zan, Northwest A&F University, China

****Correspondence:***
Yun Ma
mayun_666@126.com

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *05 July 2021*
Accepted: *30 November 2021*
Published: *22 December 2021*

Citation:
Wang S, Pan C, Ma X, Yang C, Tang L, Huang J, Wei X, Li H and Ma Y (2021) Identification and Functional Verification Reveals that miR-195 Inhibiting THRSP to Affect Fat Deposition in Xinyang Buffalo. Front. Genet. 12:736441. doi: 10.3389/fgene.2021.736441

INTRODUCTION

Fat formation is a highly complex process in mammals. Several studies have shown that miRNAs regulated (activation and inhibition) adipocyte differentiation and lipid accumulation (Remsburg et al., 2019). For example, *miR-21* was one of the earliest identified miRNAs which associated with adipogenesis and obesity, upregulated in white adipose tissue of obese subjects (Keller et al., 2011), promoted adipogenic differentiation (Kim et al., 2009), and facilitated lipogenesis (Mei et al., 2013). *miR-33a* and *miR-33b* played essential roles in

controlling cholesterol homeostasis (Najafi-Shoushtari et al., 2010). Over-expression of *miR-33b* could reduce proliferation, impair differentiation, and decrease lipid droplets in precursor adipocytes (Price et al., 2016). *Let-7* could directly target *HMGA2* to inhibit the proliferation and differentiation of 3T3-L1 precursor adipocytes (Wei et al., 2014).

Historically, Asian wild buffalo were found throughout South and Southeast Asia but are classified as endangered, with less than 4,000 worldwide and probably less than 200 purebred Asian wild buffalo (Barker, 2014). Domestic buffalo are mainly distributed in Asian countries and are classified into *Bubalus bubalis* and *Bubalus carabanensis* based on morphology, behavior, and karyotype. The most recent data show that the *Bubalus bubalis* population is growing at a steady rate of 1,800,000 head/year, while the *Bubalus carabanensis* population is declining at a rate of 180,000 head/year (Zhang et al., 2020). It is due to the increase in the dairying of *Bubalus bubalis* and the decrease in the servitude of *Bubalus carabanensis* in agriculture. The Xinyang buffalo is characterized by largeness, robustness, sturdiness, and docility. The adults have a average body weight of more than 500 kg. However, traditional selection and breeding have led to inferior intra-muscular fat deposition ability and rough meat quality (Zuo and Liu, 2008). According to a previous study, the amount of Xinyang buffalo population was less than 150. Promoting the meat performance of Xinyang buffalo is a critical way to improve its breeding status and increase its quantity. Intramuscular fat (IMF), as a determinant of beef tenderness and succulence, is crucial for improving the quality of meat (Yamada et al., 2020). Given the comprehensive and vital roles of miRNAs in fat deposition and lipid accumulation, their expression profiles in humans, mice, pigs, cattle, and buffalo (Sun et al., 2014; Guller et al., 2015; Oclon et al., 2016) have recently been characterized in adipose tissue or adipose-associated cells, providing essential information for further research on fat deposition. The differences in miRNAs between buffalo adipose and muscle tissues have been explored in our previous study (Huang et al., 2019). However, details of the expression profiles and mechanisms of miRNAs in fat accumulation in buffalo are limited and require further investigation.

The growth and development of buffalo can be divided into the embryonic, juvenile, adolescence, and adulthood (Funston et al., 2010) period. The juvenile period lays the foundation for the production performance and physical appearance, and the adult buffalo has a strong body and fat deposition ability (Du et al., 2013). miRNAs in buffalo adipose tissue were sequenced and analyzed at two developmental stages (calf and adult buffalo) to screen for potential miRNAs regulating fat deposition. They were also subjected to weighted gene co-expression analysis to screen the miRNAs associated with growth and development. The conservatism of miRNA target genes between buffalo and cattle was compared, and the miRNA-mRNA interaction network related to adipocyte differentiation was mapped. Furthermore, the expression regulation mechanism of *miR-195* on key adipocyte differentiation genes was analyzed using a dual-luciferase reporter system and overexpression techniques. Our study lays the theoretical foundation for further revealing the mechanism of fat deposition in buffalo and improving the utility value of buffalo meat.

MATERIALS AND METHODS

Ethics Statement

Animal experiments were conducted in accordance with the guidelines of the *Regulations for the Administration of Affairs Concerning Experimental Animals* (Ministry of Science and Technology, China, 2004). It is authorized by the Animal Ethics Committee of Ningxia University (permit number NXUC20190105).

Animals and Sample Collection

The buffaloes were raised under the same dietary and environmental temperature conditions and given free water and food at the Xinyang buffalo farm (Xinyang, Henan, China). Subcutaneous adipose tissue was collected after slaughter and frozen immediately in liquid nitrogen. Three 6-month-old calves and three 30-month-old adult Xinyang buffaloes were slaughter to separate Rib eye upper back subcutaneous fat.

Total RNA Extraction

Frozen tissue samples were homogenized in TRIzol reagent (Ambion, Life Technologies, NY), and total RNA was extracted from the supernatant according to the manufacturer's protocol (Invitrogen, Carlsbad, CA). The quantity and integrity of RNA were confirmed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Only RNA samples with integrity scores ≥7 were used for sequencing. Total RNA was stored at −80°C until further use.

miRNA Bioinformatics Analysis

miRNAs were analyzed as follows: with aptamers removed by cutadapt software (version = V2.7) to obtain clean data; alignment analysis and filtering of remaining sequences using the RFam database (https://rfam.org/); miRDeep2 (version = 2.0.0.8) comparison of precursor sequences to identify miRNAs(Andrés-León et al., 2016).

Differentially expressed analysis of miRNAs was analyzed using the R package DEseq2 (version = 3.6.1) (Love et al., 2014). Target genes of DE miRNAs were predicted by TargetScan7.2 (http://www.targetscan.org/vert_72/), and the functional enrichment analysis of the target genes was performed by the DAVID website (https://david.ncifcrf.gov/) with a p-value ≤ 0.05 as the threshold.

To identify the critical modules and key miRNAs associated with adipose deposition, The R package (version = 4.0.3) WGCNA (Yao et al., 2019) was used.

Validation of DE miRNA by qPCR

After removal of genomic DNA, total RNA was used for cDNA library construction, followed by reverse transcription at 37° 15 min, 85° 5 s; miRNAs were reverse transcribed by replacing

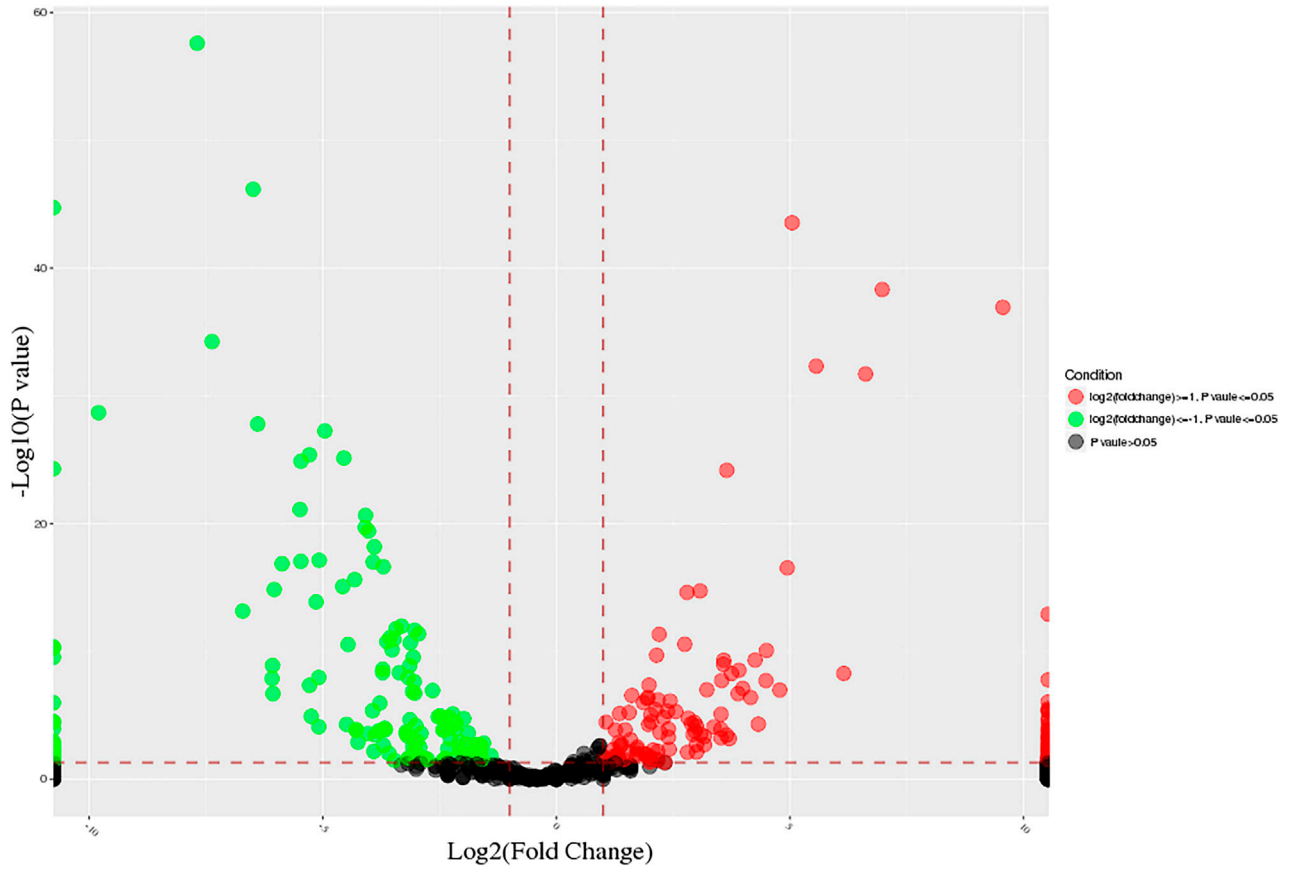


FIGURE 1 | Volcano plot of all expressed miRNAs at different developmental stages. The "green" and "red" dots represent the down- and up-regulation miRNAs in the adult group, respectively.

the universal primer with a loop primer at 42° 15 min, 85° 5 s (Takara RR047A).

The qRT-PCR was performed using pre-denaturation for the 30s, PCR reaction for 90° 5 s, 60° 30 s, 45 cycles, followed by dissociation stage on an Applied Biosystems 7,500 Sequence Detection System (Takara RR0820A).

qRT-PCR was performed using SYBR Green PCR Master Mix on an Applied Biosystems 7,500 Sequence Detection System. The threshold cycle values indicated the quantity of the target gene in each sample. Negative controls used water as the template instead of cDNA. Three independent biological samples were used, and each run was performed in triplicate on plates. All related primers were designed using Primer 5.0 and synthesized by GeneScript (**Supplementary Table S1**).

Cell Culture and Transfection

Subcutaneous adipose tissue from calves' backs was collected for primary adipocyte culture according to the following procedure. After collection, the subcutaneous adipose tissue was immediately placed in 1% double antibodies (Hyclone, Scotland) in PBS (Hyclone, Scotland) and returned to the laboratory. The tissue pieces were washed three times in a sterile environment using PBS containing 1% double antibodies and separated into soybean-sized fat parts. The fat pieces were inoculated in 90 mm Petri dishes with inversion and incubated orthotopically after 6 h in DMEM (Hyclone, Scotland) high sugar medium with 10% fetal bovine serum (Gibico, Australia) and 1% double antibodies. Primary adipocytes can be freed after 4–5 days of culture.

The 293T cell line was maintained in this laboratory. All cells and cell lines were incubated and cultured in a jacketed water incubator at 37°C and 5% CO_2. According to the manufacturer's instructions, primary adipocytes of the third and fourth subculture generation were used for transfection by lipofectamine 3,000 (Thermo, United Staes). Add 10 μg/ml Insulin, 1 μM Dexamethasone, 0.5 mM 3-Isobutyl-1-methylxanthine, and 1 μM rosiglitazone for primary adipocyte induction (sigma, GRE).

Dual-Luciferase Reporter Assay

The wild-type and mutant segments of thyroid hormone response protein (*THRSP*) 3′UTR containing the bta-miR-195 binding site were cloned by gamma two plasmids and used to construct dual-luciferase reporter gene vectors. The 293T cells were used, and the wild-type and mutant plasmids were co-transfected with bta-mir-195 agomiR.

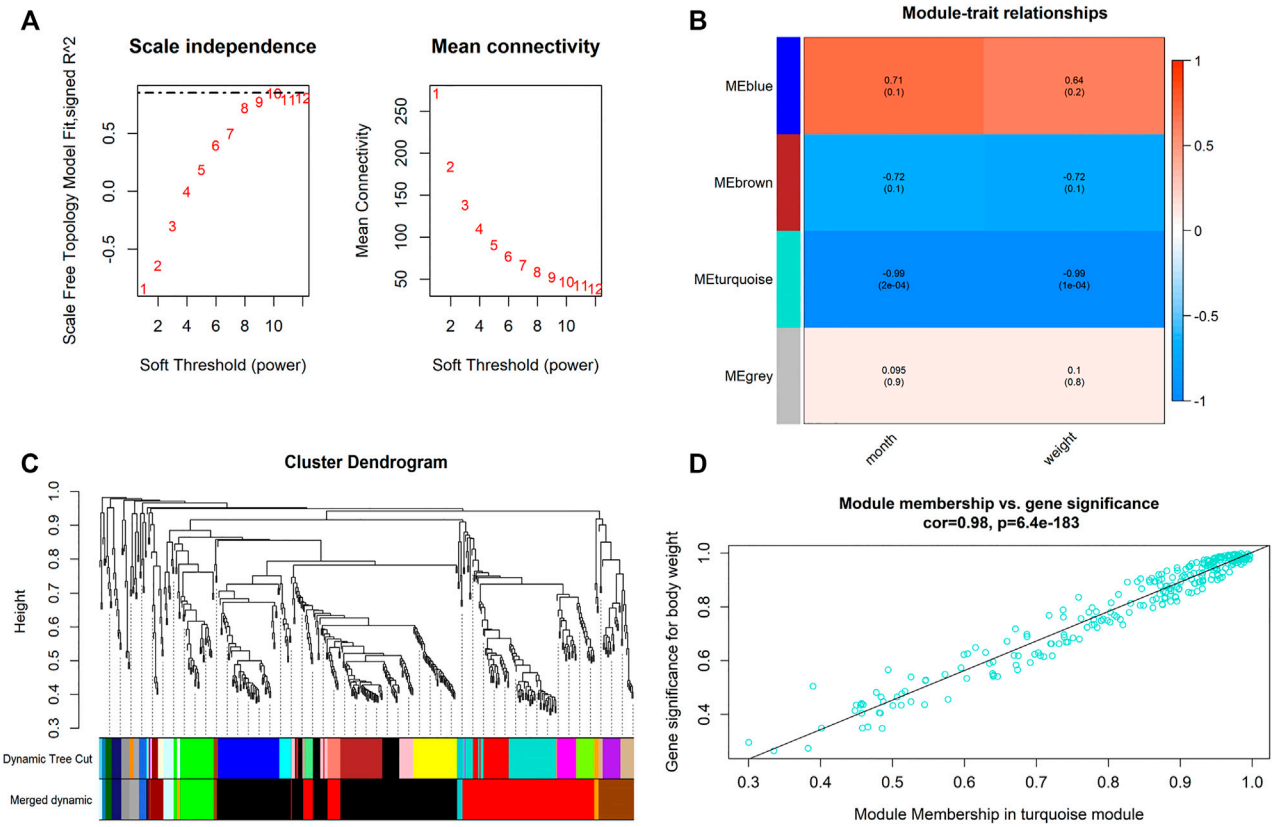


FIGURE 2 | Main findings in the module-trait correlations analyses. **(A)** Analysis of the scale-free fit index for various soft-thresholding powers **(Left)** and analysis of the mean connectivity for various soft-thresholding powers **(Right)**; **(B)** Heatmap between the correlation between modules and weight (Each cell contained the correlation coefficient and corresponding *p* value; **(C)** Clustering dendrogram of differentially expressed genes related to Month and Weight in the deodunal tissues of six Xinyang buffalo; **(D)** The gene significance for weight in the turquoise module (One dot represents one gene in the turquoise module).

Statistical Analysis

qRT-PCR results were calculated using a $2^{-\Delta\Delta Ct}$ method. Statistical significance was tested using Graphpad Prism 7.0 software, and the Significant level was set as $p < 0.05$.

RESULTS

Overview of miRNA Sequencing

Six miRNA libraries of buffalo subcutaneous adipose tissue from three calves and three adults were constructed and divided into two groups. A total of 10 GB of data was obtained. The size of libraries ranges from 10,702,344 to 30,835,129 raw reads. After filtering and aligning to mRNAs, RFam, and Repbase databases, 10,414,935–29,398,118 good reads were obtained. The libraries yielded 180,103–308,002 effective unique copies, accounting for 61.15–69.12% of the total sample (**Supplementary Table S2**). Base preference analysis of the miRNA sequencing (**Supplementary Figures S1A,B**) showed that the identity of the first base had an extremely high proportion of U (98%), which was consistent with the miRNA sequence characteristics.

Length distribution analysis demonstrated that miRNAs in the six libraries showed similar length patterns, varying from 18 nt to 26 nt, with 22 nt being the most common (**Supplementary Figure S1C**). Analysis of specific miRNA lengths was based on filtered datasets. Sequences with a length of 22 nt accounted for 44.13 ± 3.45% of the miRNAs, followed by 21 nt (14.96 ± 1.33%), and 20 nt (15.3 ± 1.64%) in the buffalo calf subcutaneous fat libraries. Sequences of 22 nt accounted for 44.13 ± 3.45%, followed by 22 nt (59.06 ± 7.09%) and 23 nt (6.1 ± 0.15%) in the adult buffalo libraries. These results were consistent with the characteristics of Dicer enzyme cleavage.

Annotation and Identification of miRNAs

Genome annotation showed that 93.92 ± 1.43% of the total miRNAs could be mapped to the bovine reference genome (**Supplementary Table S3**). To obtain conserved miRNAs in subcutaneous fat tissue of buffalo, the ACGT101-miR tool was used to compare mapped reads in the reference genome with the known mature miRNAs in the miRase database. A total of 807 miRNAs (**Supplementary Table S4**) were obtained from the two adipose tissue developmental stages. Analysis of sequence source revealed that the number of miRNAs from exons or introns of coding genes was 13,977,471, accounting for 14% of the total reads. In contrast, rRNA, tRNA, snRNA, and other non-coding RNA accounted for 3.6%, indicating that the quality of the total

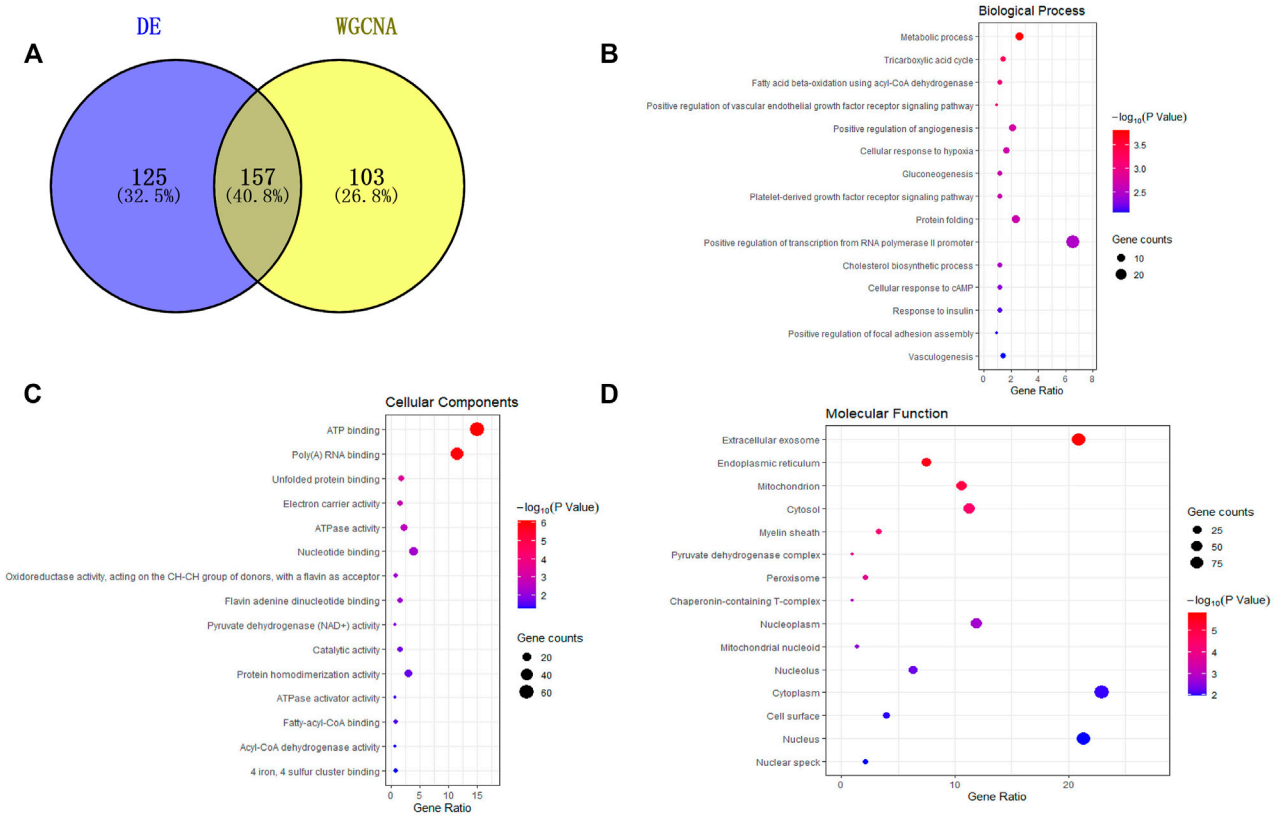


FIGURE 3 | Functional analysis of miRNAs target genes. **(A)** Venn diagram of differentially expressed miRNAs and miRNAs in the turquoise module; **(B–D)** are the gene functional enrichment analysis of the target gene of miRNAs, and B–D are biological process, cellular components, and molecular function, respectively. The dot size represents gene amount enriched in a particular process, and the color represents significance.

RNA was favorable. In addition, miRNA expression varied across the two stages of growth, 665 miRNAs being obtained from adipose tissue of 6-month-old buffalo and 763 from 30-month-old buffalo. According to principal component analysis (PCA) (**Supplementary Figure S2**), the samples were well grouped and had high correlation coefficients. There were 618 overlap miRNAs in the two growth stages. These results showed that the expression of miRNAs was relatively high in the adipose tissue of adult buffalo.

Identification of DE miRNAs

DE miRNAs in calves and adult buffalo were identified using DEseq2 package in R with the screening standard set as | log2FC(fold change) | ≥ 1and adjusted *p*-value (FDR) ≤ 0.05. To elucidate the expression pattern of DE miRNAs, hierarchical clustering was conducted on the 618 overlap miRNAs. A total of 282 significantly differential expression miRNAs were screened from 807 miRNAs in the two groups, of which 148 were up-regulated, and 134 were down-regulated (**Figure 1**). These miRNAs showed significant differential expression during adipogenesis ($p < 0.05$), suggesting that these miRNAs are dynamically regulated during adipogenesis.

Furthermore, there were 282 differential expressed miRNAs across the two ages, of which 11 conserved miRNAs were specific to adult bovine adipose tissue, and 18 miRNAs were specific to calf adipose tissue.

Identification of Hub miRNAs by WGCNA

WGCNA (Weighted Gene Co-expression Network Analysis) is used to identify gene modules with similar expression patterns. It dissects the association between gene sets and sample phenotypes or groups and identifies hub genes. In this study, we performed a WGCNA analysis of miRNAs and associated the identified modules with developmental stages to find hub miRNAs regulating adipose growth and development.

When 0.99 was used as the correlation coefficient threshold, the soft-thresholding power was selected as ten (**Figure 2A**). Through WGCNA analysis, four co-expression modules were constructed (**Figures 2B,C**). The turquoise module was comprising the most miRNAs (508 miRNAs), followed by the blue module (144 miRNAs), the brown module (88 miRNAs), and the grey module (15 miRNAs).

Module-trait correlations analyses were performed between modules and phenotypes after obtaining co-expressed modules (**Figure 2D**). Correlation analyses showed that turquoise modules were significantly associated with both weight and development stages (month) ($p < 0.01$). In this module, *bta-miR-129-5p*, *bta-miR-148a*, *bta-miR-195*, *bta-miR-93*, and *bta-miR-504* had high

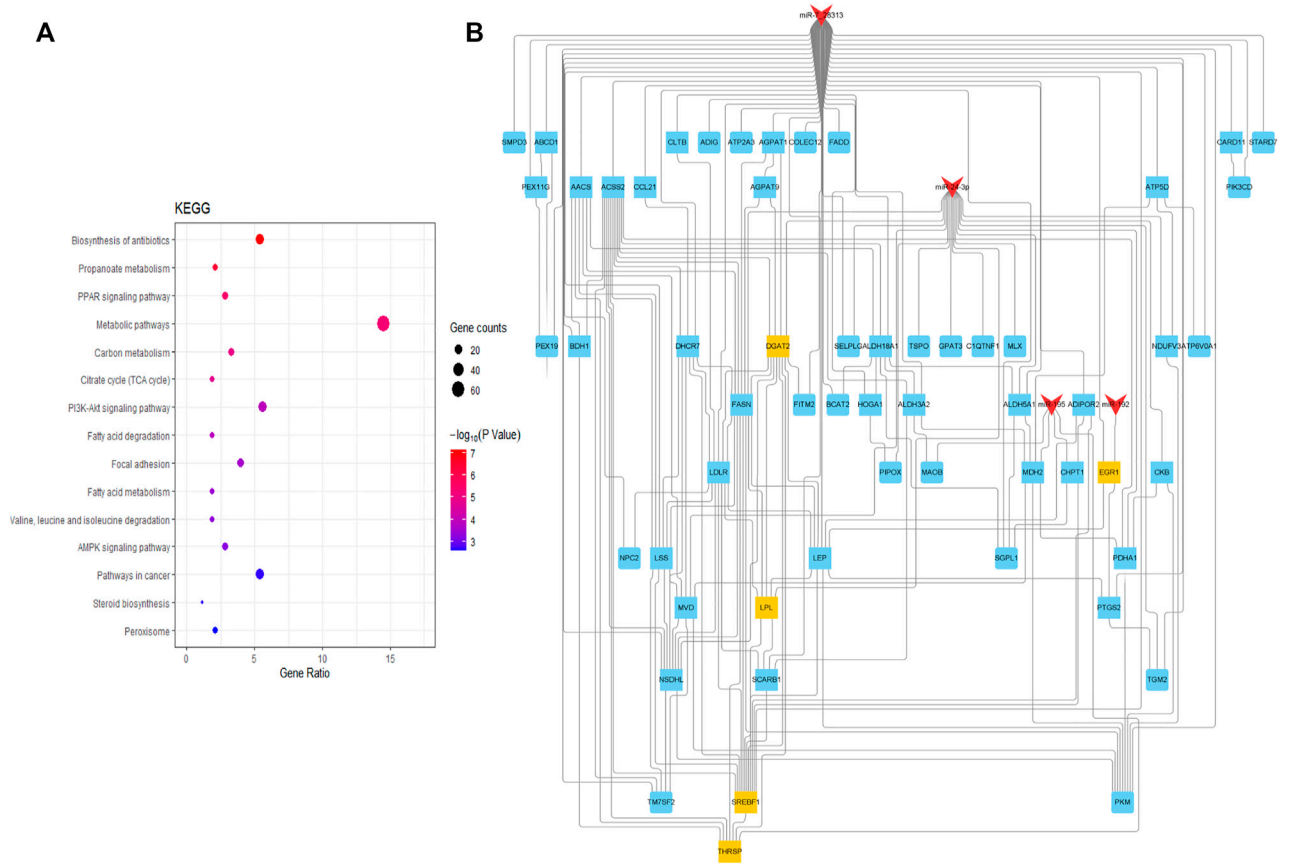


FIGURE 4 | Diagram for KEGG analysis and PPI of the target gene. **(A)** Analysis of signaling pathways of miRNAs target genes. The dot size represents gene amount enriched, and the color represents significance; **(B)** PPI plot of the miRNAs-target gene and the red shape "V" represents miRNAs; the shape "squares" are target genes, yellow and blue squares are key target genes.

gene significance (GS > 0.2) with phenotype and showed high module-membership (MM) > 0.8, which were select as candidates for hub miRNAs.

Enrichment Analysis of the Target Genes of miRNAs

A total of 157 miRNAs related to fat deposition were obtained by intersecting the miRNAs obtained from differential expression analysis and WGCNA analysis. (**Figure 3A**). Prediction of target genes showed these miRNAs could target 6,501 genes. The target genes were enriched in ten GO terms, including 83 biological process terms, 26 cellular component terms, and 35 molecular function terms. Of these, in total, 76 terms were significantly enriched ($p < 0.05$). Further analysis showed that the target genes were significantly enriched in translation, lipid transport, ATP synthesis coupled electron transport and endodermal cell differentiation (**Figures 3B–D**).

KEGG pathway enrichment analysis was performed to comprehensively understand the functions of the target genes at the two growth stages. The target genes of DE miRNAs were significantly enriched in 31 pathways associated with fatty acid ($p < 0.05$), including the fatty acid degradation, fatty acid metabolism, and PPAR signaling pathway (**Supplementary Table S5**) (**Figure 4A**).

TargetScan and RNAHybrid further predicted the target genes, and 98 target genes with extremely significant differential expression were selected for intersection with the above-predicted target genes. Finally, 103 credible candidate target genes were selected (**Figure 4B**). Four candidate target genes, *THRSP*, *EGR1*, *DGAT2*, *SREBF1*, and *LPL*, were screened and determined as hub genes related to fat deposition by interacting network analysis, which could be candidate genes for subsequent verification.

Validation of DE miRNAs Using qRT-PCR

To validate the accurancy of the sequencing data, eight miRNAs (*bta-miR-192*, *bta-miR-195*, *bta-miR-24-3p*, *bta-miR-30a*, *bta-miR-130a*, *bta-miR-218*, *novel-miR-7_28313*, and *novel-miR-26_20059*) were randomly selected from the 282 DE miRNAs to identify their expression pattern at the two growth stages by qRT-PCR. All eight were differentially expressed across the two stages ($p < 0.05$) and the expression pattern was consistent across miRNA sequencing and qPCR (**Figure 5A**), showing that the

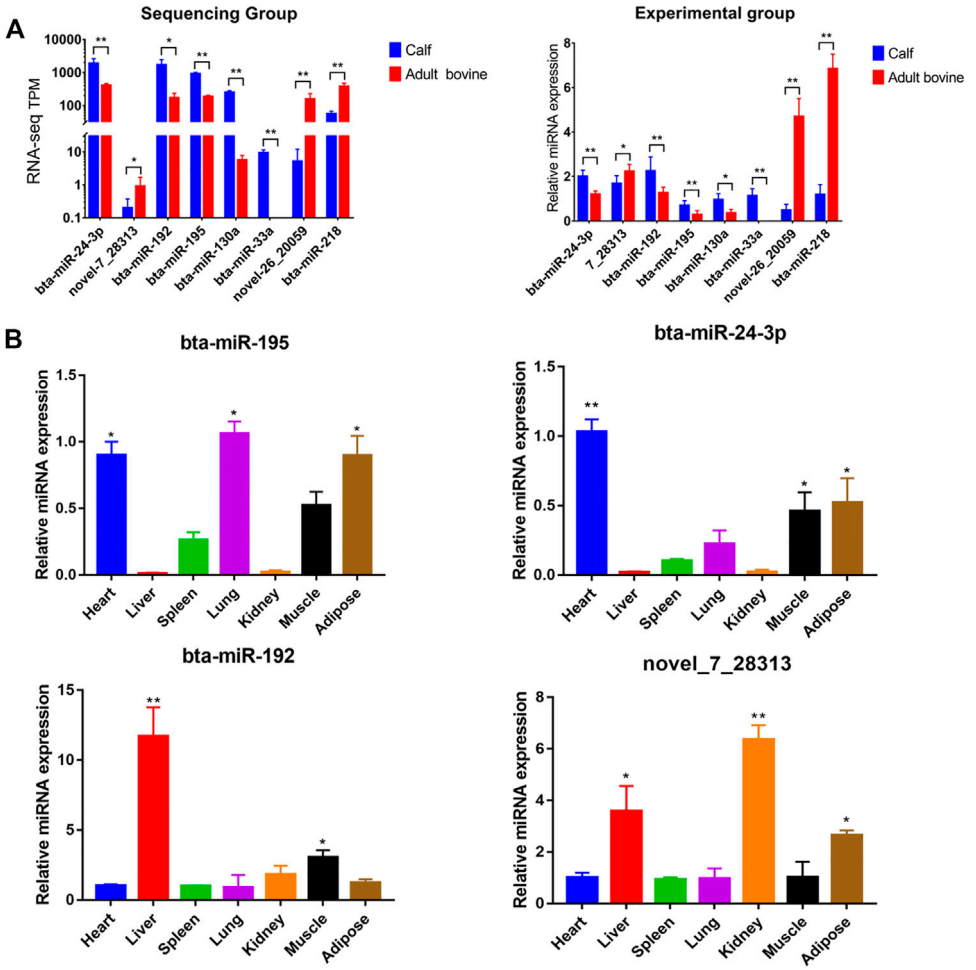


FIGURE 5 | Validation of sequencing accuracy and expression profiles of key miRNAs. **(A)** Expression of miRNAs by RNA-seq **(left)** and qPCR **(right)** at different developmental stages, and the color "blue" and "red" represents calf and adult Xinyang buffalo, respectively; **(B)** Relative expression profile of four interested miRNAs (*bta-miR-195*, *Top left*; *bta-miR-24-3p*, top right; *bta-miR-192*, bottom left, *novel_7_28313*, bottom right) in heart, liver, spleen, lung, kidney, muscle, and adipose tissue (mean ± SD, $n = 3$). The symbol "*", "**", and "****" represent $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

expression values obtained from miRNA sequencing were reliable.

Adipocyte Differentiation Induction and Bta-miR-195 Expression Analysis

Analysis of the tissue expression profile of four miRNAs that could target fat function genes showed that bta-miR-195 was highly expressed in fat (**Figure 5B**). We selected bta-miR-195 for subsequent functional verification on adipocyte differentiation. To explore its expression pattern, primary adipocytes collected from subcutaneous adipose tissue were induced to differentiation. Oil red O staining showed that adipocytes induced for 10 days were stained while preadipocytes were not, indicating that the primary adipocytes were successfully induced to differentiated adipocytes (**Figure 6A**). *PPARγ*, *CEBPα*, and *FABP4*, as marker genes of adipocyte differentiation, were also upregulated, confirming the successful induction of differentiation (**Figure 6B**). Expression values of *bta-miR-195* at days 0, 2, 4, 6, and 10 of adipocyte differentiation were determined by qPCR. The expression of bta-mir-195 was significantly upregulated ($p < 0.05$) on day 4 of induction and maintained high expression during adipocyte differentiation until day 10 (**Figure 6C**), suggesting it may play an important role in the later stages of differentiation.

Effects of Bta-miR-195 on Lipid Accumulation in Adipocytes

Overexpression of bta-mir-195 was induced to further explore its effects on the adipogenic differentiation of bovine adipocytes. The transfection efficiency of *miR-195* was determined by qRT-PCR on days 0, 2, 4, 6, and 10 of induction. Expression of *bta-miR-195* increased significantly ($p < 0.001$) on day 2 of transfection and maintained high expression until day 10, indicating that the agomiR used was effective and stable and that transfection was

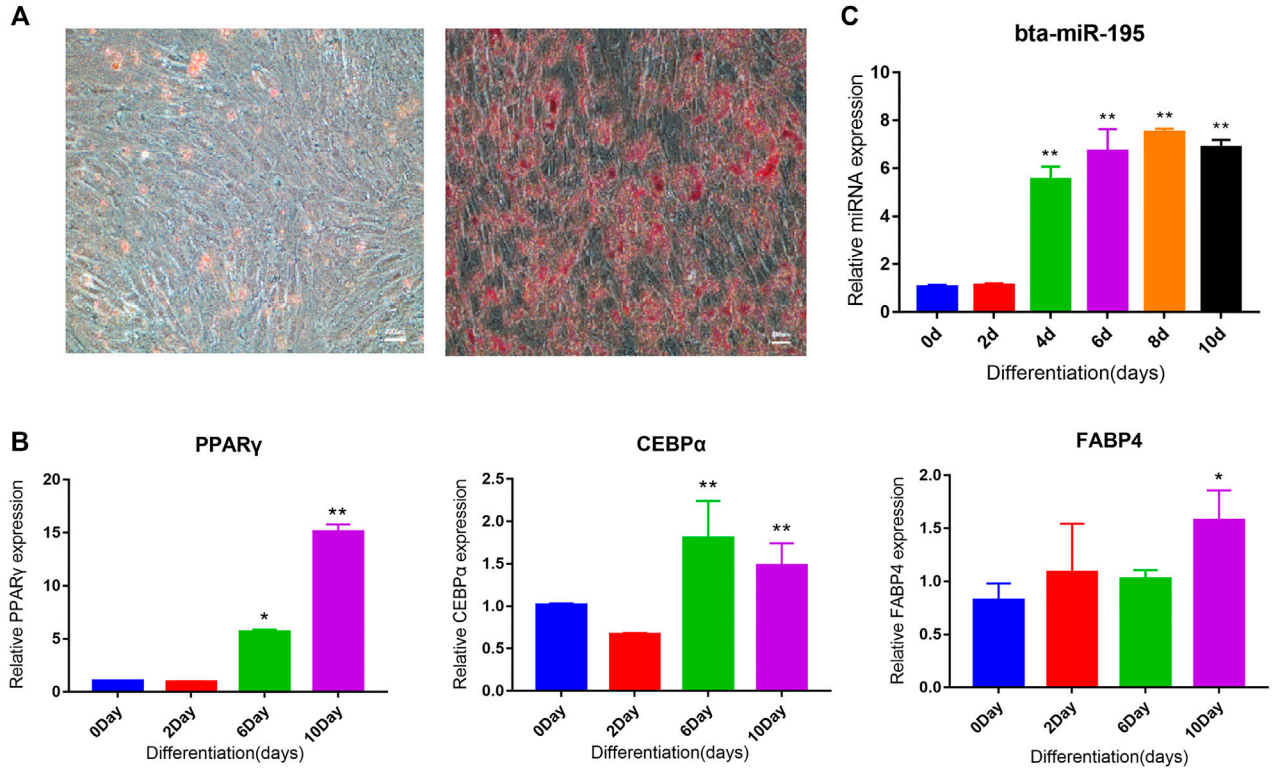


FIGURE 6 | Adipocyte differentiation induction and bta-miR-195 expression analysis. **(A)** Oil red O staining of adipocytes 10 days after induction of bovine primary adipocytes differentiation, with the control group (right, 200X); **(B)** and **(C)** Real-time quantitative detection of *PPARγ*, *C/EBPα*, *FABP4*, and miR-195 markers of adipocyte differentiation. Gene expression plotted as fold-change relative to day 0 (mean ± SD, $n = 3$, the significance symbols are the same as **Figure 5**).

successful (**Figure 7A**). The expression of lipogenic markers during adipocyte differentiation was quantified by qPCR, adipocytes were stained with Oil red O, and lipid accumulation was quantified using OD values after 10 days of induction. It was found that the expression of lipogenic marker genes *PPARγ* and *CEBPα* was significantly inhibited ($p < 0.001$) and the quantity of lipid decreased significantly with overexpression of bta-mir-195 (**Figure 7B**), indicating the negative regulatory effect of *bta-miR-195* on adipocyte differentiation (**Figure 7C**). The expression of *THRSP*, a predicted target gene of *bta-miR-195*, was significantly reduced ($p < 0.001$) after day 2 of transfection (**Figure 7D**), which was contrary to the overexpression of *bta-miR-195*.

Bta-miR-195 Inhibition of Lipid Accumulation by Targeting *THRSP*

It was predicted that *THRSP* was a target gene of *bta-mir-195*. A dual-luciferase reporter assay was carried out to detect the targeting and binding of *bta-mir-195* to *THRSP*. The fluorescence of *bta-mir-195* agromiR was significantly lower than in the NC and mutant groups ($p < 0.001$). Fluorescence in the mutant vector group was lower than in NC, but significantly higher than in the wild vector group ($p < 0.01$). This may be due to excess inserts in the mutant vector affecting the transcriptional activity of *bta-mir-195*. These results indicate that *bta-mir-195* targeted *THRSP* 3′UTR and inhibited the expression of *THRSP* (**Figure 8**).

DISCUSSION

Buffalo breeds in China are plentiful and have high meat products, which can be considered as an essential source of meat (Sun et al., 2020). However, its rough muscle fibers and low IMF abundance have resulted in a meager market share. Fat deposition abundance is one of the major determinants of meat tenderness and juiciness. Still, the dorsal subcutaneous fat above the longest dorsal muscle can reflect the effect of IMF deposition. The measurement of dorsal fat thickness (dorsal subcutaneous fat thickness) is used to predict intramuscular fat deposition *in vivo*, as shown in pig (Monziols et al., 2006; Knecht and Duziński, 2016), cattle (Nogalski et al., 2017), sheep (Chay-Canul et al., 2016), providing reference to study intramuscular fat deposition in buffalo. Fat deposition is regulated by many small molecules, such as lncRNAs and miRNAs. Still, there exists a shortage of studies on related miRNAs and their regulatory mechanisms in the development of buffalo fat deposition. Therefore, the transcriptome sequencing of Xinyang buffalo at different developmental stages was performed in the hope of identifying miRNAs and exploring their molecular mechanisms associated with fat deposition.

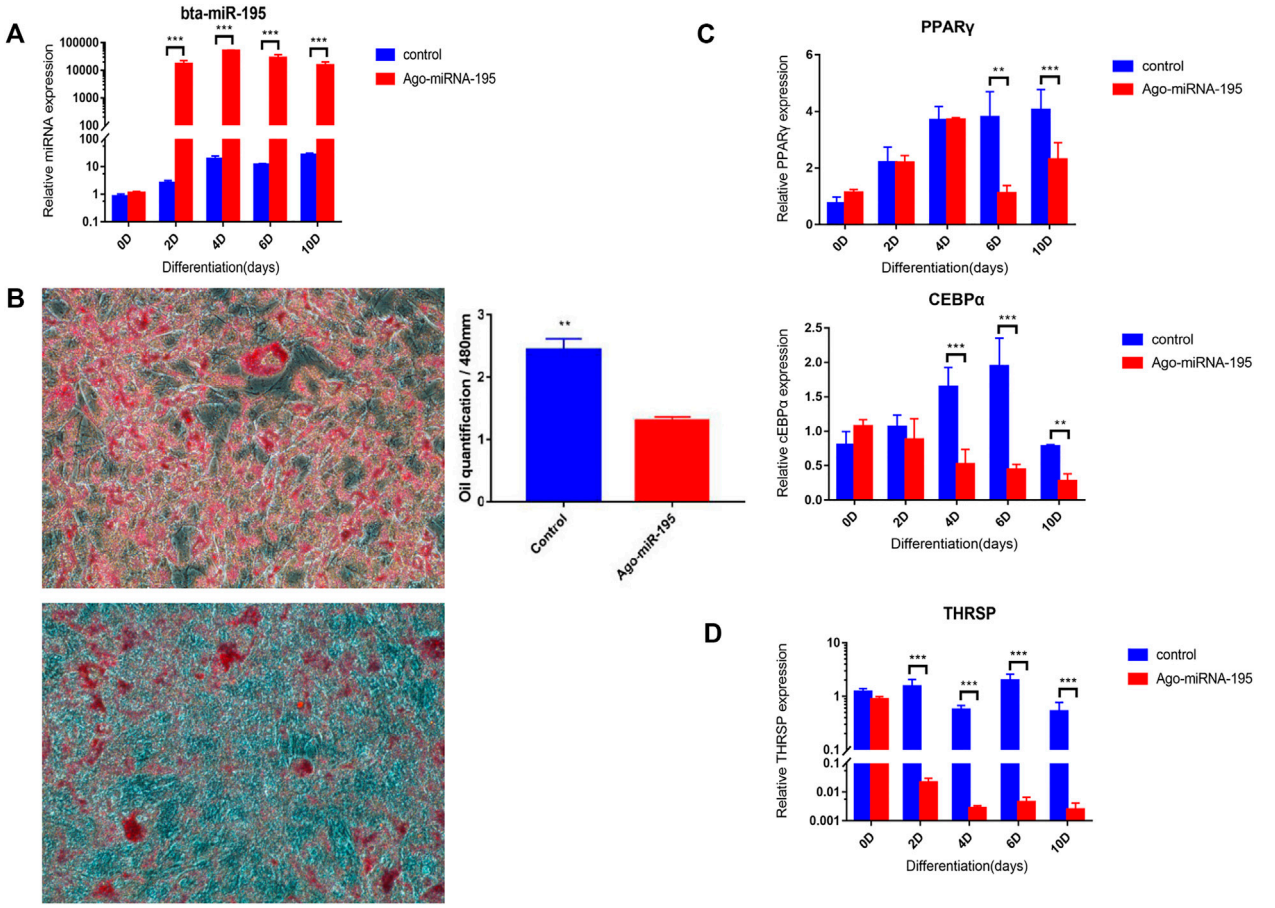


FIGURE 7 | Analysis of the functional role of bta-miR-195 in fat deposits. **(A)** Transfection efficiency of *miR-195* agonist agomiR; **(B)** Quantitative Oil red O staining (200X) after 10 days of high expression of *miR-195*; **(C)** Expression of adipogenic marker genes in primary adipocytes after high expression of *miR-195*; **(D)** Expression of candidate target gene *THRSP* in primary adipocytes after high expression of *miR-195* (mean ± SD, *n* = 3). The significance symbols are the same as **Figure 5**.

As expected, we identified several miRNAs associated with adipogenesis, such as *let-7*, *miR-10b*, *miR-148a*, *miR-27b*, *miR-29b*, and *miR-195*. It has been shown that the *let-7* miRNAs family was highly expressed in adipose tissue and that family members *let-7f*, *let-7e* could target *HMGA2* and *MMP9*, key genes for lipogenesis, and thus affect triglyceride synthesis and lipid metabolism (Sun et al., 2009; Ventayol et al., 2014). The expression of *miR-10b* was negatively correlated with the lipogenic marker genes *CEBPα*, *PPARγ* and *AP2*, and *miR-10b* could play a key regulatory role in hADSC lipogenic differentiation by inhibiting *SMAD2* gene expression and consequently affecting the TGF-β pathway (Li et al., 2018). During adipogenic differentiation of ADSCs, upregulation of *miR-29b* could promote adipogenesis by enhancing *SP1* indirectly targeting *TNF-α* (Li et al., 2018). The up-regulation of these miRNAs may indirectly target fat deposition-related pathways and proteins by affecting the translation of their target genes, which could finally regulate fat deposition and improve buffalo meat quality.

Considering that some important miRNAs (miRNAs involved in fat deposition but not significantly differential expression) may be lost by only significant expression of miRNAs, we further used WGCNA to uncover important miRNAs associated with fat deposition and found that *miR-93*, *miR-199a-3p*, and *miR-195* were closely associated with adipogenesis development. It was shown (Li et al., 2018) that decreased *miR-93* expression increased lipid droplet content, and *miR-93* restricted early adipocyte precursor self-renewal by targeting *Tbx3*. Meanwhile, *miR-93* targeted *Sirt7* to induce precursor adipocyte differentiation and maturation and affect lipid redox. *miR-199a-3p* overexpression attenuated lipid accumulation and adipogenic gene expression and impaired brown adipocyte metabolism by reducing mitochondrial DNA content and respiration (Gao et al., 2018).

In addition, to further elucidate the mechanisms associated with adipogenesis regulation by these miRNAs (*miR-195*, *miR-192*, *miR-7_28313*, *miR-24-3p*), we focused on this on the functional enrichment of target genes of key miRNAs. Excitingly, these target genes were indeed enriched in pathways related to adipogenesis, such as fatty acid metabolism (Mika et al., 2019), PPAR signaling pathway (Fanale et al., 2017). Not only could these miRNAs regulate

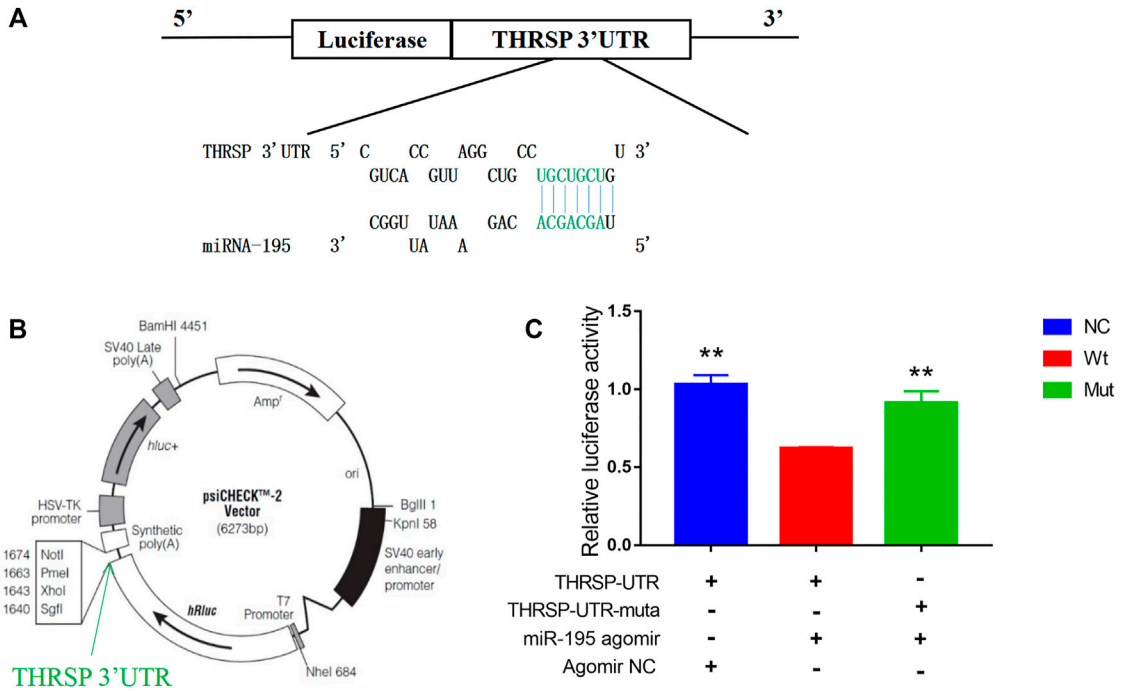


FIGURE 8 | Dual-luciferase reporter assay for *THRSP*-targeting by *bta-miR-195*. **(A)** AgomiR pairing schematic of *miR-195* and *THRSP* 3′UTR. Nucleotides in green represent the "seed agomiR sequence" of *miR-195*, with mutation nucleotides in turquoise. **(B)** psiCHECK-2 agomiR vector map with the insertion site of *THRSP* 3′UTR marked. **(C)** *THRSP*-3′ UTR and its agomiR mutation luciferase reporter vectors co-transfected with *miR-195* agomiR (or negative agomiR control) into 293T cells. Dual-luciferase assay performed 48 h after transfection. Results expressed as relative luciferase activity (mean ± SD, $n = 6$, ANOVA, the significance symbols are the same as **Figure 5**).

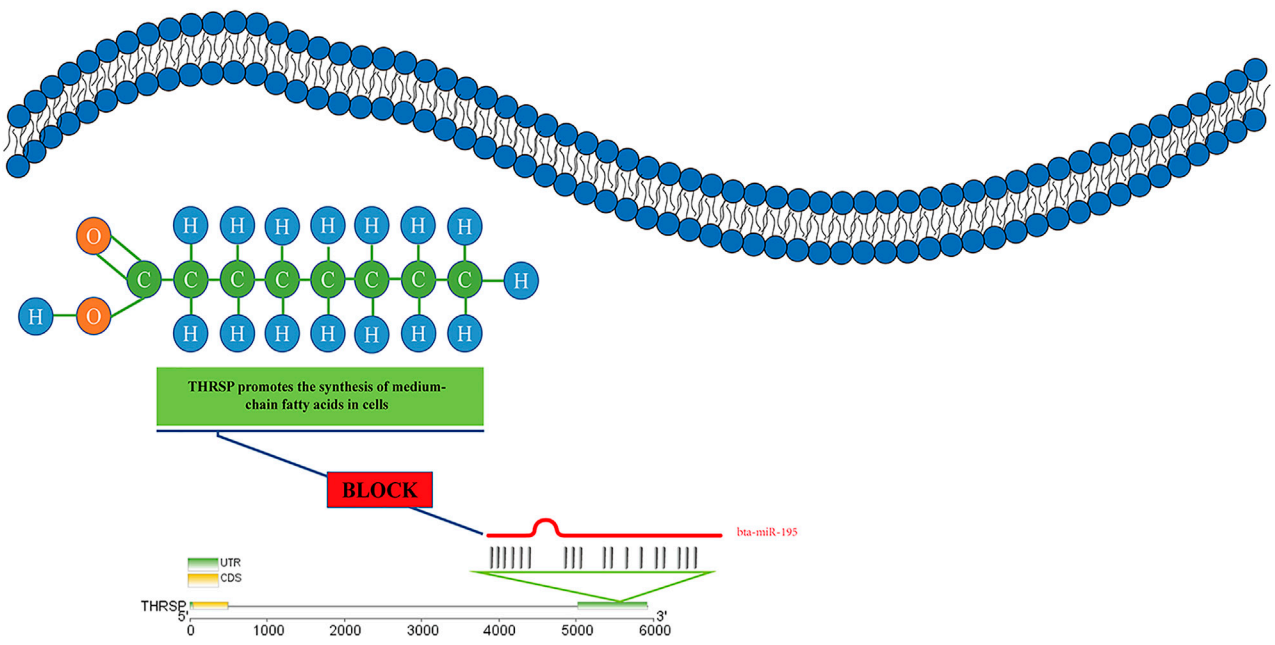


FIGURE 9 | *miR-195* inhibits medium-chain fatty acid synthesis by targeting *THRSP*.

the expression of target genes to participate in the adipogenic process, but these miRNAs may be collectively involved in several comprehensive signaling pathways affecting fat deposition.

In this study, we finally focused *miR-195* according to differential analysis and co-expression network analysis. Expression features showed a high abundance of *miR-195* expression in adipose and a significant dynamic expression of *miR-195* with the maturation of adipocyte differentiation, suggesting a possible regulatory role in adipose differentiation; this was confirmed by its overexpression. The marker of fat deposition is the differentiation of many adipocytes, and this process is the accumulation of lipid droplets after the differentiation of primary adipocytes (Moreno-Navarrete and Fernández-Real, 2017). During the differentiation of primary adipocytes, the cells first change from the shuttle to round shape and begin to differentiate and gradually produce fine lipid droplets, which accumulate and fuse with the time of differentiation and eventually form large lipid droplets that can be stained. The ability of primary adipocytes to be induced into a round shape by inducers is indicated by the upregulation of the lipogenic marker gene. The eventual inability to form large lipid droplets under the sustained effect of overexpressed *miR-195* suggests that it can inhibit adipocyte differentiation and thus affect fat deposition. It has been reported that the ectopic expression of *miR-195* suppresses the expression of *INSR*, thereby impairing the insulin signaling cascade and glycogen synthesis in HepG2 cells (Yang et al., 2014). Meanwhile, *miR-195* is able to target *FNAS* to regulate ATP consumption and cell migration (Mao et al., 2012; Lin et al., 2021). This suggests that the same miRNA will serve different functions depending on the target genes. In our study, *miR-195* affected lipid droplet formation by targeting *THRSP*. Although there was a temporary differentiation in the pre-lipogenesis-induced differentiation phase, *THRSP* was significantly suppressed all the time with the continuous overexpression of miRNAs, thus inhibiting lipid droplet formation.

THRSP is a small nuclear protein acting as a key lipogenic activator and is activated by thyroid hormone triiodothyronine (T3), carbohydrates, glucose, and insulin (Wu et al., 2012). *THRSP* is abundant in lipogenic tissues such as fat, liver, and mammary gland (Chou et al., 2007; Ren et al., 2017). Many studies show that *THRSP* plays a vital role in regulating lipogenesis in goat mammary epithelial cells by increasing the concentrations of C12:0, C14:0 and synthesizing medium-chain fatty acids *in vitro* (Yao et al., 2016). In mice, *THRSP* acts in the regulation of diet-induced obesity, while *THRSP*-knockout leads to deficiencies in *de novo* lipogenesis in the lactating mammary gland (Yao et al., 2016). T*HRSP* is a potential molecular marker for fat deposition in cattle since its mRNA abundance was shown to be elevated in skeletal muscle with high IMF concentration (Schering et al., 2017). These results suggest that *miR-195* may inhibit the synthesis of medium-chain fatty acids by inhibiting *THRSP*, thus leading to the failure of lipid accumulation (**Figure 9**). In addition, the comparison showed that the *THRSP* 3′UTR binding sites were highly consistent, indicating that *miR-195* had the same function in cattle and buffalo (**Supplementary Figure S3**). Therefore, *bta-miR-195* plays a significant role in inhibiting lipogenesis by suppressing the expression of *THRSP*, and knockdown of *bta-miR-195* may positively affect lipid deposition.

While this study validated miRNAs associated with fat deposition from different perspectives, it still leaves something to be desired. We were unable to isolate any primary adipocytes from the longest muscle of the buffalo dorsum due to poor fat deposition in Xinyang buffalo (Huang et al., 2020), so we adopted samples obtained from the subcutaneous fat of the dorsum for small RNA sequencing. Therefore, our experiments were just able to reflect the results of intramuscular fat deposition in cattle from the side. Second, we should use public database resources to validate the expression of candidate miRNAs in other bovine species to confirm the accuracy of our results.

Fat deposition is regulated by some miRNAs, which inhibit adipocyte differentiation and hinder adipocyte formation by regulating their target genes, resulting in less intramuscular fat deposition and inferior meat quality in Xinyang buffalo, but this result needs more functional experiments to prove. This study used gene sequencing technology to systematically and comprehensively identify miRNAs that regulate adipogenesis in buffalo and explore their regulatory mechanisms, providing a theoretical basis for enriching muscle fat deposition in buffalo. In addition, a miRNA (*miR-195*), which may be related to adipose function, was screened, and its mechanism of action in adipocytes was elucidated, which laterally confirmed the function of the screened miRNAs in the adipose deposition. This study enriches the theory of fat deposition in buffalo and provides a theoretical basis for the selection and breeding of buffalo for meat.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: GSE171146, GSE119064, GSE112744.

ETHICS STATEMENT

The animal study was reviewed and approved by The Animal Ethics Committees of Ningxia University.

AUTHOR CONTRIBUTIONS

SW and CP made the same contribution to the paper. Conceived and designed the research: YM and SW; Analyzed the data and conducted the experiment; Modified article: CY, YM, XW, HL, and LT. All authors read and approved the final article.

FUNDING

This study was funded by the National Natural Science Foundation of China (No. 32072720), the Leading Talents

Fund in Science and Technology Innovation in Henan Province (No. 194200510022), Cultivation of leading talents in science and technology innovation in Ningxia Autonomous Region (No. 2020GKLRLX02), Ningxia Autonomous Region Key Laboratory Infrastructure Construction Project (No. 2019ZDJS001), the Key Research and Talent Introduction Project of Ningxia Hui Autonomous Region (No. 2019BEB04006), and the Science and Technology Innovation Team Projects of Ningxia Hui Autonomous Region (No. 03010360052). The funding bodies played no role in the design of the study, collection, analysis, and interpretation of data and writing the manuscript.

ACKNOWLEDGMENTS

We would like to thank all the participants in our study. I thank Mr. Liu Yuxuan for his linguistic assistance during the preparation of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.736441/full#supplementary-material

REFERENCES

Andrés-León, E., Núñez-Torres, R., and Rojas, A. M. (2016). MiARma-Seq: A Comprehensive Tool for miRNA, mRNA and circRNA Analysis. *Sci. Rep.* 6 (1), 25749–25758. doi:10.1038/srep25749

Barker, J. S. F. (2014). Water Buffalo: Domestication. *L'Anthropologie* 94, 619–642. doi:10.1007/978-1-4419-0465-2_2217

Chay-Canul, A. J., Garcia-Herrera, R., Meza-Villalvazo, V. M., Gomez-Vazquez, A., Cruz-Hernandez, A., Magaña-Monforte, J. G., et al. (2016). Body Fat Reserves and Their Relationship to Ultrasound Back Fat Measurements in Pelibuey Ewes. *Ecosistemas y recursos agropecuarios* 3 (9), 407–413.

Chou, W.-Y., Cheng, Y.-S., Ho, C.-L., Liu, S.-T., Liu, P.-Y., Kuo, C.-C., et al. (2007). Human Spot 14 Protein Interacts Physically and Functionally with the Thyroid Receptor. *Biochem. Biophysical Res. Commun.* 357 (1), 133–138. doi:10.1016/j.bbrc.2007.03.103

Du, M., Huang, Y., Das, A. K., Yang, Q., Duarte, M. S., Dodson, M. V., et al. (2013). Meat Science and Muscle Biology Symposium: Manipulating Mesenchymal Progenitor Cell Differentiation to Optimize Performance and Carcass Value of Beef Cattle1,2. *J. Anim. Sci.* 91 (3), 1419–1427. doi:10.2527/jas.2012-5670

Fanale, D., Amodeo, V., and Caruso, S. (2017). The Interplay between Metabolism, PPAR Signaling Pathway, and Cancer. *PPAR Res.* 2017, 1–2. doi:10.1155/2017/1830626

Funston, R. N., Larson, D. M., and Vonnahme, K. A. (2010). Effects of Maternal Nutrition on Conceptus Growth and Offspring Performance: Implications for Beef Cattle Production1. *J. Anim. Sci.* 88 (13 Suppl. l), E205–E215. doi:10.2527/jas.2009-2351

Gao, Y., Cao, Y., Cui, X., Wang, X., Zhou, Y., Huang, F., et al. (2018). MiR-199a-3p Regulates Brown Adipocyte Differentiation through mTOR Signaling Pathway. *Mol. Cell Endocrinol.* 476, 155–164. doi:10.1016/j.mce.2018.05.005

Güller, I., McNaughton, S., Crowley, T., Gilsanz, V., Kajimura, S., Watt, M., et al. (2015). Comparative Analysis of microRNA Expression in Mouse and Human Brown Adipose tissueResearch Support, Non-U.S. Gov't]. *[Journal ArticleBMC Genomics* 16, 820. doi:10.1186/s12864-015-2045-8

Huang, J., Feng, X., Zhu, R., Guo, D., Wei, Y., Cao, X., et al. (2020). Comparative Transcriptome Analysis Reveals that PCK1 Is a Potential Gene Affecting IMF Deposition in buffalo. *BMC Genomics* 21 (1). doi:10.1186/s12864-020-07120-w

Huang, J., Wang, S., Feng, X., Liu, X., Zhao, J., Zheng, Q., et al. (2019). miRNA Transcriptome Comparison between Muscle and Adipose Tissues Indicates Potential miRNAs Associated with Intramuscular Fat in Chinese Swamp buffalo. *Genome* 62 (11), 729–738. doi:10.1139/gen-2018-0178

Keller, P., Gburcik, V., Petrovic, N., Gallagher, I. J., Nedergaard, J., Cannon, B., et al. (2011). Gene-chip Studies of Adipogenesis-Regulated microRNAs in Mouse Primary Adipocytes and Human Obesity. *BMC Endocr. Disord.* 11, 7. doi:10.1186/1472-6823-11-7

Kim, Y. J., Hwang, S. J., Bae, Y. C., and Jung, J. S. (2009). MiR-21 Regulates Adipogenic Differentiation through the Modulation of TGF-β Signaling in Mesenchymal Stem Cells Derived from Human Adipose Tissue. *Stem Cells* 27 (12), 3093–3102. doi:10.1002/stem.235

Knecht, D., and Duziński, K. (2016). The Effect of Sex, Carcass Mass, Back Fat Thickness and Lean Meat Content on Pork Ham and Loin Characteristics. *Arch. Anim. Breed.* 59 (1), 51–57. doi:10.5194/aab-59-51-2016

Li, H., Fan, J., Fan, L., Li, T., Yang, Y., Xu, H., et al. (2018). MiRNA-10b Reciprocally Stimulates Osteogenesis and Inhibits Adipogenesis Partly through the TGF-B/smad2 Signaling Pathway. *Aging Dis.* 9 (6), 1058–1073. doi:10.14336/AD.2018.0214

Lin, J., Maimaitiyiming, A., Chen, S., Xiao, M., and Xian, Z. (2021). Hypoxia-induced miR-27 and miR-195 Regulate ATP Consumption, Viability, and Metabolism of Rat Cardiomyocytes by Targeting PPARγ and FASN Expression. *Aging* 13 (7), 10158–10174. doi:10.18632/aging.202778

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 550–621. doi:10.1186/s13059-014-0550-8

Mao, J. H., Zhou, R. P., Peng, A. F., Liu, Z. L., Huang, S. H., Long, X. H., et al. (2012). MicroRNA-195 Suppresses Osteosarcoma Cell Invasion and Migration *In Vitro* by Targeting FASN. *Oncol. Lett.* 4 (5), 1125–1129. doi:10.3892/ol.2012.863

Mei, Y., Bian, C., Li, J., Du, Z., Zhou, H., Yang, Z., et al. (2013). miR-21 Modulates the ERK-MAPK Signaling Pathway by Regulating SPRY2 Expression during Human Mesenchymal Stem Cell Differentiation. *J. Cel. Biochem.* 114 (6), 1374–1384. doi:10.1002/jcb.24479

Mika, A., Macaluso, F., Barone, R., Di Felice, V., and Sledzinski, T. (2019). Effect of Exercise on Fatty Acid Metabolism and Adipokine Secretion in Adipose Tissue. *Front. Physiol.* 10, 26. doi:10.3389/fphys.2019.00026

Monziols, M., Collewet, G., Bonneau, M., Mariette, F., Davenel, A., and Kouba, M. (2006). Quantification of Muscle, Subcutaneous Fat and Intermuscular Fat in Pig Carcasses and Cuts by Magnetic Resonance Imaging. *Meat Sci.* 72 (1), 146–154. doi:10.1016/j.meatsci.2005.06.018

Moreno-Navarrete, J. M., and Fernández-Real, J. M. (2017). "Adipocyte Differentiation Adipose Tissue Biology," in *Adipose Tissue Biology*. Editor M. Symonds(Cham: Springer), 69–90. doi:10.1007/978-3-319-52031-5_3

Najafi-Shoushtari, S. H., Kristo, F., Li, Y., Shioda, T., Cohen, D. E., Gerszten, R. E., et al. (2010). MicroRNA-33 and the SREBP Host Genes Cooperate to Control Cholesterol homeostasis[Journal Article; Research Support. *Science* 328 (5985), 1566–1569. doi:10.1126/science.1189123

Nogalski, Z., Pogorzelska-Przybyłek, P., Białobrzewski, I., Modzelewska-Kapituła, M., Sobczuk-Szul, M., and Purwin, C. (2017). Estimation of the Intramuscular Fat Content of M. Longissimus Thoracis in Crossbred Beef Cattle Based on Live Animal Measurements. *Meat Sci.* 125, 121–127. doi:10.1016/j.meatsci.2016.11.026

Ocłoń, E., Latacz, A., Zubel–Łojek, J., and Pierzchała–Koziec, K. (2016). Hyperglycemia-induced Changes in miRNA Expression Patterns in Epicardial Adipose Tissue of pigletsResearch Support, Non-U.S. Gov't]. *[Journal Articlej. Endocrinol.* 229 (3), 259–266. doi:10.1530/JOE-15-0495

Price, N. L., Holtrup, B., Kwei, S. L., Wabitsch, M., Rodeheffer, M., and Bianchini, L. (2016). SREBP-1c/MicroRNA 33b Genomic Loci Control Adipocyte Differentiation[Journal Article; Research Support. *Mol. Cel. Biol.* 36 (7), 1180–1193. doi:10.1128/MCB.00745-15

Remsburg, C., Konrad, K., Sampilo, N. F., and Song, J. L. (2019). Analysis of microRNA Functions. *Methods Cel Biol* 151, 323–334. doi:10.1016/bs.mcb.2018.10.005

Ren, J., Xu, N., Zheng, H., Tian, W., Li, H., Li, Z., et al. (2017). Expression of Thyroid Hormone Responsive SPOT 14 Gene Is Regulated by Estrogen in Chicken (Gallus gallus). [Journal Article; Research Support, Non-U.S. Gov't]. *Sci. Rep.* 7 (1), 10243. doi:10.1038/s41598-017-08452-6

Schering, L., Albrecht, E., Komolka, K., Kühn, C., and Maak, S. (2017). Increased Expression of Thyroid Hormone Responsive Protein (THRSP) Is the Result but

Not the Cause of Higher Intramuscular Fat Content in Cattle. *Int. J. Biol. Sci.* 13 (5), 532–544. doi:10.7150/ijbs.18775

Sun, J., Zhang, B., Lan, X., Zhang, C., Lei, C., and Chen, H. (2014). Comparative Transcriptome Analysis Reveals Significant Differences in MicroRNA Expression and Their Target Genes between Adipose and Muscular Tissues in Cattle. [Comparative Study; Journal Article; Research Support, Non-U.S. Gov't]. *PLoS One* 9 (7), e102142. doi:10.1371/journal.pone.0102142

Sun, T., Fu, M., Bookout, A. L., Kliewer, S. A., and Mangelsdorf, D. J. (2009). MicroRNA Let-7 Regulates 3T3-L1 adipogenesis[Journal Article; Research Support. *Mol. Endocrinol.* 23 (6), 925–931. doi:10.1210/me.2008-0298

Sun, T., Wang, S., Chanthakhoun, V., Dang, R., Huang, Y., Chen, H., et al. (2020). Multiple Domestication of Swamp buffalo in China and South East Asia. [Journal Article]. *J. Anim. Breed. Genet.* 137 (3), 331–340. doi:10.1111/jbg.12445

Ventayol, M., Vinas, J. L., Sola, A., Jung, M., Brune, B., Pi, F., et al. (2014). MiRNA let-7e targeting MMP9 is involved in adipose-derived stem cell differentiation toward epithelia. *Cell Death Dis* 5, e1048. doi:10.1038/cddis.2014.2

Wei, J., Li, H., Wang, S., Li, T., Fan, J., Liang, X., et al. (2014). Let-7 Enhances Osteogenesis and Bone Formation while Repressing Adipogenesis of Human Stromal/mesenchymal Stem Cells by Regulating HMGA2. [Journal Article; Research Support, Non-U.S. Gov't]. *Stem Cell Dev* 23 (13), 1452–1463. doi:10.1089/scd.2013.0600

Wu, J., Li, J., and Men, X. L. (2012). The Role of Thrsp in the Regulation of Lipogenesis. *Sheng Li Ke Xue Jin Zhan* 43 (5), 393–397. doi:10.1002/hep.26272

Yamada, T., Kamiya, M., and Higuchi, M. (2020). Fat Depot-specific Effects of Body Fat Distribution and Adipocyte Size on Intramuscular Fat Accumulation in Wagyu Cattle. *Anim. Sci. J.* 91 (1), e13449. doi:10.1111/asj.13449

Yang, W. M., Jeong, H. J., Park, S. Y., and Lee, W. (2014). Saturated Fatty Acid-Induced miR-195 Impairs Insulin Signaling and Glycogen Metabolism in HepG2 Cells. *FEBS Lett.* 588 (21), 3939–3946. doi:10.1016/j.febslet.2014.09.006

Yao, D. W., Luo, J., He, Q. Y., Wu, M., Shi, H. B., Wang, H., et al. (2016). Thyroid Hormone Responsive (THRSP) Promotes the Synthesis of Medium-Chain Fatty Acids in Goat Mammary Epithelial cellsResearch Support, Non-U.S. Gov't]. *[Journal Articlej. Dairy Sci.* 99 (4), 3124–3133. doi:10.3168/jds.2015-10632

Yao, Q., Song, Z., Wang, B., Qin, Q., and Zhang, J. (2019). Identifying Key Genes and Functionally Enriched Pathways in Sjögren's Syndrome by Weighted Gene Co-expression Network Analysis. *Front. Genet.* 10, 1142. doi:10.3389/fgene.2019.01142

Zhang, Y., Colli, L., and Barker, J. (2020). Asian Water buffalo: Domestication, History and Genetics. [Journal Article; Review]. *Anim. Genet.* 51 (2), 177–191. doi:10.1111/age.12911

Zuo, C., and Liu, T. (2008). Reasons and Solutions to Reduction in Quantity of Xinyang Water-buffalo. *J. Xinyang Agric. Coll.* 1, 130–131. doi:10.16593/j.cnki.41-1433/s.2008.01.053

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Wang, Pan, Ma, Yang, Tang, Huang, Wei, Li and Ma. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 January 2022
doi: 10.3389/fgene.2021.807234

Episodes of Rapid Recovery of the Functional Activity of the *ras85D* Gene in the Evolutionary History of Phylogenetically Distant *Drosophila* Species

*A. I. Chekunova[1], S. Yu. Sorokina[1], E. A. Sivoplyas[2], G. N. Bakhtoyarov[3], P. A. Proshakov[1], A. V. Fokin[1], A. I. Melnikov[1] and A. M. Kulikov[1]**

[1]Evolutionary Genetics of Development, N.K. Koltzov Institute of Developmental Biology of the Russian Academy of Sciences, Moscow, Russia, [2]Department of Biochemistry, Molecular Biology and Genetics, Institute of Biology and Chemistry of Moscow Pedagogical State University (MPGU), Moscow, Russia, [3]Laboratory of Genetics of DNA Containing Viruses, Federal State Budgetary Scientific Institution «I. Mechnikov Research Institute of Vaccines and Sera», Moscow, Russia

As assemblies of genomes of new species with varying degrees of relationship appear, it becomes obvious that structural rearrangements of the genome, such as inversions, translocations, and transposon movements, are an essential and often the main source of evolutionary variation. In this regard, the following questions arise. How conserved are the regulatory regions of genes? Do they have a common evolutionary origin? And how and at what rate is the functional activity of genes restored during structural changes in the promoter region? In this article, we analyze the evolutionary history of the formation of the regulatory region of the *ras85D* gene in different lineages of the genus *Drosophila*, as well as the participation of mobile elements in structural rearrangements and in the replacement of specific areas of the promoter region with those of independent evolutionary origin. In the process, we substantiate hypotheses about the selection of promoter elements from a number of frequently repeated motifs with different degrees of degeneracy in the ancestral sequence, as well as about the restoration of the minimum required set of regulatory sequences using a conversion mechanism or similar.

Keywords: *Drosophila*, evolutionary conservative motifs, *ras85D*, transcription start site, broad peaked promoter, conversion, core promoter elements, non-coding regions

OPEN ACCESS

Edited by:
William C. Cho,
QEH, Hong Kong SAR, China

Reviewed by:
Claudia Carareto,
São Paulo State University, Brazil
Carl Hjelmen,
Utah Valley University, United States

****Correspondence:***
A. M. Kulikov
amkulikov@gmail.com

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: *01 November 2021*
Accepted: *21 December 2021*
Published: *12 January 2022*

Citation:
Chekunova AI, Sorokina SY, Sivoplyas EA, Bakhtoyarov GN, Proshakov PA, Fokin AV, Melnikov AI and Kulikov AM (2022) Episodes of Rapid Recovery of the Functional Activity of the ras85D *Gene in the Evolutionary History of Phylogenetically Distant* Drosophila *Species. Front. Genet. 12:807234. doi: 10.3389/fgene.2021.807234*

INTRODUCTION

Bifurcation points on the phylogenetic tree of species often correspond to critical periods in the evolutionarily history of a population, resulting from abrupt changes in the environment and accompanied by physiological and genomic stress. At the genetic level, these events are associated with a disruption of genome stability and a sharp increase in the mutation rate (Galhardo et al., 2007; Malkova and Haber, 2012). The molecular mechanisms of mutagenesis caused by genomic instability are diverse: a weakening of the error repair by DNA polymerases and changes in the contribution of different polymerases to DNA replication, changes in the conversion mechanisms for the repair of deletions and double-strand breaks, an increase in the transposition activity of mobile elements, associated changes in the expression activity of the target genes of such transpositions and the

formation of chimeric sequences, as well as large-scale rearrangements of the genome (Wong and Choo, 2004; Feschotte and Pritham, 2007; Garcia Guerreiro, 2012).

The transpositions of mobile elements can be accompanied by inversion events. The direct participation of mobile elements in the formation of rearrangements has been shown in *Drosophila* both in laboratory experiments (Alonso-Gonzalez et al., 2006; Kovalenko et al., 2006) and in evolutionary studies (Zelentsova et al., 1999; Evgen'ev et al., 2000). Such rearrangements lead to sequence replacements in the border region and to significant changes in regulatory sequences of downstream genes. The genome is literally saturated with the consequences of such rearrangements, which is manifested in a sharp decrease in the degree of gene colocalization on the synteny plots of chromosomes or Muller elements, as the phylogenetic distance between the compared species increases (Drosophila 12 Genomes Consortium and Clark, 2007). The consequences of rearrangements are manifested as a significant increase in linkage disequilibrium in their inner regions (Wallace et al., 2013) and in the level of polymorphism at the edges (Rozas et al., 1999; Sanchez-Gracia and Rozas, 2011). The authors associate the increase in the mutation rate in the latter case with the action of directed selection. Of interest are the following questions: the fixation rate of such consequences of genomic instability, the possible degree of disruption in the functional activity of genes bordering on insertions and rearrangements, and the influence of these events on further evolution of the local sequence.

Selection on coding sequences is in most cases associated with the maintenance or change of the functional activity of the protein encoded. The level and specificity of expression activity are associated with the variation of regulatory sites in non-coding sequences. For species with a distant relationship, a significant variation of key regions of the gene regulatory sequence—transcription start sites (TSSs)—has been shown. Using a large sample of loci from *Drosophila* species with different degrees of relationship, Main and co-authors have shown a high similarity of TSS sequences in orthologs for species of the *melanogaster* group and their significant differences from *D. pseudoobscura* (Main et al., 2013). The evolutionary shift of TSSs can supposedly be explained by an increased mutation rate and the fixation of mutations in these regions and is therefore a consequence of the effect of positive selection. Considering the significant number of chromosome rearrangements when comparing *D. pseudoobscura* and *D. melanogaster*, and the even larger number of degenerate remnants of mobile elements, it can be argued that at least some of these inconsistencies are associated with the emergence of sequences of independent evolutionary origin. The following question remains open: how and to what extent does the restoration of the expression activity of such genes occur?

The region of the *ras85D* gene (*Drosophila Ras1*) in *Drosophila* species with different degrees of relationship seems to be an interesting model for testing the hypothesis about the role of mobile elements and rearrangements in the accumulation of mutations in non-coding sequences. The *Ras* genes are a family of evolutionarily conserved genes encoding proteins of the group of small G-proteins (small GTPases) that play the role of "molecular switches" (Valencia et al., 1991; Rommel and Hafen, 1998; Rojas et al., 2012). The vast majority of signaling cascades dependent on *Ras* proteins are conserved (Slack et al., 2015) and typical for all eukaryotes. The coding sequence of the *ras85D* gene is under the influence of strict stabilizing selection (Chekunova et al., 2008) and is characterized by neutral polymorphism and low population variation (Gasperini and Gibson, 1999). At the same time, analysis of the nature of interspecific mutation accumulation in a fragment of this locus in closely related *Drosophila* species of the *virilis* group showed a significant impairment in the operation of the molecular clock (Kulikov et al., 2010). Rearrangements and insertions of mobile elements occurring in functionally significant regions of the genome must inevitably lead to dramatic changes in the activity and regulation of these regions. In the case of the *ras85D* gene, one of the edges of the large inversion is located on the pre-promoter region. Disruption of the expression activity of the *ras85D* gene, as one of the most important genes of the cell cycle regulation system, should have a significant effect on the viability of the cell and the organism as a whole. Significant structural changes in the promoter region of this gene are associated with the formation of lethal mutations and pronounced severe morphological disorders abnormalities. Thus, the insertion of mobile elements and regulatory fusion constructs into the intergenic region upstream of the *ras85D* transcription start site (TSS) or into the region of the non-coding 5′UTR sequence leads to the occurrence of lethal and deleterious mutations, female sterility, and impairments in the structure of neuromuscular junctions (Schnorr and Berg, 1996; Spradling et al., 1999; Prober and Edgar, 2000; Koh et al., 2002; Wu et al., 2009).

The objectives of this study are to test the hypothesis of the evolutionarily independent origin of the promoter region in *Drosophila* species with different degrees of relationship, to confirm the rather obvious assumption about the participation of mobile elements in the formation of different variants of this sequence, and to reveal the most probable mechanisms for restoring the functional activity of the gene. In the latter case, the question is divided into two parts as follows. 1) How is the new promoter selected if a part of the sequence including the promoter is replaced? 2) How is the minimum necessary set of enhancers selected and restored for an adequate expression regulation?

MATERIALS AND METHODS

Research Objects

The *ras85D* gene with upstream intergenic region of 29 *Drosophila* species, including *Drosophila erecta, D. yakuba, D. sechellia, D. simulans, D. melanogaster, D. rhopaloa, D. ficusphila, D. biarmipes, D. takahashii, D. elegans, D. suzukii, D. eugracilis, D. obscura, D. pseudoobscura, D. persimilis, D. ananassae, D. bipectinata, D. kikkawai, D. serrata, D. willistoni, D. albomicans, D. hydei, D. grimshawi, D. navojoa, D. mojavensis, D. virilis, D. americana, Dorsilopha busckii, Scaptodrosophila lebanonensis*

(the Genomes—NCBI Datasets database), and nine species of the *virilis* group, including *D. lacicola, D. montana, D. borealis, D. flavomontana, D. kanekoi, D. ezoana, D. littoralis, D. lummei, D. novamexicana* (Yusuf et al., in prep.) were used for the analysis. The boundaries of the fragment for the species with annotated genomes were revealed according to the GDV genome browser data. The homologous sequences in unannotated genomes were found using BLASTn or tblastn with the *D. melanogaster ras85D, Rlb1* and *CG31344* coding or amino acid sequence. The boundaries of non-coding sequences were revealed by the SRA-BLAST results according to the Sequence Read Archive Database NCBI (if any), or by alignment with the sequences of most related species. To find sequences in *D. virilis* genome that are homologous to inverted repeats, specific for the *ras85D* upstream intergenic region of the *virilis* group of *Drosophila*, we used Nucleotide BLAST with fragment size of at least 75 bp.

Sequence Alignment and Extraction of Evolutionarily Conserved Motifs

The sequence alignment was carried out in MEGA X software (Kumar et al., 2018) using the ClustalW method and manually for the inner area of the intergenic region of the *virilis* species group enriched with multiple deletions. Because of different evolutionary origin of noncoding sequences (intergenic region and 5′UTR of *ras85D*) of different *Drosophila* species groups, the alignment was carried out only for closely related species and for fragments with significant homology. Evolutionarily conserved motifs (ECMs) were used as markers of homology of noncoding regions. ECMs were identified in the sequences of all species studied using the MEME analytical platform version 5.4.1—the MEME (Bailey and Elkan, 1994; Bailey et al., 2009) and the MAST (Bailey et al., 2009) tools. ECMs were searched independently for two sets of sequences: Set A includes 23 species from different groups, excluding closely related species that demonstrate high homology of noncoding regions; set B includes all 37 species. When the total amount of ECMs per sequence was 14, the number of common ECMs for both sets of sequences was maximal.

The number of homologous motifs in each sequence was set as any. In order to normalize for biased distribution of nucleotides, the background model was used by default. The homologous regions within the intergenic sequence of each species, such as forward and inverted repeats, were revealed using the YASS tool (Noe and Kucherov, 2005). The parameters of sequence fragment comparison were used by default.

Revealing of the Sequence Homology With Mobile Elements

The search for regions of homology with mobile elements was carried out using the GIRI Repbase library and the Censor tool (Kohany et al., 2006) with the Hexapoda filter (Insects). The Censor tool does not allow to decrease the Score values and to change the values of false positive homology (E-values). Because of that, the search for homology with degenerate motifs was performed using the YASS tool (Noe and Kucherov, 2005) and the library of sequences of 57 transposon superfamilies from 52 *Drosophila* species of total amount—2280 records.

Analysis of the Promoter Structure

To determine the TSS of the *ras85D*, we used the publicly available NCBI databases: SRA (for cDNA libraries), EST, Gene, TSA. The enrichment of the promoter region with promoter elements was estimated using the AME algorithm (McLeay and Bailey, 2010), letter-probability matrix (LPM) of JASPAR POLII motifs library for core promoter elements (Wasserman and Sandelin, 2004; Stormo, 2013) and LPM calculated on the base of positional weight matrices (PWM) for Ohler elements (Rach et al., 2009). Control sets of sequences were obtained in two ways: by cutting fragments of similar length from the upstream region of the intergenic sequence in species that form groups with common pattern of promoter elements (user-provided control sequences), and by generating random permutations based on the composition of the analyzed sequences (shuffled input sequences). To analyze small or highly heterogeneous samples of promoter sequences in the *obscura*, *ananassae*, and *montium* groups the Xstreme algorithm was used (Grant and Bailey, 2021). To visualize the data FIMO algorithm was applied (Grant et al., 2011) using all the identified motifs for each set of sequences and lowering the match *p*-value to 0.01. The obtained coordinates and structures of motifs for each fragment of the sequence with the putative promoter were mapped on the aligned sequences in the MEGA X software and were visualized on the scheme using coordinates of the general consensus sequence.

GO Enrichment Analysis

Estimations of the enrichment of ECMs with the binding sites of transcription factors (TF) and with promoter elements were obtained using the AME algorithm. LPMs from Combined *Drosophila* Databases represented by the MEME SUITE tool were used as a library of motifs of TF binding sites. The results obtained were filtered in accordance to the list of TFs confirmed for the *ras85D* locus of *D. melanogaster* presented in the NCBI and modENCODE genome browsers (Kudron et al., 2018). GO enrichment analysis was performed using the FlyEnrichr Database (Chen et al., 2013; Kuleshov et al., 2016).

Phylogenetic Analysis and Timetree Construction

Phylogenetic constructions and timetree calculations were performed using the MEGA X software. The Maximum Likelihood method was used with the Tamura 3-parameter model (Tamura, 1992) for the coding sequences of the *ras85D* gene of 38 *Drosophila* species including *Scaptodrosophila lebanonensis* as an outgroup. Gamma distribution was used to model evolutionary rate differences among sites [5 categories (+G, parameter = 0.8107)]. The rate variation model allowed for some sites to be evolutionarily invariable [(+I), 19.28% sites]. A timetree was inferred by applying the RelTime method (Tamura et al., 2018). Divergence times for all branching points in the topology were calculated using the Maximum Likelihood method

and Tamura-Nei model (Tamura and Nei, 1993). This analysis involved 28 nucleotide sequences. The estimated log likelihood value of the topology shown is −4498.05. A discrete Gamma distribution was used to model evolutionary rate differences among sites [3 categories (+G, parameter = 3.0235)]. The timetree was computed using 3 calibration constraints. Estimates of the divergence time of *D. lummei—D. novamexicana* (2.9 Ma), species of the *montana* subgroup (4.8 Ma), and all species of the *virilis* group (9–9.5 Ma) were chosen as calibration (Morales-Hojas et al., 2011).

RESULTS

Evolutionarily Conserved Motifs of the Promoter Region and 5′UTR Sequence of Different *Drosophila* Species

The locus of our interest consisting of the *ras85D* upstream intergenic region and 5′UTR, subdivided into two parts by intron 1 contains cis-regulatory elements of *ras85D*. The formal criterion for the common evolutionary origin is a significant sequence homology obtained by alignment. The use of standard sequence alignment algorithms (Unipro UGENE v. 36, Rose et al., 2019; MEGA X, Kumar et al., 2018) made it possible to identify significant areas of homology only in closely related species, for example, in the *melanogaster* and *virilis* groups. For the sequences of all 37 species used in the analysis, it was possible to show the homology of small regions flanking intron 1, the 3′end of intron 1, and the 5′UTR region at the border with exon 1, containing different sets of deletions. This result was confirmed and supplemented using the MEME and MAST algorithms for finding evolutionarily conserved motifs (Bailey and Elkan, 1994; Bailey et al., 2009). The boundaries of ECMs and the significance of homology depend both on the degree of their similarity and on their total number in the total set of sequences. Adding closely related species with high sequence homology to the analysis increases the chances of identifying regions that are overrepresented in these sequences. Therefore, the search for ECMs was carried out in two versions: for 23 species of *Drosophila* from different phylogenetic clades (Set A) and for 37 species, including closely related species from the *melanogaster* and *pseudoobscura* subgroups and the *virilis* group (Set B). The results of the sequence analysis of the 23 species are presented graphically in **Figure 1**, and those for the 37 species are provided in the **Supplementary Materials** (**Supplementary Figure S1**). In the figures, the sequences are aligned from the promoter to exon 1. The species of the subgenera *Drosophila* and *Sophophora* are grouped in the upper and lower parts of the scheme, respectively; the species *D. busckii*, a representative of the subgenus *Dorsilopha*, is located between them. The values of the statistical significance of ECM patterns for each sequence are given, estimated after their extraction and mapping by the MEME algorithms and after additional alignment by the MAST algorithms. In accordance with these values, only the *D. willistoni* sequence has low, albeit significant, E-values, because of the degeneracy of most of the identified ECMs in this species. For each dataset, 14 ECMs were obtained. Their names correspond to the ordinal numbers of each set, assigned in accordance with the results of the two-tiered significance analysis of the motifs selection.

A number of the identified conserved motifs are typical for both versions of the analysis, not exactly coinciding, but showing significant similarity. In the schemes shown in **Figure 1** and **Supplementary Figure S1**, they occupy coinciding or overlapping positions relative to the boundaries of the structural elements of the gene. The accordance of their ordinal numbers to each other is given in **Table 1**. A comparison of sequence logos is given in **Supplementary Figure S2**. The overlap of ECMs sets obtained from both sets of sequences allows revealing the ECMs that are evolutionarily conservative and most likely highly significant, regardless of whether they have a common origin. Similar patterns of the ECMs distribution were obtained for both sets of sequences (for 24 and 37 species) when the fewer expected ECMs were selected in the settings. However, the best ECM topological match was obtained with the selection of 14 expected ECMs.

When specifying the ECM number, hereinafter we will mean the ordinal numbers of the motifs obtained by analyzing the set of sequences from the 23 species, unless otherwise indicated. ECM1 – ECM5 and ECM12 are present in all the species. ECM3 and ECM12 are represented by fragments enriched with AC-dinucleotide repeats and are located in the middle part of intron 1 on the plus strand and in the second half of the 5′UTR both on the plus strand and on the minus strand. The distribution of ECM1, ECM2, ECM3, and 5 suggests that the area of intron 1 and the 5′UTR region, from the small fragment flanking the intron upstream and corresponding to ECM2 to exon 1, has a common evolutionary origin in all the species. It can be seen that this area has numerous deletions and possible degenerate duplications of sequence fragments. ECM4, located near the TSS, is present in all the species, but its location differs among species from different *Drosophila* subgenera. In species of the subgenus *Drosophila*, this motif is located on the minus strand and borders on ECM7, which is located in the promoter region and is specific to these species. ECM7 is also present in these species immediately behind the intron region on the opposite strand. In species of the subgenus *Sophophora*, ECM4 is located on the plus strand and is displaced by 49 bp downstream of the TSS. An exception is the species *D. willistoni*, which diverged earlier than the others from the common ancestor of the presented species of the subgenus *Sophophora*: in *D. willistoni*, a degenerate ECM4 sequence is also identified on the minus strand, as in the species of the subgenus *Drosophila*. In *D. busckii*, a representative of the subgenus *Dorsilopha*, a degenerate form of this motif is displaced upstream of the TSS and is located on the minus strand at the beginning of the intergenic region (**Figure 1**). The lack of homology of the left 5′UTR fragment in the species of the subgenera *Drosophila* and *Sophophora* is also confirmed by the fact that this region features ECM6, which is unique to *Sophophora*, and ECM13, which is unique to *Drosophila*.

The intergenic region is the most variable of all, its size varies from more than 2000 bp in species of the *virilis* group to an almost complete absence in a number of species of the *obscura* group (*D.*

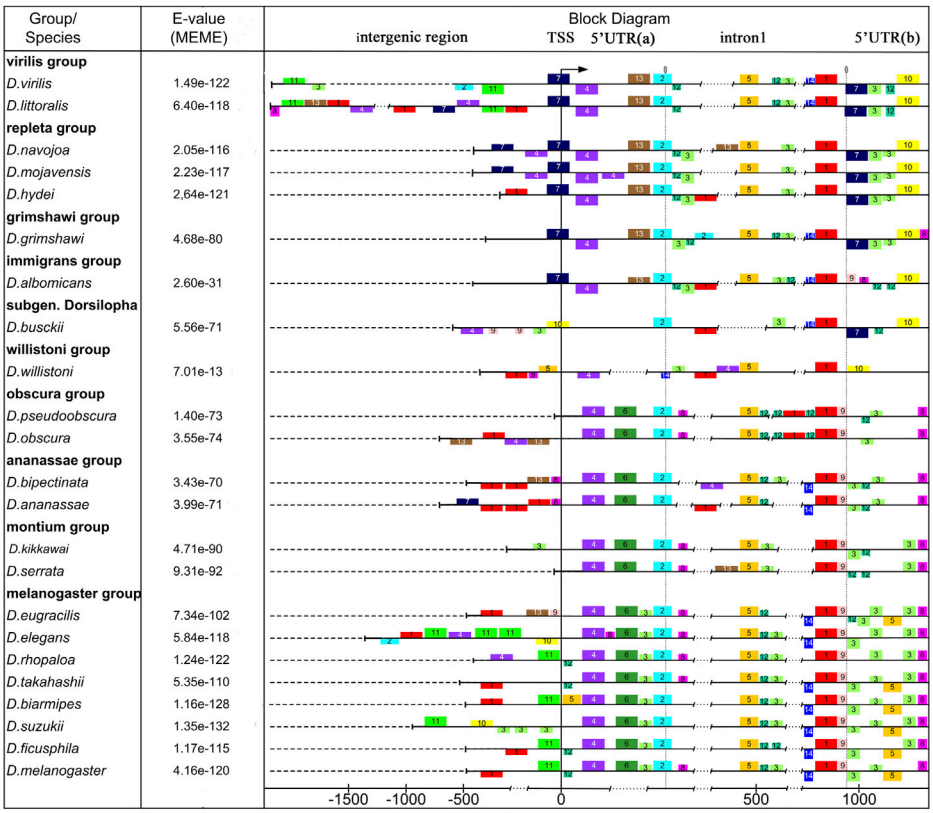


FIGURE 1 | Polymorphism of the intergenic region, promoter, 5′UTR, and intron 1 in *Drosophila* species with different degrees of relationship. Scheme of the distribution of evolutionarily conserved sequences and insertion-deletion polymorphism in the analyzed region of the *ras85D* gene. The structure of evolutionarily conserved sequences which were obtained using the MEME algorithm. In the scheme, the boxes above the line show the location of ECMs on the plus strand, while the boxes below the line show ECM location on the minus-strand. The reduced box height marks degenerate ECMs. The upper part shows species of the subgenus *Drosophila (D. virilis—D. albomicans),* and in the lower part, there are species of the subgenus *Sophophora (D. willistoni—D. melanogaster). D. busckii (*subgenus *Dorsilopha)* is located between the subgenera *Drosophila* and *Sophophora*.

TABLE 1 | Homology of evolutionarily conserved motifs isolated from the Set A (37 species) and the Set B (23 species) of the *ras85D* promoter region sequences.

Set A (23 sp.), conservative_sequence	Strand	Set B (37 sp.), conservative_sequence	Strand
ECM2 (E = 3.6e-257)	+	ECM1 (E = 1.4e-734)	+
ECM1 (E = 2.7e-315)	+	ECM2 (E = 1.3e-738)	+
ECM3 (E = 1.2e-234)	–	ECM3 (E = 1.8e-356)	+
ECM12 (E = 2.4e-028)	–	ECM3 (E = 1.8e-356)	+
ECM4 (E = 7.9e-135)	+	ECM4 (E = 7.9e-296)	+
ECM4 (E = 7.9e-135)	+	ECM13 (E = 4.7e-197)	–
ECM5 (E = 1.3e-109)	+	ECM5 (E = 2.8e-441)	+
ECM7 (E = 9.7e-069)	–	ECM8 (E = 3.5e-220)	+
ECM7 (E = 9.7e-069)	+	ECM9 (E = 1.9e-294)	+
ECM4 (E = 7.9e-135)	–	ECM7 (E = 3.7e-271)	+/–*
ECM13 (E = 3.8e-026)	+	ECM11 (E = 1.2e-257)	+

**motif 7 from Set B is represented by a degenerate palindromic sequence homologous in the 5′-region to a fragment of motif 4 from the minus strand from Set A, and in the 3′-region to the same fragment from the plus strand.*

pseudoobscura, *D. persimilis*) and in *D. serrata* from the *montium* group. Sequence homology can be detected only in selected groups of species with significant phylogenetic similarity. Thus, ECM11 was found in species of the *melanogaster* group of the *melanogaster*, *rhopaloa*, *ficusfila*, and *suzukii* subgroups. The position of ECM11 can be significantly shifted relative to the promoter, in particular, due to insertions and the multiplication of AC-repeats in ECM3, as in *D. suzukii*. In the subgenus *Drosophila*, ECM11 is found only in the *virilis* group: it is located on the minus strand distally of the promoter. All the other ECMs of the intergenic sequences identified in species from different phylogenetic clades are represented by degenerate copies.

The general picture of the variation of the entire non-coding sequence suggests that the main role in the regulation of the *ras85D* gene expression is played by the motifs located in the 5′UTR and the intron 1 regions. Even when there is rearrangement in this area, the ECM4 regains its position.

The *ras85D* Gene Neighbourhood and Rearrangements

The *ras85D* gene is located in the Muller E element, which corresponds to chromosome 2 in *Drosophila* of the *virilis* group (Mitrofanov et al., 2011) and the right arm of chromosome 3 in *Drosophila* of the *melanogaster* group. The ras85D gene in *D. virilis is inverted in relation to neighboring genes-orthologs in D. melanogaster* (**Supplementary Figure S3**). As a result of the fixed chromosomal rearrangements, the closest gene neighbourhood of *ras85D* in species of the *virilis* group differs from that of all other *Drosophila* species. The genes located upstream of the *ras85D* gene in species of the *virilis* group [*GJ23372* (ortholog of the *D. melanogaster CG31344* gene), *GJ23373* (*D.m. Rpb7*) and *GJ23843* (*D.m. CG12241*)] are located at a distance of 5710 kB in the *D. melanogaster* genome (**Supplementary Figure S3**). The genes located in this place in other species (except for *D. willistoni*) are as follows: *Rlb1* (ortholog of the *GJ10856* gene in *D. virilis*), *mRpL47* (*D.v. GJ10316*), *JHDM2* (*D.v. GJ10857*), and *CG8176* (*D.v. GJ10858*), all of which are located at a distance of 680 Kb from the *ras85D* gene in the *D. virilis* genome (**Supplementary Figure S3**). The 3′-intergenic region also underwent rearrangement in the ancestor of the *virilis* group. There is an insertion of two genes between the 3′UTR of the *D. virilis ras85D* gene and the downstream set of genes typical for other species. One of them, *CJ23371* (ortholog of the *D. melanogaster AOX2* gene), is located at a distance of 6039 kB from the *ras85D* gene in the *D. melanogaster* genome. The second gene, *GJ26119*, is a chimeric gene that contains a fragment of the coding sequence of the retrotransposon *CR1-1* gene for pol-like protein with endonuclease and reverse transcriptase (RT) domains more than 2500 bp in length. This also points to frequent rearrangement events associated with the activity of transposons. The genetic environment characteristic of the species of the *melanogaster* group remains unchanged in other *Drosophila* species from both subgenera, for example, *D. pseudoobscura*, *D. ananassae*, *D. grimshawi*, and *D. mojavensis*. Only in *D. willistoni*, two genes—*Rlb1* and *mRpL47*—were reversed as a result of an inversion and are located on the opposite strand of the chromosome. Thus, the divergence of species is often accompanied by periods of genome instability and chromosomal rearrangements, which is associated with the activity of mobile elements (Hedges and Deininger, 2007).

The Participation of Mobile Elements in Evolution of the *ras85D* Intergenic Region, Promoterand downstream 5'UTR sequence

To search for traces of mobile element insertions, we used the GIRI Repbase database and its software Censor tool (Kohany et al., 2006). The search was carried out along the entire length of the sequence from the putative TSS of the upstream gene to exon 1 of *ras85D*. The analysis results are shown in **Table 2** and in more detail in the **Supplementary Materials** (**Supplementary Table S1**). The tables also include the search results for inverted repeats flanking some species-specific and species group-specific sequences. The analysis of repeats was carried out using the software tool YASS (Noe and Kucherov, 2005). The complete information on the analysis of the intergenic region in *Drosophila* of the *virilis* group is given in **Supplementary Table S2**. Non-random similarity with the sequences of mobile elements was found in 25 species of *Drosophila*, including 11 species of the *virilis* group.

Sequence elements of all the analyzed species showing homology with transposon sequences were excluded from the results of the analysis. Homology variants represented mainly by microsatellite repeats were also excluded. On the contrary, homology was considered characteristic of a group or subgroup of species if the consensus of a sequence element shared by the given group could be restored from fragments preserved in different species of this group.

According to the results, the following mobile elements may have participated in the formation of the evolutionary variation of the *ras85D* gene promoter region: DNA transposons belonging to the superfamilies *P, Helitron, hAT, Mariner, PiggyBac, Transib, DNA8*, and *Polinton*; LTR retrotransposons such as *Gypsy* and *BEL*; and non-LTR retrotransposons such as *L2, R1* and *CR1*. The events of transposon incorporation in the studied non-coding sequence during the evolutionary divergence of the *Drosophila* species are mapped on the phylogenetic tree (**Figure 2**).

The *ras85D* gene promoter region in species of the subgenera *Drosophila* and *Sophophora* is represented by three main patterns of ECMs location, all of which have independent origins. The 54 nucleotides before ECM4 in the species of the *montium*, *obscura*, and *ananassae* groups, which are part of the promoter, demonstrate a significant homology not only with each other, but also with a representative of the genus *Scaptodrosophila*, *S.lebanonensis*, which indicates the ancestral state of this region in these three groups of species. In this region, homology is found with the beginning of LTR retrotransposon BEL: this homology continues in the distal direction in *D. ananassae* and *D. bipectinata*. In the upstream part of the intergenic region, these species display homology with the terminal repeats of *Transib* DNA transposon and *Gypsy* LTR retrotransposon. In *D. pseudoobscura*, *D. persimilis*, and *D. serrata*, homology in this region is disrupted by large deletions, which may be a consequence of recombination of similar motifs of the mobile elements involved in the formation of this region. Thus, in *D. kikkawai*, traces of incorporation of a *Gypsy* group transposon are found in this region, while in *D. obscura*, inverted repeats flank

TABLE 2 | Homology with the transposon sequences of the *ras85D* noncoding and the upstream intergenic regions in *Drosophila* species.

Species	DNA transposon	LTR retrotransposon	Non-LTR Retrotr.
D. willistoni	*Homo5_hAT*^{IS}	*Gypsy1*^{P}	-
D .pseudoobscura	*P-1*^{IN*}	-	*L2-10*^{IN*}
D. persimilis	-	-	*L2-10*^{IN*}
D. obscura	*P-3*^{IS}, *Helitron-N1*^{IS}, *HelitronN-1*^{IS}, In.rep^{IS}	-	-
D. suzukii	*Homo10 hAT*^{IS}	*Gypsy-5*^{IS}, *Gypsy-23*^{IN***}	-
D. takahashii	*Hoana5/hAT*^{P*}	-	-
D. eugracilis	-	*invaider3-Gypsy*^{IS}, *Gypsy-11*^{IS}	-
D. elegans	*P-1N1*^{IS}, *piggyBac-N1*^{IS}, *Hoana5/hAT*^{P*}, *hAT-1*^{P*}	*Gypsy-1*^{IS}, *Gypsy-N1B*$^{5'UTR*}$	-
D. ananassae	-	*Gypsy1-LTR*^{IS*}, *Gypsy20-I*^{IS*}	-
D. bipectinata	*Transib-8*^{IS*}, *Transib-2*^{IS*}	*Gypsy-37*^{IS*}, *BEL-17*^{IS*}	-
D. kikkawai	*Transib-1*^{IS}	*Gypsy-24*^{IS}, *Gypsy-2*^{IS}	-
D. busckii	-	*Gypsy-1*^{IS}	-
D. albomicans	*Mariner-4*^{IS}	*Gypsy-24*^{P***}	-
D. hydei	-	-	*R1-4B*^{IS**}
D. mojavensis	*hAT-2*^{IS*}	-	*Jockey-7*^{IS*}, *R1-3*^{IN***}
D. grimshawi	-	*BEL-7*^{IS**}	*CR1-1*^{IS***}
Sp. gr. virilis	Uncharacterized X –DNA TE In.rep.^{IS*}	-	
D. kanekoi	*hAT-8*$^{IS*V/K}$, *Helitron-like-5*$^{IS*V/K}$	-	*L2-10*^{IN*}***
D. ezoana	*DNA8-78*$^{IS*V/K}$, *Helitron-N10*$^{IS*V/K}$	*Gypsy-22*^{IN****}	-
D. littoralis	*hAT-11*$^{IS*V/K}$, *Helitron-N10*$^{IS*V/K}$, *hAT-N36*$^{IS*V/K}$	*Gypsy-22*^{IN****}	-
D. lummei	-	*Gypsy-1*$^{IS*V/K}$, *Gypsy-22*^{IN****}	-
D. virilis	-	*BEL-10*^{IN****}	-
ancestral *virilis- kanekoi* s/gr.	*Polinton-2*$^{IS*V/K}$, *Helitron-N10*$^{IS*V/K}$, *Helitron-1*$^{IS*V/K}$	*Gypsy-22*^{IN****}	-

*The superscripts indicate the region-specific position of the motif (IS, intergenic sequence; P, promoter, 5′UTR; IN, intron 1) and asterisks mark the homology of the aligned sequences at the level: *V/K—subgroups, *—own group of species, **—two related groups of species, ***—own Subgenus, ****—two related Subgenera.*

the intergenic region and in the proximal part are located close to the region of common homology. These repeats are located in close proximity to or overlap with motifs characteristic of DNA transposon terminal repeats.

On the evolutionary branch of the *obscura* group, an extended insertion into intron 1 shows possible homology with either the 3′UTR region of an *L2* retrotransposon or with the terminal repeats of DNA transposon *P* (**Supplementary Table S1**). At the same time, in the species of the *pseudoobscura* subgroup, the homology with the transposon is more significant than in *D. obscura*.

For the species of the *melanogaster* group, according to the sequences of *D. takahashii* and *D. elegans*, it was possible to identify homology of the promoter region with inverted repeats of DNA transposons from the *hAT* superfamily. This points to a common evolutionary origin of the intergenic sequence, including the *ras85D* promoter, in most species of this group, by insertion of a mobile element prior to the divergence of the main subgroups of the *melanogaster* group. The noted variation of this region is related both to point mutations and to microindels. In species of the subgroups *melanogaster*, *ficusphila*, and *rhopaloa*, the short basal part of this insertion is deleted, about 10 bp. In the species *D. elegans*, *D. suzukii*, and *D. eugracilis*, later independent insertions of mobile elements belonging to the *P, hAT*, and *Gypsy* groups were noted in the area of the first insertion in the promoter region (**Table 2**). Moreover, in *D. eugracilis*, this led to the loss of the area of general homology with the other species (**Supplementary Figure S4**). The fragment of the intergenic region, which has a common origin in the species of this group, is polymorphic and has accumulated multiple deletions and point mutations in the distal part. In this case, the boundaries of deletions in different species do not coincide, which indicates independent events of sequence region loss. In phylogenetically more distant species of the *melanogaster* and *suzukii* subgroups, the area of homology accumulates deletions, so the common evolutionary origin of these regions is difficult to detect without prior alignment with other sister species.

In *D. willistoni*, the area of reliable homology of the non-coding sequence begins in the region of intron 1, but a small sequence fragment of 20 bp in length in the region of the promoter shows a strict homology with the species of the *montium*, *obscura*, and *ananassae* groups. The intergenic region in this species was formed as a result of an inversion rearrangement, and the most likely participant of this event is a retrotransposon related to *Gypsy*. Homology with the LTR of the retrotransposon is shown for the region of the *D. willistoni* sequence immediately behind the promoter region. In this case, the rearrangement completely replaced the ancestral sequence of the intergenic region and the 5′-part of the 5′UTR. The question of how the area of the promoter sequence showing homology with the species of the *montium*, *obscura*, and *ananassae* groups was preserved in this region remains open and will be discussed below.

In species of the subgenus *Drosophila* (*D. hydei* and *D. grimshawi*), ECM7 in the promoter region exhibits homology (65–70%) with the 5′UTR of non-LTR retrotransposons *R1* and *CR1*. Sequence differences appear in the region upstream of the promoter by 70 bp and lead to the formation of four independent patterns characterizing the species of the groups *virilis*, *repleta*, *grimshawi*, and *immigrans*. In *D. albomicans* (the *immigrans* group), this region shows homology with DNA transposon

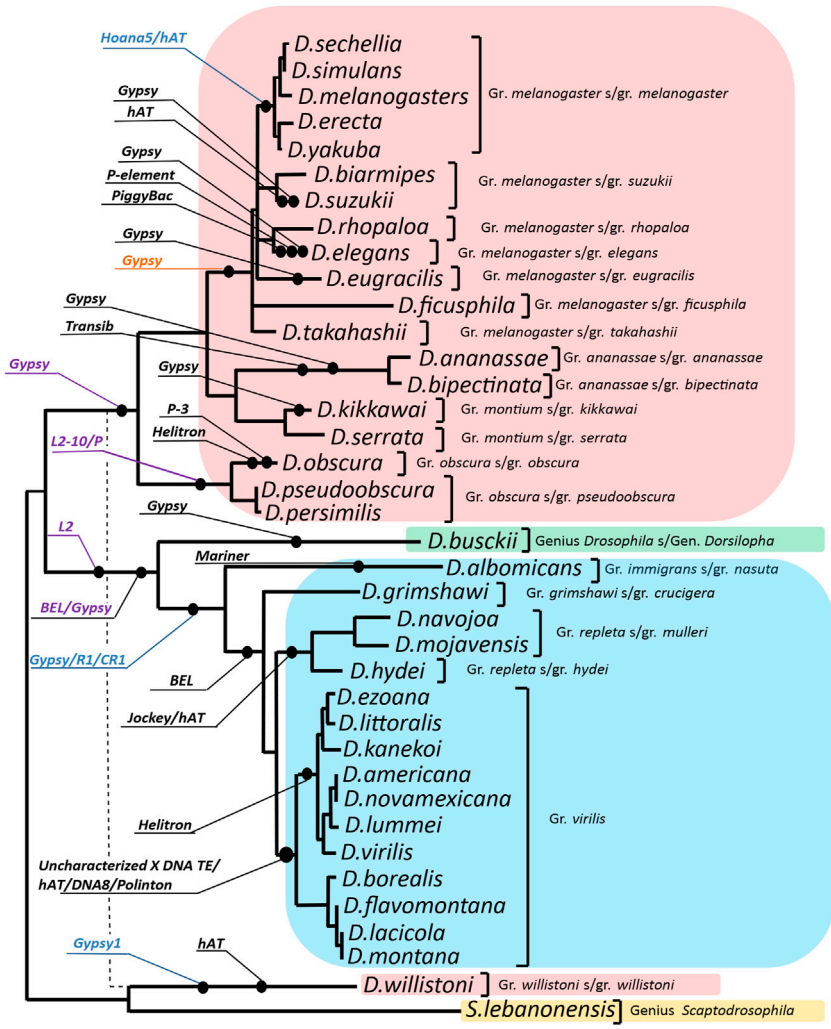


FIGURE 2 | Distribution of the identified transposon insertion events on the evolutionary tree of the analyzed *Drosophila* species. The phylogenetic tree was constructed using the Maximum Likelihood method on the base of the Tamura 3-parameter model. Sequences of exons and 3′UTR of the *ras85D* gene of 37 *Drosophila* species were used. Gamma distribution was used to model evolutionary rate differences among sites [5 categories (+G, parameter = 0.8107)]. The rate variation model allowed some sites to be evolutionarily invariable [(+I), 19.28% sites]. The structure of the resulting phylogenetic tree is generally consistent with the species tree obtained for 155 genomes of *Drosophila* species (Suvorov et al., 2021) excluding *D. willistoni*, for which the position on the species tree is indicated by a dashed line. The callouts specify the transposable element and insertion events on the branches of the evolutionary tree. Callout-designated insertions in the following regions: intergenic region, promoter, 5′UTR, and intron 1, were highlighted in black, blue, purple, and orange, respectively.

Mariner, in *D. grimshawi*, with LTR retrotransposon *BEL*, in *D. mojavensis*, there is a similarity with non-LTR retrotransposon *Jockey* or DNA transposon *hAT*. In species of the *virilis* group, an inversion rearrangement that affected the *ras85D* gene promoter region was accompanied by the appearance of inverted repeats flanking almost the entire intergenic region. Assessments of the position and structure of the repeats in 11 species of the *virilis* group, carried out using the YASS tool, indicate the fundamental similarity of the inverted repeats and their presence in all the species of this group (**Supplementary Table S2**). Both ends of the *ras85D* upstream intergenic sequence contain two to four fragments located sequentially and demonstrating significant homology of the inverted sequences (Score> 82; bit-score> 25.3; E-value <0.044). The alignment of the entire area of the intergenic region in species of the *virilis* group is shown in **Figure 3**. The region of direct repeats occupies the central part of the intergenic sequence. There is an individual direct repeats area in the middle part of the left inverted repeat (hereinafter referred to as Ir-a). As a result, Ir-a is 730 bp in length and turns out to be longer than the right inverted repeat (hereinafter referred to as Ir-b), which is 380 bp in length. Both repeats underwent various deletions in different species, as can be seen from the scheme of the arrangement of homologous regions in different species (**Figure 3**). The deletion mechanism of repeat divergence is confirmed by a significant homology of overlapping fragments of these repeats (**Supplementary Figure S5**).

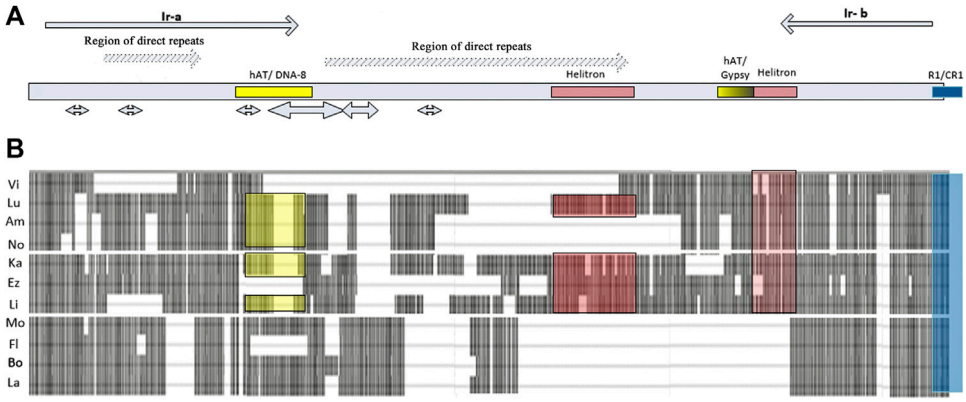


FIGURE 3 | The structure of the *ras85D* upstream intergenic region in *Drosophila* species of the *virilis* group. **(A)** Repeats, palindromes, and areas of homology with mobile elements: inverted repeats flanking the sequence are designated as Ir-a and Ir-b and are represented in the scheme by gray arrows; regions of direct repeats are represented by light gray arrows with diagonal hatching; palindromes are shown by double-headed arrows; areas of homology with terminal repeats of transposons are marked by colored rectangles. **(B)** Insertion-deletion polymorphism of the regulatory region: rectangular gray blocks represent the DNA sequence, dashed lines and gaps between the blocks show the deletions.

The Ancient Insertion of *Uncharacterized X DNA* Transposon and its Involvement in the Rearrangement of Chromosome 2 in the *Virilis* Group

There is no reliable homology of the inverted repeats of the *virilis* group with any known mobile element. However, in several species of the group, the 3′-end of Ir-a, 40–44 bp in length, resembles the terminal repeats of DNA transposons belonging to the *hAT, DNA8,* and *Polinton* superfamilies (**Figure 2**, **Supplementary Table S1**). In addition, numerous sequences of inverted repeats are found in the *D. virilis* genome. 37 degenerate sequences homologous to the inverted repeats Ir-a and Ir-b from closely related species of the *virilis* group have been found in the *D. virilis* genome by the BLAST method within different scaffolds. The length of these sequences ranges from 54 to 310 bp, while the length of 16 fragments exceeds 190 bp. (**Supplementary Table S3**). All the fragments are located in non-coding regions of the genome, and more than half of them have double confirmation either by homology with both repeats of the same species, or by homology with one of the repeats from two species. We have hypothesized that traces of inverted terminal repeats of some mobile element can be found along the boundaries of ancient rearrangements, including the rearrangement that led to a change in the gene environment of the *ras85D* gene in the ancestor of the *virilis* group, in regard to the gene order characteristic of other species in the subgenera *Drosophila* and *Sophophora*. To obtain a structure characteristic of the *virilis* group from the ancestral structure preserved in *D. melanogaster* and *D. grimshawi*, at least three inversions are required. One falls onto the *ras85D—Rlb1* and *CG31344—Caf1-55* intergenic regions, and the other two capture the regions of the formed *Rlb1—Caf1-55* and *ras85D—CG31344* blocks and turn them over into the opposite direction (**Figure 4**). In *D. virilis*, the order of genes in the area of chromosome 2 corresponding to scaffolds sc_13047 and sc_12855 differs from that of other species of the *virilis* group due to one more inversion (2A) described earlier (Reis et al., 2018), which occurred with the participation of DNA transposon *DAIBAM* (*hAT* superfamily) and implicated the same *Caf1-55—Rlb1* intergenic region on one side and the *invadolysin* gene on the other side.

As expected, the prepromoter region of the *D. virilis Rlb1* gene contains one of the inverted Ir-b repeat copies of 190 bp in length (sequence no. 5) (**Supplementary Table S3**), which exhibits the highest homology with the Ir-b repeat of *D. lacicola*. It occupies the area from the inverted repeat of *DAIBAM* to the TSS of the *Rlb1* gene and continues further into the region of the 5′UTR of the *Rlb1* gene. In the genomes of *D. montana* and *D. americana* that did not undergo the 2A rearrangement, the homology area with repeats is longer and consist of about 300 bp, i.e., more than 80% of the Ir-b repeat (**Supplementary Table S4**, **Supplementary Figure S6**).

Another copy, which should be located in the *Caf1-55* gene region and correspond to Ir-a, was not found in the *D. virilis* genome. However, there is a fragment from the right side of the Ir-a repeat of 107 bp in length in the central part of the *Caf1-55—Rlb1* intergenic region of the *D. americana* and *D. montana* genomes. This fragment is located upstream of the area of homology with the Ir-b element by 36 bp, but is oriented with its tail towards the *Caf1-55* gene (**Supplementary Table S4**, **Supplementary Figure S6**), i.e., in the direction opposite to the expected. A possible scenario of events leading to this result is considered in experiments with induced genomic instability with chimeric DNA transposons *P* in *D. melanogaster*, which have disruptions of inverted terminal repeats (Georgiev et al., 1997; Pomerantseva et al., 2006). Activation of transposase in this case leads to the formation of complete or partial deletions of the copies of transposons and downstream sequences, and to insertions of inverted sequence fragments. It can be concluded that the change in the *ras85D* upstream intergenic region in *Drosophila* of the *virilis* group occurred under the influence of insertion of some DNA

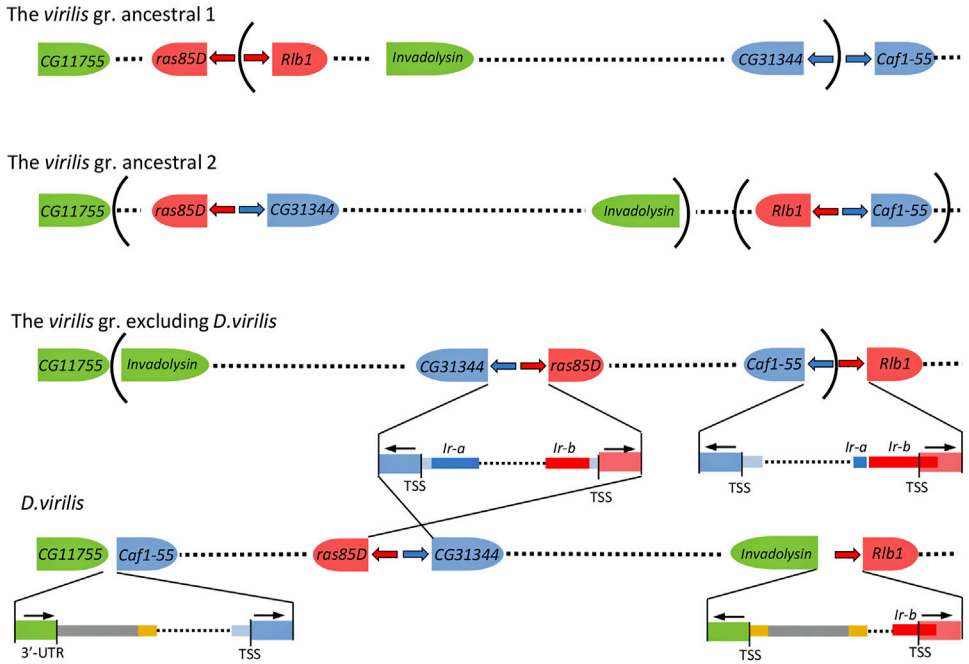


FIGURE 4 | The minimal number of rearrangements required to transform the ancestral state of chromosome 2 to specific for the *virilis* group. Blue and red arrows—the inverted repeats Ir-a and Ir-b, respectively, that marked insertion sites of two copies of the Uncharacterized X DNA Transposon into the *ras85D—Rlb1* and *CG31344—Caf1-55* intergenic regions in the ancestor of the *virilis* species group. *DAIBAM* elements are colored with yellow (inverted terminal sequences) and gray (Reis, et al., 2018). Remnants of ancestral sequences are colored with light blue. The nomenclature of the genes corresponds to the names of *D. melanogaster* orthologs. The representation is not to scale.

transposon (uncharacterized *X* DNA TE) with inverted terminal repeats and subsequent inversion is associated with this insertion.

Taking into account the common origin and homology of Ir-b repeats from the *ras85D—CG31344* and *Rlb1—invadolysin* intergenic regions in *D. virilis*, it is possible to estimate the divergence time of these repeats, which has elapsed since the beginning of the accumulation of polymorphism between the two copies of transposons, as well as the time of the divergence between the right and left inverted terminal repeats of one transposon. The divergence of two repeats (Ir-b) of the ancestral copy of the transposon from the *ras85D—Rlb1* intergenic region is shown relative to the repeats (Ir-a) from another copy of transposon corresponding to the *Caf1-55—CG31344*intergenic region, that were taken as outgroup (**Figure 5**). The timetree was constructed taking into account the calibration estimates of the divergence time of *D. lummei—D. novamexicana* (2.9 million years), the species of the *D. montana* subgroup (4.8 million years), and all the species of the *virilis* group (9–9.5 million years), obtained by Morales-Hojas using multilocus data (Morales-Hojas et al., 2011). It should be noted that the divergence time of Ir-b repeats located close to the *ras85D* and *Rlb1* genes, i.e., the time elapsed from the moment when the transposon had lost its autonomy, is 20 million years. The picture of species relationship, similar to the generally accepted one in the composition and relationship of subphylads, is reproduced using gaps as the fifth base, but taking into account the length of repeats shortened by a quarter in the *Rlb1* region. Such an estimate leads to a revised assessment of the number of substitutions in these repeats and to an inflation of the divergence time (**Figure 5**). Nevertheless, the result corresponds to the time period when the common ancestor of the *virilis* group species was existed, after the division of the *virilis—repleta* clade, which occurred, according to various estimates, from 48 to 24 million years ago (Ross et al., 2003; Adryan and Russell, 2012).

It can be concluded that mobile elements marked by their presence almost all events associated with significant changes in the sequence of the intergenic region located upstream of the *ras85D* gene. When comparing related species, the degradation of areas formed as a result of insertions and rearrangements can be seen. Moreover, the absence of traces of the coding sequences specific for the corresponding transposons superfamilies in the intergenic region sequence of related species indicates the rapid degradation of a significant part of the mobile element sequence.

Structure and Origin Sources of the *ras85D* Promoter in Different Species

Replacement of the promoter region suggests that in the absence of a new point of assemblage of the preinitiation complex, the gene will be inactivated. In the case of conserved genes with key functions in significant biological processes, the resulting lethal allele will be displaced from the population. The presence of such

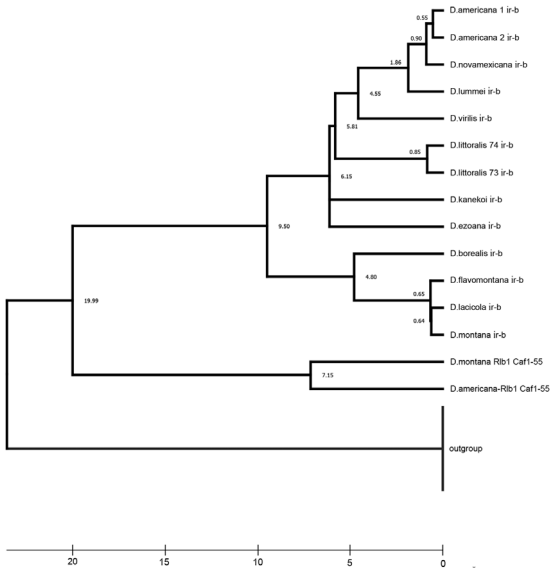


FIGURE 5 | Estimated time of the ancestral rearrangement occurrence, based on the time of mutation accumulation of homologous Ir-b repeats. The Timescale at the bottom of the figure is in millions of years. A timetree inferred by applying the RelTime method (Tamura et al., 2012). The timetree was computed using three calibration constraints. This analysis involved 28 nucleotide sequences. Evolutionary analyses were conducted in MEGA X (Kumar et al., 2018). Divergence times for all branching points in the topology were calculated using the Maximum Likelihood method and Tamura-Nei model (Tamura and Nei, 1993). The estimated log-likelihood value of the topology shown is −4498.05. A discrete Gamma distribution was used to model evolutionary rate differences among sites [three categories (+G, parameter = 3.0235)]. The tree is drawn to scale, with branch lengths measured in the relative number of substitutions per site. This analysis involved 28 nucleotide sequences of the inverted repeats. There were a total of 700 positions in the final dataset.

alleles, fixed in different species, indicates a high rate of restoration of the functional activity of the gene after rearrangement of the promoter region. Let us consider the structure of the promoters and TSSs of the *ras85D* gene and its orthologs in the studied species.

The use of publicly available NCBI databases, such as SRA (for cDNA libraries), EST, Nucleotide, and TSA, nevertheless, leaves open the question of the correspondence of the cDNA data presented in them to the real TSS position. To assess the accuracy of TSS revealing from these data, we compared the results obtained with the TSS analysis data we received using the standard methods for revealing capped transcript ends: cap-trapped expressed sequence tag (EST), cap analysis of gene expression (CAGE), and 5′-end serial analysis of gene expression (SAGE). The results of detailed cross-genomic analysis of TSSs for *D. melanogaster* by the 5′-SAGE method in combination with NGS of the obtained libraries according to the Illumina/Solexa protocols are published in the MachiBase database (Ahsan et al., 2009). The picture of the distribution of reads in the promoter region obtained in this study has shown the need to take into account their frequencies when revealing the reference TSS. Later Hoskins and co-authors, using the methods CAGE, RACE, and RE EST, and Rach and co-authors, using computational methods, have shown that according to the distribution pattern of TSS reads, promoters are differentiated into classes of broad (or weak peaked), narrow peaked, and broad peaked promoters (Hoskins et al., 2011; Rach et al., 2011). The form of TSS distribution is strictly related to the features of expression regulation: narrow peaked promoters determine the expression of luxury genes with spatial and temporal expression restrictions, while broad promoters determine the constant and ubiquitous expression of housekeeping genes (Hendrix et al., 2008; Juven-Gershon and Kadonaga, 2010; Ni et al., 2010; Hoskins et al., 2011).

The *ras85D* gene is expressed in all cells of an organism throughout its lifetime, but the expression levels of the gene, depending on the tissue type and stage of development, can differ by an order of magnitude or more, according to the SRA data from the public ENA database (European Nucleotide Archive), thus occupying an intermediate position between housekeeping genes and regulated genes. The promoter shape of this gene shows signs of a broad peaked promoter: the reads are densely located in an area of about 100 bp in length and have a pronounced peak in the middle of the area (**Figure 6**). The figure shows that the SAGE and CAGE data are in good agreement with each other, differing only in the more pronounced TSS peaks for the SAGE data, which is probably caused by an underestimation of the biochemical background formed by degraded and re-capped fragments. The EST data also confirm the noted shape of the promoter, somewhat narrowing the TSS area and bringing it closer to narrow peaked promoters, which is associated with insufficiency of EST data (Hoskins et al., 2011). SRA data do not allow us to determine the shape of the promoter precisely, but mark the area of major gene expression rather accurately.

For most species, only SRA data and the results of predicted estimates of possible TSSs (predicted data) are available, to which, in some cases, data from the TSA and EST libraries are added. The numbers of the SRA and TSA libraries, which were used to perform a BLAST to determine TSSs in *Drosophila* species, as well as the EST data and the accession numbers of predicted mRNA used in the analysis, are presented in **Supplementary Table S5**. The analysis results do not always coincide, and SRA data in many cases demonstrate the presence of short sequence fragments marked with reads preceding the area of continuous coverage with reads. Although the analysis of the *D. melanogaster* sequence suggests a broad peaked form of the promoter, significant changes in the sequence in other species can lead to both a change in the type of the promoter and the appearance of additional promoters that share in the regulation of the functional activity of the gene (Haberle et al., 2019). The beginning of each fragment of this kind can be considered to be either an erroneous result, or a minor, additional TSS, caused by a random and defective combination of promoter elements. When analyzing the SRA, EST, and TSA libraries, promoter elements were searched in the area of 100 bp from all detected TSSs and in the area of a sharp rise in read frequencies in the

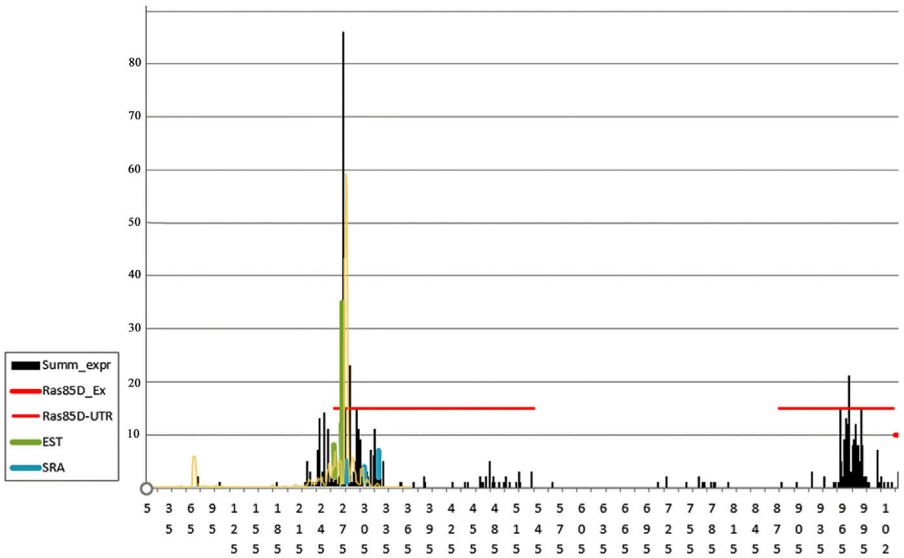


FIGURE 6 | Frequencies of reads around TSS in the intergenic region and 5′UTR of the *ras85D* gene in *D. melanogaster.* Horizontal red lines mark the boundaries of the 5′UTR fragments broken by intron 1. The number of reads is indicated on the Y-axis. The total data on the TSS read frequencies, obtained by the 5′-SAGE method from samples of embryos, larvae, young females, young males, old females, and old males (MachiBase), are highlighted in black; the data obtained from a sample of embryos using the CAGE method are marked in yellow (Hoskins et al., 2011), the reads of the EST and SRA libraries are shown in green and blue **(Supplementary Table S5)**.

promoter neighbourhood. Obviously, such an analysis does not allow the detection of double promoters in the absence of a significant effect of the second promoter, because the most of the data were obtained at one stage of development: imago. However, this analysis allows us to assess the randomness of the coincidence of major and minor TSSs in related species. We considered the following as confirmation of the correct localization of the promoter: coincidence of the TSS localization in the analyzed species according to data from different sources, coincidence with the predicted TSS, and position coincidence in closely related species.

For information on sequence enrichment with promoter elements in the putative TSS region, see **Supplementary Table S6**. The distribution of potential promoters in the analyzed species is schematically shown in **Figures 7**, **8**, **9**. The positions of the identified promoter elements are shown on aligned sequences, presented in **Figure 1**. Several typical structures can be distinguished according to pattern similarities in the promoter structures located at the beginning of the area of continuous coverage with reads.

Most of the species belonging to the *melanogaster* group of the subgenus *Sophophora*, with the exception of *D. eugracilis* and *D. suzukii*, show a high homology of the sequence area marked by the beginning of the area of continuous coverage with reads. This area roughly corresponds to the predicted TSS of *ras85D* in *D. melanogaster*, covers an area of 100 bp around the TSS, including ECM11 located upstream, and is closely adjacent to ECM4 downstream (**Figure 7**, **Supplementary Figure S4**, the first 10 species). The presence of point mutations, insertions, and deletions does not undermine the conclusion about the general evolutionary origin of this area in these species. A characteristic feature of the putative promoter is the obligatory presence of Ohler7 and DRE elements, with which this area is enriched. The enrichment estimates were obtained both in regard to the intergenic sequence and to randomly generated permutations based on the nucleotide composition of the analyzed sequences using the AME algorithm (McLeay and Bailey, 2010). Such distribution of Ohler7 and DRE elements indicates the effect of selection, which maintains the conserved structure and position of these elements in the promoter region. This region also contains the following elements: BRE down (BRE^{d}), Initiator (Inr), and Downstream Promoter Element (DPE). These elements are randomly distributed within the promoter region, and the enrichment of them was statistically insignificant with respect to the intergenic sequence (**Supplementary Table S6**). Location of reads marking distal TSS positions can vary within 50 bp, but a significant increase in the number of SRA reads, suggesting the highest start frequency, is in most cases associated with the position of the Ohler7 element. Independent transcription starts, located upstream from the main promoter and not connected with it by the area of continuous coverage with reads, were noted for the species *D. erecta*, *D. sechellia*, *D. melanogaster*, *D. takahashii*, and *D. suzukii*. In the area of these TSSs, the elements Inr, TCT, TATA-box, BRE^{d}, and DPE may be present, but the canonical arrangement of these elements relative to the transcription start (Lee et al., 2005; Gershenzon et al., 2006; Theisen et al., 2010; Danino et al., 2015; Vo Ngoc et al., 2019) is observed only in some cases. In *D. suzukii*, the element Ohler1 was also found in the area of single starts.

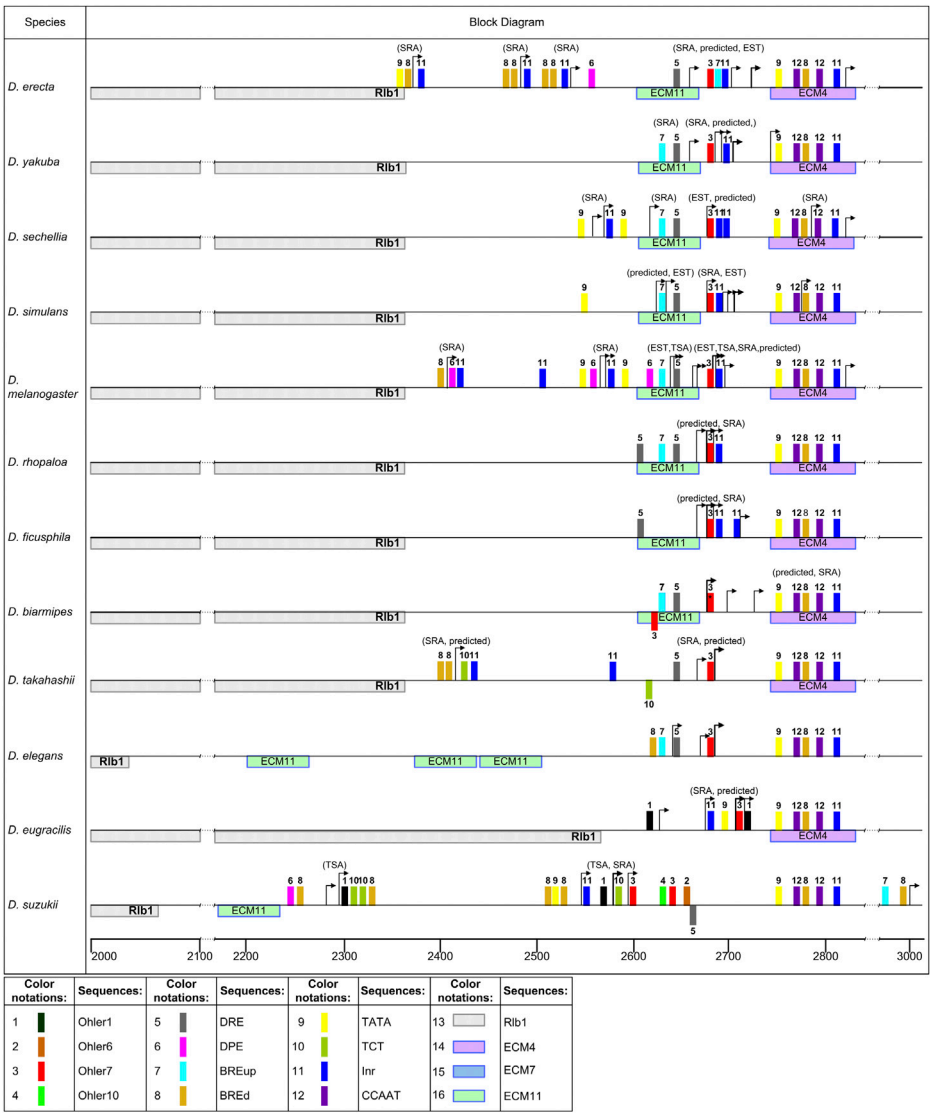


FIGURE 7 | Patterns of promoter elements in the TSS regions of the *ras85D* gene and its orthologs in the studied *Drosophila* species. Distribution of the elements in the promoter region in the *melanogaster* group. The scale below the line shows the position from the start of the overall alignment. Colored rectangles represent promoter elements. The elements shown under the straight line indicate homology with the minus-strand of DNA. TSS are marked with arrows. The gene located upstream is marked on the left side of the diagrams.

In the species *D. eugracilis* and *D. suzukii* of the *melanogaster* group, insertions of transposones into the intergenic region partially or completely cover the region of the promoter (**Figure 7**). In *D. eugracilis*, most part of the intergenic region was deleted, and the promoter region remained in the same position relative to ECM4, but lost the DRE and acquired the TATA-box and Ohler1 elements. At the same time, a sharp increase in the number of reads is also associated with the position of the Ohler7 element. The beginning of the area of continuous coverage with reads in *D. suzukii* is shifted upstream on 100 bp from ECM4, in accordance with SRA data. The distribution of TSSs in the range of 50-90 bp in this area suggests that the main promoter in *D. suzukii* belongs to the broad promoter class. It should also be noted that there is a set of TSSs located distally at a distance 250 bp upstream assuming the second promoter. In both cases, the 5′-ends of the distal reads are located to the left of the Ohler1 elements. In the second, most distant promoter, TCT and BRE^{d} elements are also found, while in the main promoter, Inr with TATA and BRE^{d} in the canonical position are additionally observed, as well as Ohler6, Ohler7, and Ohler10 located downstream and DRE positioned on the complementary strand. The predicted TSSs for this species are located deep downstream in the area of continuous coverage with reads.

The region of the putative promoter in seven species of the *obscura*, *montium*, and *ananassae* groups is shifted by about 40 bp

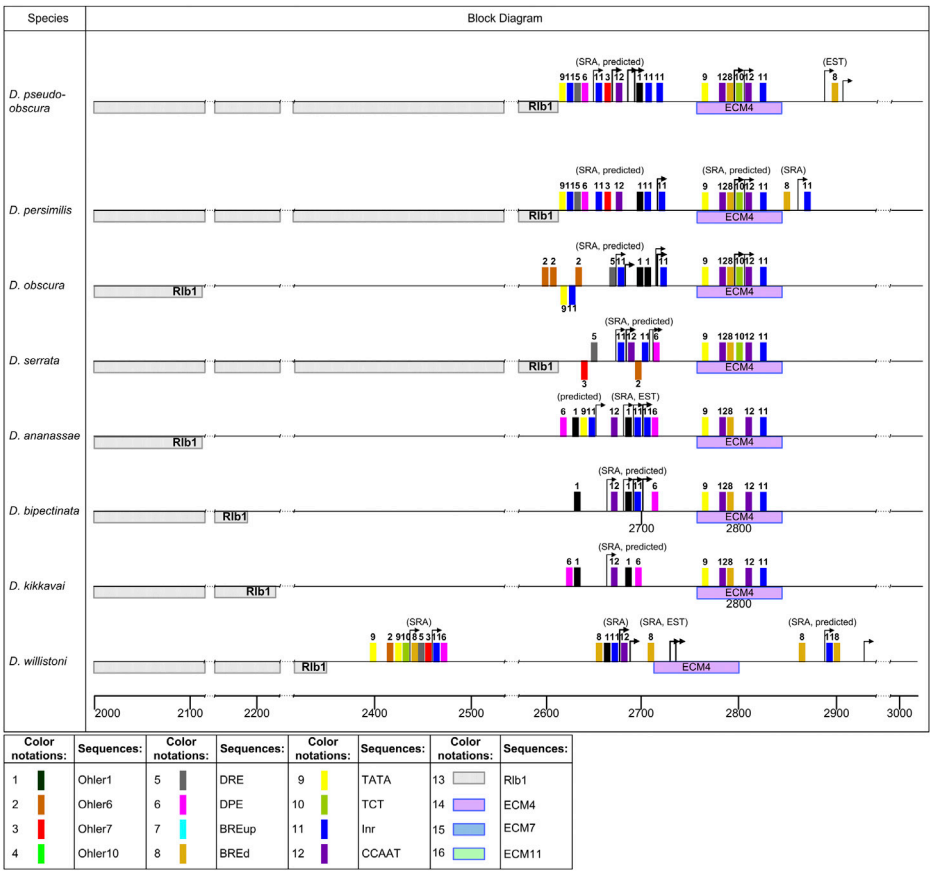


FIGURE 8 | Distribution of the elements in the promoter region in the *obscura, montium, ananassae*, *willistoni* groups. Designations as in **Figure 7**.

from ECM4, in comparison with species of the *melanogaster* group, and has a different composition of promoter elements and a divergent pattern of their reciprocal arrangement (**Figure 8**). The distribution of distal TSSs in this region also varies within 50–60 bp in the analyzed species. The promoter region in these species is characterized by the presence of CCAAT-box, TATA-box, Inr, DPE, Ohler1, and Ohler7 elements, and has a significant enrichment with the latter two elements. The composition of promoter elements in the species of the *montium* and *ananassae* groups changes insignificantly, and is represented mainly by the elements Ohler1, CCAAT-box, Inr, and DPE. An exception is the species *D. serrata*, in which the deletion of a significant part of the intergenic region reaches the promoter, and is accompanied by the appearance of the Ohler1 element on the minus strand and the DRE element on the plus strand. Degenerate sequences of elements Ohler6 and Ohler7 are noted only on the minus strand. The promoter of species in the *obscura* group has the following characteristic features: in the species *D. pseudoobscura* and *D. persimilis*, the intergenic region is almost entirely deleted; the promoter is on the border with the non-coding sequence of the *Rlb1* gene; the area of Ohler1 elements and downstream Inr elements is shifted to the right; the promoter contains Ohler7 and DRE elements. With an almost complete homology of the promoter sequences of these species, the TSS border of *D. persimilis* is shifted downstream by 70 bp, according to SRA data. The sequence of the *D. obscura* intergenic region was formed with the participation of DNA transposons, and in the left part of the promoter, the fragment from the TATA-box and Inr elements located on the plus strand to the Ohler7 and CCAAT-box elements was replaced by a fragment containing the TATA-box and Inr elements on the minus strand, and several closely spaced elements Ohler6, DRE, and Inr on the plus strand. For all the seven species, no independent transcription starts upstream of the main promoter were found.

The rearrangement of the intergenic region in *D. willistoni* penetrates deep into the intron 1 area, capturing the entire 5′-half of the 5′UTR. At the same time, the TSSs which mark the area of continuous coverage with reads are located only in one area of the sequence of 27 bp in length, which exhibits a considerable homology with the sequences of species of the *montium* and *ananassae* groups corresponding to their characteristic *ras85D* gene promoter. Downstream, this fragment is limited by a region homologous to conserved ECM4, but located on the complementary strand. The promoter of *D. willistoni* is similar to that described for these species in composition of main elements, and includes the elements BRE^{d}, Ohler1, Inr, and CCAAT-box (**Figure 8**). The structure of the element arrangement is similar to the left part of the promoter of these species, where the Ohler1 element is located distally of the Inr and

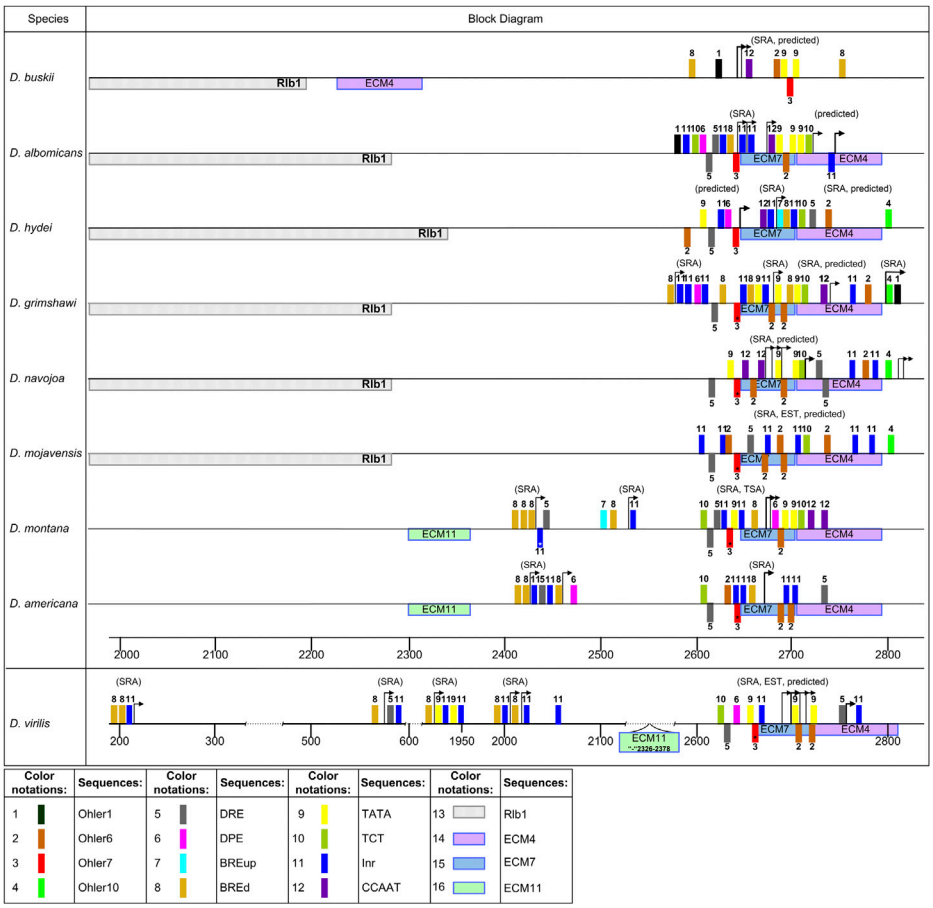


FIGURE 9 | Distribution of the elements in the promoter region in the subgenus *Drosophila*. Designations as in **Figure 7**.

CCAAT-box elements. The area of single starts in *D. willistoni* is located at a distance of 200 bp upstream of the main promoter. It corresponds to the localization of the Ohler7 element, paired with the DRE element, as well as to the canonical position of the TCT and Inr-DPE elements.

Species of the subgenus *Drosophila* exhibit a strong homology and a promoter region pattern unique to these species (**Figure 9**). It includes ECM7, a fragment of about 40 bp in length located upstream, and a half of ECM4, located downstream on the complementary strand. Distal TSS positions vary within 50 bp and are distributed mainly on the distal, part of the ECM7. The start positions marking the areas of maximum or sharply increasing coverage with reads within the discussed promoter also vary considerably, from the middle of ECM7 to the 3′-end of the inverted ECM4 or even further to the right. In this case, we can talk about the conformity of the promoter shape to a broad promoter, but there is no reason to believe that it has a peak. The composition of the promoter elements is significantly enriched with Ohler6 and Ohler7 motifs, both in terms of their occurrence relative to the rest of the intergenic region and in relation to a randomly generated sample based on the analyzed fragments (**Supplementary Table S6**). The DRE element does not display a significant enrichment relative to the sequence of the entire inergenic region, but its location at a certain section of the promoter is not accidental. The Ohler7 element in this promoter is always located before the ECM7 sequence on the minus strand, while the DRE elements are located as follows: the first one is placed before the Ohler7 element, and the second is found at the beginning of ECM4. The DRE element is a palindrome, and the sequences showing the greatest homology are present on both strands. Most TSSs show no dependence on the position of the Ohler7 element. Ohler6 is also found predominantly on the minus strand. The promoter is enriched with the core elements INR, DPE, CCAAT-box and BRE^{d}, relative to their occurrence in the intergenic region, but their non-random distribution within the promoter sequence itself is not confirmed (**Supplementary Table S6**). All the analyzed sequences also contain the TATA-box and TCT elements, but their enrichment of the promoter has not been shown. In the species of the groups *repleta*, *grimshawi*, and *immigrans* presented in the analysis, the area of the greatest coverage with reads is shifted downstream and is associated with the elements Ohler6 and Ohler10. Random start sites with the Inr and DRE elements

upstream of the promoter are noted for *D. grimshawi* and species of the *virilis* group.

The species *D. busckii* of the subgenus *Dorsilopha* does not have any noticeable homology with the rest of the species in the sequence marked with TSSs. However, the size of the 5′UTR transcripts roughly corresponds to those of the other species, and the elements Ohler1, CCAAT-box, Ohler6, and Ohler7 (on the opposite strand) are present in the start region, as is the TATA-box element (**Figure 9**). The Ohler1 element preceding the TSS in *D. busckii* is noted in this position in *D. albomicans* and in the species of the groups *ananassae*, *montium*, and *willistoni*.

Thus, the putative region of the main promoter in different *Drosophila* species was formed on the basis of sequences of different origins, which corresponds to the conclusions about the regular capture of the noncoding region of the *ras85D* gene up to intron 1 by rearrangements. We have noted at least six variants of differing promoter structures characteristic of the studied species, which are as follows: the structure noted for the species of the *melanogaster* group, the independent variant in *D. suzukii*, the structure characteristic of the *montium*, *ananassae*, and *obscura* groups, the independent structures in *D. willistoni* and *D. busckii*, as well as the general structure for the species of the subgenus *Drosophila*. Interestingly, *D. willistoni* has a 16 bp section of the promoter region which shows a high homology with the corresponding promoter sequences of the *montium*, *ananassae*, and *obscura* groups, against the background of the surrounding sequence associated with an inversion rearrangement under the probable influence of the *Gypsy* transposon and extending deep into the 5′UTR region. Of interest is also the position of ECM4, which accompanies the main promoter in most species, associated in some cases with a significant increase in expression and located in the species of the subgenus *Drosophila* on the opposite strand as part of the surrounding unrelated sequence.

The composition of the related promoter structures present in each of the three sets of species (**Figures 7–9**) may change due to accumulated point mutations, insertions or deletions, and rearrangements bordering on the promoter, nevertheless retaining its main features, especially in the right part of the promoter. All the identified structures carry Ohler elements, most often Ohler1, Ohler6, and Ohler7, as the main, constantly present motifs, and the latter two elements are also found on the opposite strand. As noted elsewhere, the DRE element often accompanies Ohler7 (Ohler et al., 2002; Ohler, 2006; Rach et al., 2009). The presence of these elements is typical for broad promoters with low peaks (Rach et al., 2009; Kadonaga, 2012; Danino et al., 2015). There are three main patterns with independent origins in most cases. They are located at a similar distance from the coding sequence and have entirely different structures. Altogether, it implies the possibility of rapid formation (or selection based on the available elements) of an active promoter in an evolutionarily new sequence after the events of insertions of mobile elements or rearrangements.

The Functional Significance of Evolutionarily Conserved Motifs

According to modENCODE database data obtained using high throughput ChIP-chip and ChIP-seq methods, the *ras85D* gene area of *D. melanogaster* the contains binding sites for more than 40 transcription factors (TFs) and their transcription complex partners from 20 TF superfamilies and four families of coactivator and corepressor proteins (modENCODE Consortium and Roy, 2010; Negre et al., 2011). This set was used as a filter to confirm the specificity of the identified transcription factor binding sites has shown factors. It turned out that only 23 transcription factors were non-randomly distributed over the ECM sequences (**Table 3**). Transcription factor *lola* and kni binding sites are found most frequently and have been identified in the composition of six and four ECMs, respectively. Enrichment with transcription factor D, disco, Dll, en, Med, pnr, and Ubx binding sites has been detected for three ECMs and enrichment with transcription factor bab1, BEAF-32, sens, Trl, and twi binding sites has been detected for two ECMs.

The identified set of TFs has a wide range of functions, from the organization of chromatin and the co-activation and co-repression of transcription complexes, to the control of cell division and cell differentiation processes, cell death and segmentation, as well as the morphogenesis of organs and body systems. The control of the *ras85D* gene expression activity is thus directly related to ontogenetic processes during embryogenesis and the development of the larva and pupa of *Drosophila.* Evolutionary conservatism assumes the support of ECM composition by selection, which determines the set of mandatory regulatory functions. To reveal the specific functions for each of the studied ECMs, we have undertaken a GO enrichment analysis of the biological processes characterizing the sets of specific TFs obtained for these ECMs (**Supplementary Table S7**).

The biological processes revealed as a result of GO enrichment analysis can be divided into two groups: those that determine the general gene expression activity, and those that depend on the tissue, on the type and state of cells, and on the stage of development. For the analysis, GO processes were ranked according to the adjusted *p*-values; then the first 10 GO-processes associated with tissue and age specificity, as well as all general GO processes falling within the boundaries of the obtained *p*-values were selected. The sets of analyzed genes are represented exclusively by transcription factors and chromatin proteins; therefore, all processes characterizing the formal connection with promoters, DNA, and transcription were removed from the general GO processes. Positive regulation of transcription is characteristic of the TF sets obtained for ECMs 2A/1B and 3A/3B/12A (**Table 1**), while the ECM1A/2B is not related to an increase or decrease in the basic gene expression. For the remaining ECMs 4A/4B/13B, 5A/5B, 7A/8B/9B, 4A/7B, 13A/11B, both directions of changes in the basic gene expression are shown.

GO processes associated with tissue and cell specificity show functional differences in ECMs. Thus, the effect of the ECM2A/1B sequence on the expression of the *ras85D* gene mediates its

TABLE 3 | Enrichment of evolutionarily conserved sequences with transcription factor binding sites (MEME Suite 5.4.1, AME algorithm).

Motif/control	2A/1B	1A/2B	3A/ 3B/12A	4A/4B/13B	5A/5B	7A/ 8B/9B	4A/7B	13A/ 11B
P-val.[1,2] < 0.01; E-val.[1,2] < 0.05 P -val.[1] < 0.01; E -val.[1] < 0.05	*D, kni, lola, Trl, disco Med*	*Dll, en, Ubx, bab1, fru*	*Trl, twi, Med h, lov, ato/da*	*BEAF-32, kni, pnr, ttk, Ubx, lola, sens, Dll, en pho*	*lola, BEAF-32, disco, dl, pnr, kni*	*kni, lola, sc/da, twi*	*Ubx, bab1, disco, lola, Med, sens, D, Dll, en*	*D, lola, pnr*

P, E-val.[1, 2]: the corrected p*-value of the probability of random enrichment with transcription factor binding sites and the expected number of false positives obtained using two variants of control sequence sets.*

[1]Obtained by cutting fragments of similar length from the intergenic region, the 5′UTR, and intron 1, except for fragments of the analyzed motif.

[2]Generated by random permutations based on the analyzed conserved set of fragments. P, E-val.[1]: indicators p and E, significant only for assessments using the control sequences obtained by method (1). The letter next to the ECM number designates the sequence set that was analyzed. A: Set A (23 species), B: Set B (37 species).

participation in the control of cell division, the morphogenesis of salivary glands and tracheal system, retinal innervation and embryo segmentation, as well as cellular responses to BMP and ecdysone stimuli. ECM1A/2B is required to control the development of the leg, wing, ocular, and genital discs and the anteroposterior cell polarization. ECM3A/3B/12A mediate the involvement of the gene in dorso-ventral cell polarization, the development of the nervous system, the differentiation of neurons and photoreceptors, and cuticle formation; ECM5A/5B mediates the participation of the gene in embryonic segmentation and the development of the heart, circulatory system, lymph gland, and peripheral nervous system; ECM7A/8B/9B are needed for the regulation of the mitotic cycle, the formation of neuroblast and stem cell differentiation, the differentiation of chaetes and the development of the peripheral nervous system and Malpighian tubules; ECM4A/4B/13B bring about the role of the gene in the mediated regulation of macromolecule metabolism, the development of genitals, antennae, halteres, wings, and legs, as well as photoreceptor differentiation. ECM4A/7B is similar in functional activity to ECM3A/3B/12A and ECM4A/4B/13B, while ECM13A/11B is similar to ECM5A/5B.

It can be noted that ECM7A and ECM13A, both in terms of the probable composition of transcription factor binding sites and in regard to the participation in GO processes, largely duplicate the regulatory functions of the other motifs that are mandatory in the non-coding sequences of the *ras85D* gene and retain their order and position on the strand. ECM4 differs in that it possesses a set of unique regulatory functions for the given cluster of conserved motifs, accompanies the downstream promoter region in most species, retains its localization in relation to the mandatory motifs, but may be found in an inverted position.

DISCUSSION

The presented picture of the divergence of the structure of the intergenic region and 5′UTR of the *ras85D* gene in *Drosophila* suggests several important evolutionary conclusions. They concern the rate and possible mechanisms of restoration of the functional activity of the gene after replacement of the promoter and a significant part of the regulatory sequences, as well as the localization of the regulatory sequences of the *ras85D* gene in the intergenic region and, to a greater extent, 5′UTR.

The first and most unexpected conclusion is the ability of the gene to quickly and completely change a significant part of the regulatory sequence, including the promoter region. The surprise of this conclusion lies in the fact that random insertions and rearrangements that change the promoter region of the gene inevitably disrupt its normal expression. *D. melanogaster* lines carrying various *ras85D* mutations in the promoter region (Bellen et al., 1989; Parks et al., 2004 (in **Supplementary Table S2**)) exist either in a heterozygous state, maintaining a lethal allele on the balancer, or have morphological and physiological abnormalities incompatible with effective survival and reproduction in nature. In this case the time for the recovery of normal or at least minimally sufficient expression of the mutant allele is extremely limited: such alleles are present in the population with a low frequency, and the probability of their complete displacement is inversely proportional to the frequency. The restoration of the functional activity of a gene should occur almost instantly on the scale of evolutionary time, within several tens of generations.

Nevertheless, in the sequences of different species, we see various combinations of fragments of independent origin, which testify to the successful reorganization of the regulatory region, its fixation in the population, and further inheritance by descendant species. Independent rearrangements of the sequence, which covered the non-coding region of the gene up to intron 1, occurred three times: in the common ancestor of the subgenus *Drosophila*, in the subgenus *Sophophora* during the divergence of the ancestral species of the *melanogaster* group from the other groups, and during the later inversion rearrangement in *D. willistoni*. In most cases, such a "new" sequence, which means a sequence of independent origin, contains a promoter which has its own characteristic pattern of elements, distinct in the composition and sequence of element distribution in the corresponding region of the sequence. The similarity of promoter structures of a common origin can be seen primarily from the key elements of a promoter of the broad type, such as Ohler and DRE, as well as CAAT-box and Inr elements. It cannot be argued that all these elements are necessary for the effective functioning of these promoters, but the considerable similarity of patterns assumes the support of each of these variants by selection. A possible scenario for the formation of these

patterns, taking into account the short period of restoration of gene expression activity, is the selection of promoter elements from suitable motifs located at an optimal distance to the preserved downstream regulatory sequences and the coding part of the gene.

Subsequent insertions of mobile elements and the accompanying rearrangements of the intergenic region left only the basal fragments of sequences from earlier rearrangements, as, for example, in species of the subgenus *Drosophila*, or "broke up" the ancestral sequences, displacing their fragments distally (as in the species *D. elegans*, *D. suzukii*, and *D. eugracilis* from the *melanogaster* group), or else led to deletions of most of the intergenic region. Such modifications are causing a change in the position of the promoter. For example, in *D. suzukii*, the promoter shifted upstream following a similar displacement of the homology region, while in *D. eugracilis* and *D. persimilis*, the promoter moved downstream. Such events are accompanied by a complete or partial change in the composition of the promoter elements, which again suggests the possibility of a rapid selection of a new optimal combination of elements, within the boundaries of an effective control of gene expression. This assumption is supported by the presence of individual TSSs in the prepromoter region of many species, localized close to the motifs with a high homology with promoter elements, but probably not playing a significant role in the regulation of gene expression. The conclusion about the lack of functional significance of short transcripts from additional TSSs was also made by Hoskins and co-authors, based on the results of analyzing the transcriptome of *D. melanogaster* embryos by the CAGE method (Hoskins et al., 2011). Obviously, an essential condition for choosing a promoter that performs basic gene expression regulation is the accessibility of enhancers to promoter elements, which ensures a complete assembly of the transcription complex (Ohler and Wassarman, 2010; Ibrahim et al., 2018). In this case, the structure of the sequence, including the distribution of elements such as nucleosome depleted regions, insulators, etc., will provide the choice of a promoter and determine the efficiency of expression from different regions of the sequence carrying a particular set of promoter elements. At the same time, the possibility of selecting a suitable composition of promoter elements suggests a wide range of possible variants of the distribution of these elements in the promoter region, with the possibility of their synergetic interactions and an efficient assembly of the transcription complex, which was noted in a number of studies earlier (Dikstein, 2011; Mwangi et al., 2015).

The correspondence between rearrangements of the non-coding region of the gene and the conserved motifs and fragments homologous to the substituted sequence located within them, allows us to conclude that there are some mechanisms for restoring the expression activity of the *ras85D* gene. The appearance of such motifs assumes their direct copying from the replaced sequence. Two examples among the studied regulatory sequence structures directly indicate such a mechanism. One example is the presence of an inverted ECM4 immediately behind the promoter region in the species of the subgenus *Drosophila*: the position of this motif corresponds to the direct ECM4 in all the other species, but is found on the opposite strand in the sequence having an independent evolutionary origin. Another example is the homology of the promoter region in *D. willistoni* to the corresponding region of species of the *ananassae*, *montium*, and *obscura* groups. In the latter case, we are faced with a phylogenetic discordance caused by the presence of an inverted ECM4 in *D. willistoni*, as in the species of the subgenus *Drosophila*. This discordance may be caused by a false positive homology of the compared motifs. It is supported by the least significant E-values of the general ECM pattern in this species (**Figure 1**), the strong degeneracy of ECM4 itself (p-val. = 7.25e-5 compared to 7.48e-26 <p-val. <1.60 e-13 in other species), and the possibility of alignment of its right side with respect to the downstream sequences in species of the subgenus *Sophophora*. The homology of this motif in species of the subgenus *Drosophila* is highly significant (p-val. <1.60e-13); in the genomes of species of the subgenus *Drosophila*, there are no homologous motifs located outside the analyzed locus (according to BLAT Search Results Genome Browser UCSC). In the *D. melanogaster* genome, this motif is reproduced in intron 1 of the *ine* gene (UCSC Genome Browser, chr2L: 4,475,427-4,475,476) and in the *Nep4—Cpr92F* intergenic region, closer to the *Nep4* promoter (UCSC Genome Browser, chr3R: 20,660,789-20,660,838). *ine* and *Nep4* are involved in the formation of the nervous system, genitals, and excretory system, as well as muscle cell differentiation. The functional activity of these genes largely coincides with that revealed for ECM4, which confirms the Gene Ontology of biological processes indicated for it. An accidental presence of this motif in the composition of the replacement sequence seems unlikely.

The possibility of copying a fragment of the ancestor, or the sequence replaced as a result of a rearrangement, can be explained by the conversion mechanisms of sequence restoration: these mechanisms are activated during the rearrangement process. Being one of the two mechanisms of homologous recombination, the canonical mechanism of meiotic conversion occurs at the sites of double-strand breaks in the chromosome and is associated with the formation of two Holliday junctions (Chen et al., 2007; Do et al., 2014). Conversional replacements of homologous and non-homologous sequences are widespread in *Drosophila* genomes (Chen et al., 2007; Johnson-Schlitz et al., 2007; Wei and Rong, 2007; Kane et al., 2012; Velandia-Huerto et al., 2016), and a significant part of them are associated with double-strand breaks induced by the mobile elements transpositions (Lankenau, 1995; Hedges and Deininger, 2007; Robert et al., 2008; Kane et al., 2012). In the above examples, motifs homologous to ancestral ones, but surrounded by a new sequence, are located at the 3′-end of the rearrangement region, close to the point of the putative transposon insertion. A possible scenario for the formation and maintenance of such variants is associated with "successful" conversional replacement of fragments of an evolutionarily new sequence located in the place of functionally significant regions, such as enhancer blocks and promoter regions. Selection will maintain such sequences, which partially or completely restore the transcriptional activity of the gene, and eliminate lethal and partially lethal alleles.

The third conclusion is related to the analysis of the distribution of evolutionarily conserved motifs in the analyzed

sequence in different species. A strict demarcation line turns out to be the boundary of intron 1, beyond which the rearrangements that capture the intergenic region do not extend. A stable pattern of this fragment is formed by ECM1, ECM2, ECM3, and ECM5, and its variation is mainly associated with insertion-deletion polymorphism and with the multiplication of dinucleotide microsatellite repeats and poly-T tracks. A certain conservatism also characterizes the area from the promoter region to intron 1, but in this case the region includes ECM4, which has signs of a conversion transfer to a younger sequence in species of the subgenus *Drosophila*. This observation does not exclude the presence of enhancers involved in the regulation of *ras85D* gene expression in the intergenic region, but suggests that regulatory sequences that are critical for the functional activity of this gene are located downstream of the promoter.

The *ras85D* gene is conserved, its functions are just as conserved and important for the control of ontogenetic processes and the maintenance of normal vital activity of cells, tissues, and the whole organism. The presented picture of the evolution of the non-coding sequence and in particular the prepromoter region of this gene allows us to conclude that each evolutionary lineage has undergone at least one event of replacement of a significant part of this area. It seems that the picture of the distribution of regulatory sequences is not accidental. Perhaps the answers to the questions of how common the observed distribution of conserved and variable regions and regulatory motifs is for other conserved genes, and in which taxonomic groups of living organisms, will bring us closer to understanding to what extent and in what ways the formation of the structure of regulatory sequences can be controlled.

Finally, let us note one more unexpected conclusion. Based on general ideas about the formation of adaptation mechanisms and isolating barriers in the course of evolution, conserved genes should maintain the homeostasis of an organism and serve as the foundation on which all evolutionary innovations are built. However, the results obtained suggest that periodic (but not critical) changes in the regulatory activity of a gene, during the evolution of its non-coding sequence, can have their effect on the ontogenetic process, also leading to evolutionarily significant changes.

Thus, we can conclude:

1) High variation of the prepromoter region and distal part of the 5′UTR is associated with rearrangements caused by insertions of mobile elements. These parts of the sequence are of independent evolutionary origin.
2) Motifs (ECMs) that are critical for the regulation of the processes of development and functioning of cells are located downstream of the promoter and, along with the promoter region, can be transferred to evolutionarily new sequences using the mechanisms of genetic conversion.
3) Evolutionary changes in the regulatory region of a conserved gene, leading in most cases to its inactivation or abnormalities incompatible with normal development, can be restored to an acceptable level in the shortest possible time, practically instantly on the evolutionary time scale.

DATA AVAILABILITY STATEMENT

The article is based on the analysis of publicly available data. All datasets presented in this study are included in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

AK contributed to the conceptualization. AC, SS, AM, and AK contributed to the writing, preparing the original draft. AF, ES, PP, and GB contributed to the investigation and validation. All authors contributed to the article and approved the submitted version.

FUNDING

This work was conducted under the Institute of Developmental Biology Russian Academy of Sciences Government basic research program, no. 0088-2021-0019.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.807234/full#supplementary-material

REFERENCES

Adryan, B., and Russell, S. (2012). "Genome Mapping and Genomics in Drosophila," in *Genome Mapping and Genomics in Laboratory Animals*. Editors P. Denny and C. Kole (Berlin, Heidelberg: Springer), 4, 31–86. doi:10.1007/978-3-642-31316-5_3

Ahsan, B., Saito, T. L., Hashimoto, S.-i., Muramatsu, K., Tsuda, M., Sasaki, A., et al. (2009). MachiBase: a *Drosophila melanogaster* 5'-end mRNA Transcription Database. *Nucleic Acids Res.* 37 (Database issue), D49–D53. doi:10.1093/nar/gkn694

Alonso-González, L., Domínguez, A., and Albornoz, J. (2006). Direct Determination of the Influence of Extreme Temperature on Transposition and Structural Mutation Rates of *Drosophila melanogaster* mobile Elements. *Genetica* 128, 11–19. doi:10.1007/s10709-005-2480-6

Bailey, T. L., and Elkan, C. (1994). Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: Tools for Motif Discovery and Searching. *Nucleic Acids Res.* 37, W202–W208. doi:10.1093/nar/gkp335

Bellen, H. J., O'Kane, C. J., Wilson, C., Grossniklaus, U., Pearson, R. K., Gehring, W. J., et al. (1989). P-element-mediated enhancer detection: a versatile method to study development in *Drosophila*. *Genes Dev.* 3 (9), 1288–1300. doi:10.1101/gad.3.9.1288

Chekunova, A. I., Kulikov, A. M., Mikhailovskii, S. S., Lazebny, O. E., Lazebnaya, I. V., and Mitrofanov, V. G. (2008). The Relationships Among the Species of the

Drosophila virilis Group Inferred from the Gene *Ras1* Sequences. *Russ. J. Genet.* 44, 286–294. doi:10.1134/S1022795408030071

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool. *BMC Bioinformatics* 14, 128. doi:10.1186/1471-2105-14-128

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. (2007). Gene Conversion: Mechanisms, Evolution and Human Disease. *Nat. Rev. Genet.* 8, 762–775. doi:10.1038/nrg2193

Danino, Y. M., Even, D., Ideses, D., and Juven-Gershon, T. (2015). The Core Promoter: At the Heart of Gene Expression. *Biochim. Biophys. Acta (Bba) - Gene Regul. Mech.* 1849, 1116–1131. doi:10.1016/j.bbagrm.2015.04.003

Dikstein, R. (2011). The Unexpected Traits Associated with Core Promoter Elements. *Transcription* 2, 201–206. doi:10.4161/trns.2.5.17271

Do, A. T., Brooks, J. T., Le Neveu, M. K., and LaRocque, J. R. (2014). Double-strand Break Repair Assays Determine Pathway Choice and Structure of Gene Conversion Events in *Drosophila melanogaster*. *G3 (Bethesda)* 4, 425–432. doi:10.1534/g3.113.010074

Drosophila 12 Genomes Consortium, Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., et al. (2007). Evolution of Genes and Genomes on the Drosophila Phylogeny. *Nature* 450, 7167203–7167218. doi:10.1038/nature06341

Evgen'ev, M. B., Zelentsova, H., Poluectova, H., Lyozin, G. T., Veleikodvorskaja, V., Pyatkov, K. I., et al. (2000). Mobile Elements and Chromosomal Evolution in the *Virilis* Group of *Drosophila*. *Proc. Natl. Acad. Sci.* 97, 11337–11342. doi:10.1073/pnas.210386297

Feschotte, C., and Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu. Rev. Genet.* 41, 331–368. doi:10.1146/annurev.genet.40.110405.090448

Galhardo, R. S., Hastings, P. J., and Rosenberg, S. M. (2007). Mutation as a Stress Response and the Regulation of Evolvability. *Crit. Rev. Biochem. Mol. Biol.* 42, 399–435. doi:10.1080/10409230701648502

Gasperini, R., and Gibson, G. (1999). Absence of Protein Polymorphism in the *Ras* Genes of *Drosophila melanogaster*. *J. Mol. Evol.* 49, 583–590. doi:10.1007/pl00006579

Georgiev, P., Tikhomirova, T., Yelagin, V., Belenkaya, T., Gracheva, E., Parshikov, A., et al. (1997). Insertions of Hybrid *P* Elements in the *Yellow* Gene of Drosophila Cause a Large Variety of Mutant Phenotypes. *Genetics* 146, 583–594. doi:10.1093/genetics/146.2.583

Gershenzon, N. I., Trifonov, E. N., and Ioshikhes, I. P. (2006). The Features of *Drosophila* Core Promoters Revealed by Statistical Analysis. *BMC Genomics* 7, 161. doi:10.1186/1471-2164-7-161

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: Scanning for Occurrences of a Given Motif. *Bioinformatics* 27, 1017–1018. doi:10.1093/bioinformatics/btr064

Grant, C. E., and Bailey, T. L. (2021). *XSTREME: Comprehensive Motif Analysis of Biological Sequence Datasets*. doi:10.1101/2021.09.02.458722Available at: https://www.biorxiv.org/content/10.1101/2021.09.02.458722v1 (Accessed September 3, 2021)

Guerreiro, M. P. G. (2012). What Makes Transposable Elements Move in the *Drosophila Heredity* 108, 461–468. doi:10.1038/hdy.2011.89

Haberle, V., Arnold, C. D., Pagani, M., Rath, M., Schernhuber, K., and Stark, A. (2019). Transcriptional Cofactors Display Specificity for Distinct Types of Core Promoters. *Nature* 570, 122–126. doi:10.1038/s41586-019-1210-7

Hedges, D. J., and Deininger, P. L. (2007). Inviting Instability: Transposable Elements, Double-Strand Breaks, and the Maintenance of Genome Integrity. *Mutat. Research/Fundamental Mol. Mech. Mutagenesis* 616, 46–59. doi:10.1016/j.mrfmmm.2006.11.021

Hendrix, D. A., Hong, J.-W., Zeitlinger, J., Rokhsar, D. S., and Levine, M. S. (2008). Promoter Elements Associated with RNA *Pol II* Stalling in the *Drosophila* Embryo. *Proc. Natl. Acad. Sci.* 105, 7762–7767. doi:10.1073/pnas.0802406105

Hoskins, R. A., Landolin, J. M., Brown, J. B., Sandler, J. E., Takahashi, H., Lassmann, T., et al. (2011). Genome-wide Analysis of Promoter Architecture in *Drosophila melanogaster*. *Genome Res.* 21, 182–192. doi:10.1101/gr.112466.110

Ibrahim, M. M., Karabacak, A., Glahs, A., Kolundzic, E., Hirsekorn, A., Carda, A., et al. (2018). Determinants of Promoter and Enhancer Transcription Directionality in Metazoans. *Nat. Commun.* 9, 4472. doi:10.1038/s41467-018-06962-z

Johnson-Schlitz, D. M., Flores, C., and Engels, W. R. (2007). Multiple-Pathway Analysis of Double-Strand Break Repair Mutations in Drosophila. *Plos Genet.* 3, e50. doi:10.1371/journal.pgen.0030050

Juven-Gershon, T., and Kadonaga, J. T. (2010). Regulation of Gene Expression via the Core Promoter and the Basal Transcriptional Machinery. *Developmental Biol.* 339, 225–229. doi:10.1016/j.ydbio.2009.08.009

Kadonaga, J. T. (2012). Perspectives on the RNA Polymerase II Core Promoter. *Wiley Interdiscip. Rev. Dev. Biol.* 1, 40–51. doi:10.1002/wdev.21

Kane, D. P., Shusterman, M., Rong, Y., and McVey, M. (2012). Competition between Replicative and Translesion Polymerases during Homologous Recombination Repair in Drosophila. *Plos Genet.* 8, e1002659. doi:10.1371/journal.pgen.1002659

Koh, Y.-H., Ruiz-Canada, C., Gorczyca, M., and Budnik, V. (2002). The Ras1-Mitogen-Activated Protein Kinase Signal Transduction Pathway Regulates Synaptic Plasticity through Fasciclin II-Mediated Cell Adhesion. *J. Neurosci.* 22, 2496–2504. doi:10.1523/JNEUROSCI.22-07-02496.2002

Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. (2006). Annotation, Submission and Screening of Repetitive Elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7, 474. doi:10.1186/1471-2105-7-474

Kovalenko, L. V., Zakharenko, L. P., Voloshina, M. A., Karamysheva, T. V., Rubtsov, N. B., and Zakharov, I. K. (2006). Behavior of Hobo and P Transposons in Yellow 2-717 Unstable Line of *Drosophila melanogaster* and its Derivatives after Crossing with a Laboratory Strain. *Russ. J. Genet.* 42, 605–612. doi:10.1134/S1022795406060032

Kudron, M. M., Victorsen, A., Gevirtzman, L., Hillier, L. W., Fisher, W. W., Vafeados, D., et al. (2018). The ModERN Resource: Genome-wide Binding Profiles for Hundreds of *Drosophila* and *Caenorhabditis elegans* Transcription Factors. *Genetics* 208, 937–949. doi:10.1534/genetics.117.300657

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. *Nucleic Acids Res.* 44, W90–W97. doi:10.1093/nar/gkw377

Kulikov, A. M., Lazebnyi, O. E., Gornostaev, N. G., Chekunova, A. I., and Mitrofanov, V. G. (2010). Unequal Evolutionary Rates in the *Drosophila virilis* Species Group: I. The Use of Phylogeny-Based Takezaki's Tests. *Biol. Bull. Russ. Acad. Sci.* 37, 18–25. doi:10.1134/S1062359010010036

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi:10.1093/molbev/msy096

Lankenau, D.-H. (1995). Genetics of Genetics in Drosophila: *P* Elements Serving the Study of Homologous Recombination, Gene Conversion and Targeting. *Chromosoma* 103, 659–668. doi:10.1007/BF00344226

Lee, D.-H., Gershenzon, N., Gupta, M., Ioshikhes, I. P., Reinberg, D., and Lewis, B. A. (2005). Functional Characterization of Core Promoter Elements: the Downstream Core Element Is Recognized by TAF1. *Mol. Cel. Biol.* 25, 9674–9686. doi:10.1128/MCB.25.21.9674-9686.2005

Main, B. J., Smith, A. D., Jang, H., and Nuzhdin, S. V. (2013). Transcription Start Site Evolution in *Drosophila*. *Mol. Biol. Evol.* 30, 1966–1974. doi:10.1093/molbev/mst085

Malkova, A., and Haber, J. E. (2012). Mutations Arising during Repair of Chromosome Breaks. *Annu. Rev. Genet.* 46, 455–473. doi:10.1146/annurev-genet-110711-155547

McLeay, R. C., and Bailey, T. L. (2010). Motif Enrichment Analysis: a Unified Framework and an Evaluation on ChIP Data. *BMC Bioinformatics* 11, 165. doi:10.1186/1471-2105-11-165

Mitrofanov, V. G., Chekunova, A. I., Zelentsova, E. S., Gauze, L. N., and Bakhtoiarov, G. N. (2011). [Localization of the Dras1 Sequence to Polytene Chromosomes in Sibling Species of the *Drosophila virilis* Group]. *Tsitologiia* 53, 513–516. (in Russian).

modENCODE ConsortiumRoy, S., Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., et al. (2010). Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science* 330, 60121787–60121797. doi:10.1126/science.1198374

Morales-Hojas, R., Reis, M., Vieira, C. P., and Vieira, J. (2011). Resolving the Phylogenetic Relationships and Evolutionary History of the *Drosophila virilis* Group Using Multilocus Data. *Mol. Phylogenet. Evol.* 60, 249–258. doi:10.1016/j.ympev.2011.04.022

Mwangi, S., Attardo, G., Suzuki, Y., Aksoy, S., and Christoffels, A. (2015). TSS Seq Based Core Promoter Architecture in Blood Feeding Tsetse Fly (*Glossina Morsitans Morsitans*) Vector of Trypanosomiasis. *BMC Genomics* 16, 722. doi:10.1186/s12864-015-1921-6

Nègre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., et al. (2011). A Cis-Regulatory Map of the *Drosophila* Genome. *Nature* 471, 527–531. doi:10.1038/nature09990

Ni, T., Corcoran, D. L., Rach, E. A., Song, S., Spana, E. P., Gao, Y., et al. (2010). A Paired-End Sequencing Strategy to Map the Complex Landscape of Transcription Initiation. *Nat. Methods* 7, 521–527. doi:10.1038/nmeth.1464

Noe, L., and Kucherov, G. (2005). YASS: Enhancing the Sensitivity of DNA Similarity Search. *Nucleic Acids Res.* 33, W540–W543. doi:10.1093/nar/gki478

Ohler, U. (2006). Identification of Core Promoter Modules in *Drosophila* and Their Application in Accurate Transcription Start Site Prediction. *Nucleic Acids Res.* 34, 5943–5950. doi:10.1093/nar/gkl608

Ohler, U., Liao, G.-C., Niemann, H., and Rubin, G. M. (2002). Computational Analysis of Core Promoters in the *Drosophila* Genome. *Genome Biol.* 3, 1. doi:10.1186/gb-2002-3-12-research0087

Ohler, U., and Wassarman, D. A. (2010). Promoting Developmental Transcription. *Development* 137, 15–26. doi:10.1242/dev.035493

Parks, A. L., Cook, K. R., Belvin, M., Dompe, N. A., Fawcett, R., and Huppert, K. (2004). Systematic Generation of High-Resolution Deletion Coverage of the Drosophila Melanogaster Genome. *Nat. Genet.* 36, 288–292. doi:10.1038/ng1312

Pomerantseva, E., Biryukova, I., Silicheva, R., Savitskaya, E., Golovnin, A., and Georgiev, P. (2006). Transposition of Regulatory Elements by *P*-Element-Mediated Rearrangements in *Drosophila melanogaster*. *Genetics* 172, 2283–2291. doi:10.1534/genetics.105.052803

Prober, D. A., and Edgar, B. A. (2000). Ras1 Promotes Cellular Growth in the *Drosophila* wing. *Cell* 100, 435–446. doi:10.1016/s0092-8674(00)80679-0

Rach, E. A., Winter, D. R., Benjamin, A. M., Corcoran, D. L., Ni, T., Zhu, J., et al. (2011). Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level. *Plos Genet.* 7, e1001274. doi:10.1371/journal.pgen.1001274

Rach, E. A., Yuan, H.-Y., Majoros, W. H., Tomancak, P., and Ohler, U. (2009). Motif Composition, Conservation and Condition-Specificity of Single and Alternative Transcription Start Sites in the *Drosophila* Genome. *Genome Biol.* 10, R73. doi:10.1186/gb-2009-10-7-r73

Reis, M., Vieira, C. P., Lata, R., Posnien, N., and Vieira, J. (2018). Origin and Consequences of Chromosomal Inversions in thevirilisGroup of Drosophila. *Genome Biol. Evol.* 10, 3152–3166. doi:10.1093/gbe/evy239

Robert, V. J., Davis, M. W., Jorgensen, E. M., and Bessereau, J.-L. (2008). Gene Conversion and End-Joining-Repair Double-Strand Breaks in the *Caenorhabditis elegans* Germline. *Genetics* 180, 673–679. doi:10.1534/genetics.108.089698

Rojas, A. M., Fuentes, G., Rausell, A., and Valencia, A. (2012). The Ras Protein Superfamily: Evolutionary Tree and Role of Conserved Amino Acids. *Cell Biol* 196, 189–201. doi:10.1083/jcb.201103008

Rommel, C., and Hafen, E. (1998). Ras - a Versatile Cellular Switch. *Curr. Opin. Genet. Development* 8, 412–418. doi:10.1016/S0959-437X(98)80111-1

Rose, R., Golosova, O., Sukhomlinov, D., Tiunov, A., and Prosperi, M. (2019). Flexible Design of Multiple Metagenomics Classification Pipelines with UGENE. *Bioinformatics* 35, 1963–1965. doi:10.1093/bioinformatics/bty901

Ross, C. L., Dyer, K. A., Erez, T., Miller, S. J., Jaenike, J., and Markow, T. A. (2003). Rapid Divergence of Microsatellite Abundance Among Species of *Drosophila*. *Mol. Biol. Evol.* 20, 1143–1157. doi:10.1093/molbev/msg137

Rozas, J., Segarra, C., Ribó, G., and Aguadé, M. (1999). Molecular Population Genetics of the *Rp49* Gene Region in Different Chromosomal Inversions of *Drosophila subobscura*. *Genetics* 151, 189–202. doi:10.1093/genetics/151.1.189

Sánchez-Gracia, A., and Rozas, J. (2011). Molecular Population Genetics of the OBP83 Genomic Region in *Drosophila subobscura* and *D. Guanche*: Contrasting the Effects of Natural Selection and Gene Arrangement Expansion in the Patterns of Nucleotide Variation. *Heredity* 106, 191–201. doi:10.1038/hdy.2010.26

Schnorr, J. D., and Berg, C. A. (1996). Differential Activity of *Ras1* during Patterning of the *drosophila* Dorsoventral axis. *Genetics* 144, 1545–1557. doi:10.1093/genetics/144.4.1545

Slack, C., Alic, N., Foley, A., Cabecinha, M., Hoddinott, M. P., and Partridge, L. (2015). The Ras-Erk-ETS-Signaling Pathway Is a Drug Target for Longevity. *Cell* 162, 72–83. doi:10.1016/j.cell.2015.06.023

Spradling, A. C., Stern, D., Beaton, A., Rhem, E. J., Laverty, T., Mozden, N., et al. (1999). The Berkeley Drosophila Genome Project Gene Disruption Project: Single P-Element Insertions Mutating 25% of Vital Drosophila Genes. *Genetics* 153, 135–177. doi:10.1093/genetics/153.1.135

Stormo, G. D. (2013). Modeling the Specificity of Protein-DNA Interactions. *Quant Biol.* 1, 115–130. doi:10.1007/s40484-013-0012-4

Suvorov, A., Kim, B. Y., Wang, J., Armstrong, E. E., Peede, D., D'Agostino, E. R. R., et al. (2021). Widespread Introgression across a Phylogeny of 155 Drosophila Genomes. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0960982221014962?via%3Dihub. (Accessed November 16, 2021). doi:10.1016/j.cub.2021.10.052

Tamura, K., and Nei, M. (1993). Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees. *Mol. Biol. Evol.* 10, 512–526. doi:10.1093/oxfordjournals.molbev.a040023

Tamura, K., Battistuzzi, F. U., Billing-Ross, P., Murillo, O., Filipski, A., and Kumar, S. (2012). Estimating Divergence Times in Large Molecular Phylogenies. *Proc. Natl. Acad. Sci.* 109, 19333–19338. doi:10.1073/pnas.1213199109

Tamura, K. (1992). Estimation of the Number of Nucleotide Substitutions when There Are strong Transition-Transversion and G+C-content Biases. *Mol. Biol. Evol.* 9, 678–687. doi:10.1093/oxfordjournals.molbev.a040752

Tamura, K., Tao, Q., and Kumar, S. (2018). Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. *Mol. Biol. Evol.* 35, 1770–1782. doi:10.1093/molbev/msy044

Theisen, J. W. M., Lim, C. Y., and Kadonaga, J. T. (2010). Three Key Subregions Contribute to the Function of the Downstream RNA Polymerase II Core Promoter. *Mol. Cel. Biol.* 30, 3471–3479. doi:10.1128/MCB.00053-10

Valencia, A., Chardin, P., Wittinghofer, A., and Sander, C. (1991). The Ras Protein Family: Evolutionary Tree and Role of Conserved Amino Acids. *Biochemistry* 30, 4637–4648. doi:10.1021/bi00233a001

Velandia-Huerto, C. A., Berkemer, S. J., Hoffmann, A., Retzlaff, N., Romero Marroquín, L. C., Hernández-Rosales, M., et al. (2016). Orthologs, Turn-Over, and Remolding of tRNAs in Primates and Fruit Flies. *BMC Genomics* 17, 617. doi:10.1186/s12864-016-2927-4

Vo Ngoc, L., Kassavetis, G. A., and Kadonaga, J. T. (2019). The RNA Polymerase II Core Promoter in *Drosophila*. *Genetics* 212, 13–24. doi:10.1534/genetics.119.302021

Wallace, A. G., Detweiler, D., and Schaeffer, S. W. (2013). Molecular Population Genetics of Inversion Breakpoint Regions in *Drosophila pseudoobscura*. *G3 (Bethesda)* 3, 1151–1163. doi:10.1534/g3.113.006122

Wasserman, W. W., and Sandelin, A. (2004). Applied Bioinformatics for the Identification of Regulatory Elements. *Nat. Rev. Genet.* 5, 276–287. doi:10.1038/nrg1315

Wei, D. S., and Rong, Y. S. (2007). A Genetic Screen for DNA Double-Strand Break Repair Mutations in *Drosophila*. *Genetics* 177, 63–77. doi:10.1534/genetics.107.077693

Wong, L. H., and Choo, K. H. A. (2004). Evolutionary Dynamics of Transposable Elements at the Centromere. *Trends Genet.* 20, 611–616. doi:10.1016/j.tig.2004.09.011

Wu, Y., Zhuang, Y., Han, M., Xu, T., and Deng, K. (2009). Ras Promotes Cell Survival by Antagonizing Both JNK and Hid Signals in the Drosophilaeye. *BMC Dev. Biol.* 9, 53. doi:10.1186/1471-213X-9-53

Zelentsova, H., Poluectova, H., Mnjoian, L., Lyozin, G., Veleikodvorskaja, V., Zhivotovsky, L., et al. (1999). Distribution and Evolution of mobile Elements in the *Virilis* Species Group of *Drosophila*. *Chromosoma* 108, 443–456. doi:10.1007/s004120050396

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Chekunova, Sorokina, Sivoplyas, Bakhtoyarov, Proshakov, Fokin, Melnikov and Kulikov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 January 2022
doi: 10.3389/fgene.2021.804222

Non-Invasive microRNA Profiling in Saliva can Serve as a Biomarker of Alcohol Exposure and Its Effects in Humans

Edward A. Mead[1,2]*, *Nadia Boulghassoul-Pietrzykowska*[1,3,4,5], *Yongping Wang*[1,6], *Onaiza Anees*[1,7], *Noah S. Kinstlinger*[1,8], *Maximillian Lee*[1,9], *Shireen Hamza*[1,10], *Yaping Feng*[11,12] and *Andrzej Z. Pietrzykowski*[1,5]

[1]Laboratory of Adaptation, Reward and Addiction, Department of Animal Sciences, Rutgers University, New Brunswick, NJ, United States, [2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [3]Mayo Clinic Health System, NWWI, Barron, WI, United States, [4]Department of Medicine, Capital Health, Trenton, NJ, United States, [5]Weight and Life MD, Hamilton, NJ, United States, [6]Holmdel Township School, Holmdel, NJ, United States, [7]Virginia Commonwealth University Health, CMH Behavioral Health, South Hill, VA, United States, [8]Albert Einstein College of Medicine, Bronx, NY, United States, [9]George Washington University, School of Medicine and Health Sciences, Washington DC, MA, United States, [10]Department of the History of Science, Harvard University, Cambridge, MA, United States, [11]Waksman Genomics Core Facility, Rutgers University, Piscataway, NJ, United States, [12]Bioinformatics Department, Admera Health, South Plainfield, NJ, United States

OPEN ACCESS

Edited by:
William C. Cho,
QEH, Hong Kong SAR, China

Reviewed by:
Kedar N. Prasad,
Engage Global, United States
Gordon Proctor,
King's College London,
United Kingdom

***Correspondence:**
Edward A. Mead
edward.mead@mssm.edu

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: *29 October 2021*
Accepted: *13 December 2021*
Published: *20 January 2022*

Citation:
Mead EA, Boulghassoul-Pietrzykowska N, Wang Y, Anees O, Kinstlinger NS, Lee M, Hamza S, Feng Y and Pietrzykowski AZ (2022) Non-Invasive microRNA Profiling in Saliva can Serve as a Biomarker of Alcohol Exposure and Its Effects in Humans. Front. Genet. 12:804222. doi: 10.3389/fgene.2021.804222

Alcohol Use Disorder (AUD) is one of the most prevalent mental disorders worldwide. Considering the widespread occurrence of AUD, a reliable, cheap, non-invasive biomarker of alcohol consumption is desired by healthcare providers, clinicians, researchers, public health and criminal justice officials. microRNAs could serve as such biomarkers. They are easily detectable in saliva, which can be sampled from individuals in a non-invasive manner. Moreover, microRNAs expression is dynamically regulated by environmental factors, including alcohol. Since excessive alcohol consumption is a hallmark of alcohol abuse, we have profiled microRNA expression in the saliva of chronic, heavy alcohol abusers using microRNA microarrays. We observed significant changes in salivary microRNA expression caused by excessive alcohol consumption. These changes fell into three categories: downregulated microRNAs, upregulated microRNAs, and microRNAs upregulated *de novo*. Analysis of these combinatorial changes in microRNA expression suggests dysregulation of specific biological pathways leading to impairment of the immune system and development of several types of epithelial cancer. Moreover, some of the altered microRNAs are also modulators of inflammation, suggesting their contribution to pro-inflammatory mechanisms of alcohol actions. Establishment of the cellular source of microRNAs in saliva corroborated these results. We determined that most of the microRNAs in saliva come from two types of cells: leukocytes involved in immune responses and inflammation, and buccal cells, involved in development of epithelial, oral cancers. In summary, we propose that microRNA profiling in saliva can be a useful, non-invasive biomarker allowing the monitoring of alcohol abuse, as well as alcohol-related inflammation and early detection of cancer.

Keywords: abuse, alcohol, array, microRNA, profiling, saliva, biomarker

INTRODUCTION

Alcohol is the oldest addictive substance. Currently, Alcohol Use Disorder (AUD) is among the most prevalent mental disorders in the world (Grant et al., 2007). In the United States, excessive alcohol consumption is the third leading cause of preventable death (NIH-NIAAA, 2021a). Over ninety-five thousand people die every year from the consequences of alcohol consumption, which includes alcohol-related illnesses and accidents (Centers for Disease Control and Prevention, 2020). Moreover, with 14% of alcohol users suffering from alcohol dependence and over $249 billion of estimated alcohol-related costs in 2010 alone, AUD puts a substantial economic burden on the society (National Institutes of Health, 2021a). Excessive alcohol consumption affects both men and women of various ages and diverse ethnic groups (National Institutes of Health, 2021a). The current fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) classifies AUD based on its severity into several categories under the label of alcohol use disorder: mild, moderate and severe (Gruenberg, 2013; National Institutes of Health, 2021b). A key diagnostic and pathogenic criterion is an exposure to alcohol. While the labeling has changed in development of DSM-5, essentially the same core criteria exists: patients suffering from AUD drink often and in excess, despite negative consequences caused or exacerbated by drinking (National Institutes of Health, 2021b). Nationwide statistics indicate that, as of 2019, a staggering 14.5 million Americans 12 years old and up were consuming alcohol in a pathological manner (National Institutes of Health, 2021a).

Considering the widespread occurrence of AUD, a reliable biomarker of pathological alcohol consumption would be useful to healthcare providers, clinicians, researchers, as well as public safety workers or those in the criminal justice system (Litten et al., 2010; Nanau and Neuman, 2015). This biomarker should also be inexpensive and easy to use at varying socioeconomic levels. Ability to perform frequent sampling of the same individuals would permit monitoring of progress of the disease with its many diverse features like escalation of drinking, development of tolerance, abstinence or relapse, from mild to severe subtypes of AUD (National Institutes of Health, 2021b). Several compounds have been considered, or are currently used, as alcohol biomarkers (see reviews by Litten et al., 2010; Nanau and Neuman, 2015). Traditional and new biomarkers are mainly alcohol metabolites, or direct or indirect indicators of abnormal enzymatic activity. Carbohydrate-deficient transferrin (CDT) and ethyl glucuronide (EtG) are two particularly promising biomarkers due to their high sensitivity and specificity. However, their measurements still require costly methodology: high performance liquid chromatography (HPLC) to determine CDT, and liquid chromatography mass spectroscopy (LC-MS) to measure EtG (Litten et al., 2010; Nanau and Neuman, 2015).

microRNAs are intracellular molecules, belonging to a class of small, non-coding RNAs, (Stickel and Dubuquoy, 2016). microRNA binds to its mRNA target causing either mRNA degradation or impairment of mRNA translation (Lim et al., 2005; Peterson et al., 2014). Since a single microRNA can bind to several mRNA transcripts, one microRNA can simultaneously regulate expression of hundreds of mRNA targets. Moreover, efficiency of microRNA binding can vary from target to target, allowing microRNA to regulate gene expression with great finesse. microRNAs are estimated to modify expression of over 60% of the human transcriptome (Shu et al., 2017) making microRNAs very powerful post-transcriptional regulators of overall gene expression.

Because many microRNAs are differentially expressed in pathological cells compared to normal cells, attempts are being made to use microRNAs as biomarkers in several diseases, *e.g.*, in neurological disorders (Shafi et al., 2010; Wang et al., 2021) cancer (Vu et al., 2021), infections (Wang et al., 2016), aging (Kumar et al., 2017) and forensics (Silva et al., 2015). Interestingly, microRNAs emerge as attractive biomarkers of cancer as microRNA profiling seems to more accurately cluster different types of cancer rather than mRNA (Lu et al., 2005). (Gilad et al., 2008; Yang et al., 2018). It is tempting to consider microRNAs as an attractive alternative to traditional markers of alcohol consumption in AUD (Rosato et al., 2019; Lim et al., 2021).

Alcohol has a global effect on the body and its regulation of microRNAs has been shown in different organs and cell types like the liver (Bala and Szabo, 2012; Dippold et al., 2013), gastrointestinal tissue (Tang et al., 2008; Natarajan et al., 2015), the brain (Lewohl et al., 2011; Natarajan et al., 2015) and others (Natarajan et al., 2015).

De-regulation of leukocytes by alcohol can contribute to the immunosuppressive and inflammatory effects of alcohol. The association between the development of cancer and inflammatory processes has been well appreciated (Coussens and Werb, 2002; Greten and Grivennikov, 2019). Moreover, carcinogenic effects of alcohol, particularly chronic alcohol exposure, are also thought to take place, at least in part, *via* modulation of inflammatory responses (Seitz and Stickel, 2007; Ureña-Peralta et al., 2018). Specifically, cancers originating from the epithelial cells seem to be affected by alcohol (Vagst et al., 2003; Gupta et al., 2010; Michaud et al., 2010; Fedirko et al., 2011; Johnson et al., 2011; Pelucchi et al., 2011; Meadows and Zhang, 2015).

Since research has shown that microRNAs are present in saliva (Park et al., 2009; Wong, 2015; Rosato et al., 2019; Romani et al., 2021) and it is a minimally invasive, non-biohazardous biofluid to collect, saliva can be an attractive source of biomarkers for AUD, as well as related disorders including inflammation and cancer.

The aim of this study was to establish whether alcohol affects microRNA expression in saliva, and to demonstrate feasibility of using microRNA profiling in saliva as a biomarker of chronic, excessive alcohol consumption. We also aimed to determine potential functional consequences of microRNA expression altered by alcohol, providing further insight into pathogenesis of AUD.

MATERIALS AND METHODS

Subject Selection

All patients selected to this study were admitted to Capital Health Hospital, Fuld Campus, Trenton, NJ due to underlying medical

condition requiring hospitalization (e.g., pneumonia, acute kidney injury, chronic kidney disease secondary to hypertension, secondary diabetes mellitus, chest pain, ischemic heart disease, congestive heart failure or cardiac arrhythmia). All patients were adults over 21 years old, enrolled without restrictions based upon race, ethnicity or gender, living locally (within approximately 50 miles from the hospital). Since the Capital Health Hospital, Fuld Campus covers a large Polish community, in addition to the English consent form, a certified Polish consent form was also available, as well as a Polish-speaking provider. Thus, inclusion criteria included individuals able to fully communicate in English and/or Polish.

Individuals were selected based on their alcohol drinking pattern, which was determined based on a number of drinking bouts per week, a number of drinks per bout and an overall drinking period. Two groups of individuals were considered in this study: 1/controls, which drank alcohol sporadically and in low amounts with an average alcohol consumption: 16.4 gm EtOH/24 h (non-abusers), and 2/heavy drinkers with at least a ten year history of daily alcohol consumption, with current alcohol intake of 230 gm EtOH/24 h on average (abusers). A standard drink was defined as 14 g of ethanol, which is equivalent to either twelve ounces of beer, five ounces of wine or one ounce of distilled spirits (Ferreira and Willoughby, 2008). Controls consumed alcohol sporadically: not more often than once a week and not more than one drink. In order to identify suitable subjects, 1351 medical records were reviewed, from which 330 subjects were selected for an interview, which focused on determination of the alcohol drinking pattern based on DSM-IV. 202 patients were selected and presented with a thorough explanation of the study and the written consent form. 37 subjects elected to participate in the study.

Exclusion criteria included: minors (<21 years of age), patients with intellectual disability, or conditions not allowing for proper communication (*e.g.*, aphasia), non-English/non-Polish speaking individuals, and patients with known drug abuse history, including nicotine addiction (smokers). Non-smokers were permitted to participate in the study. This study (including a bilingual consent form) was approved by two Ethical Review Board Committees (Capital Health Hospital System, and Rutgers State University of New Jersey).

Sample Collection

Saliva was collected using the Oragene RNA collection kits (DNA Genotek Inc., Canada) from all patients enrolled in this study during their hospitalization no earlier than 48 h after hospitalization. Typically, saliva samples were collected on the day of discharge, when any active acute sickness for which patients were admitted was resolved, and consent forms were explained and accepted when patients were fully sober. All saliva samples were self-collected by patients under the supervision of a physician following the manufacturer's protocol. Patients did not eat or drink for an hour prior to saliva collection. Two (2) ml of saliva were collected directly into collection tubes without any stimulation of saliva secretion or scraping of cheeks. Tubes were capped immediately and mixed thoroughly with the self-releasing container cap stabilization solution, which also neutralizes SARS-CoV-2 virus responsible for COVID-19 (DNA Genotek, 2021).

RNA Isolation

For each sample, total RNA was isolated using mirVana miRNA isolation kit following manufacturer guidelines (Life Technologies, NY) with minor modifications. RNA was stored immediately at −80°C to preserve its integrity until further analysis. Briefly, 500 µL aliquot of each saliva sample from the last step was incubated at 50°C overnight and the next day the aliquot was incubated at 90°C for 15 min miRNA homogenate additive, in volumes of 1/10, was then added in each aliquot. An equal volume of acid-phenol:chloroform was added to each aliquot and mixed well. The mixtures were spun for 10 min at 10,000 × g. Next, 1.25 volumes of 100% ethanol were added to the aqueous phase. The mixture was passed through a mirVana column in sequential 700-µL aliquots. Several washing steps were carried out. Finally, RNA was recovered in 50 µL of elution buffer. The eluted RNAs were treated with DNase I to remove DNA from RNA (Life Technologies, NY). The RNA concentration was quantified using a NanoDrop 1000 (NanoDrop Technologies, DE). RNA integrity and quality control were additionally measured for all samples on an Agilent 2100 Bioanalyzer (Agilent Technologies, CA).

microRNA Expression Profiling Using TaqMan Low-Density Array

Tested subjects were not pooled, but rather an individual array was used for each subject. Thus, thirty-seven stem-loop RT-PCR based TaqMan Human microRNA Taqman low-density arrays were used (v3.0, Applied Biosystems, Foster City, CA). Each array consisted of two cards (Card A and B) with 384 TaqMan MicroRNA Assays per card. Arrays were designed based on a set of human-specific, mature microRNAs present in the Sanger miRBase v14 (http:/www.mirbase.org/). In addition, each card contains four control microRNAs including three carefully selected candidate endogenous controls for data normalization, inter- and intra-array control as well as a negative control (a Drosophila-specific microRNA).

After isolation of total RNA from saliva, all mature miRNAs were reverse transcribed into complementary DNA (cDNA) using looped-primer RT-PCR according to the manufacturer's instructions. All reagents were obtained from Applied Biosystems. Three µl of total RNA were reverse transcribed using the TaqMan miRNA RT kit in combination with Megaplex RT primers in a total volume of 7.5 µL. The cDNA templates were pre-amplified using Megaplex Pre-Amp Primers and TaqMan Pre-Amp Master Mix. Twenty-seven µl of the 1:4 diluted preamplified product were amplified using sequence-specific primers and probes on the Taqman MicroRNA Array. PCR was performed using the 7900HT Fast Real-Time PCR (Applied Biosystems, CA) or Viia 7 Real-Time PCR thermocycler systems (Applied Biosystems, CA) following manufacturer's recommended cycling conditions. Since each array results correspond to a separate subject, they were analyzed individually using the Applied Biosystems generic

ExpressionSuite Software (v1.0.2) with standard settings containing FDR correction. Cycle threshold (C_t) values were calculated using automatic baseline settings and a threshold of 0.2. GeNorm (Vandesompele et al., 2002) was used initially to determine control miRNAs. MammU6 was used as an endogenous control because of its low C_t values and very low variability amongst samples. C_t values >37 were considered to be below the detection level. miRNAs present in at least 50% of the control measurements were included in the analysis. Relative miRNA expression was calculated using the ΔΔCt method. The results were expressed as a fold change, unless stated otherwise. The Mann-Whitney *U* test was performed to detect differentially expressed miRNAs between groups (alcohol abusers vs non-abusing controls) using GraphPad (v6.0a). $p < 0.05$ was set as significant.

Quantitative Real-Time PCR Analysis of Individual miRNAs

Real-Time quantitative PCR (qPCR) was performed by Taqman assay according to manufacturer's instructions for Small RNA assays (Applied Biosystems, Foster City, CA). Ten individual microRNAs (see below) across the upregulated, expression *de novo*, downregulated, and unchanged microRNA categories were analyzed with U6 as a control from individual samples taken from each of six of individuals ($n = 6$) at random who had been screened by microarray card (3 controls, 3 alcohol abusers) and individually compared against the results for the same individuals from the arrays. miRNAs examined include hsa-miR-1, -10a-5p, -182-5p, -26a-5p, -27a-3p, -20a-5p, -29c-3p, -106a-5p, -9-5p, -618. Please note that microRNAs are named according to the nomenclature of miRBase 22.1 in which mature microRNAs coming from the complementary strands of a precursor are called miR-xx-5p and -3p, respectively, not miR-xx and miR-xx*.

Supplies used included Taqman Universal PCR master mix II (2x) no UNG, Taqman microRNA assays for each of the miRNAs under analysis, and 1:4 dilutions of the pre-amplification cDNA prepared for the microarray card assays. In short, a tube containing enough Universal Master Mix and nuclease-free water to provide 10 ul and 7.67 ul respectively of each for each well was prepared. This was gently mixed and aliquoted into separate tubes for each miRNA (or to U6) under analysis, to allow for 20x Taqman miRNA assay (primers + probe) to be added to each tube (enough for 1ul of final Taqman miRNA assay/well corresponding to that miRNA or U6). These final master mix (FMM) tubes were aliquoted into the appropriate wells, and for each well, 1.33 ul of a 1:4 dilution of cDNA corresponding to a given individual was added to each appropriate well, as directed by Applied Biosystems. To ensure the greatest comparative accuracy for a given miRNA, all of the 6 individual cDNA samples were examined together on the same plate for a given miRNA at the same time. This also served as a good control measure for assessing the arrays, which required a separate card for each individual examined: if the array results were correct, and preparation to preparation differences were non-significant, then the miRNA qRT-PCR assays which could allow for examining all samples for a given miRNA together on the same plate, would agree with the array results. For each plate, MammU6 was used for normalization across plates. Each sample was analyzed in 3 experimental replicates (3 wells loaded with FMM + cDNA) to allow for accuracy verification. Altogether, 66 individual qPCR reactions were performed in the technical triplicate, to create replicates for qualitative analysis. Technical triplicate means that each reaction was repeated 3 times and the average calculated. Individual differences among 3 repeats were smaller than 1 Ct. Positive, negative and NoRT controls were included routinely into each PCR reaction.

qRT-PCR was performed on a StepOne Plus ABI thermocycler (Applied Biosystems). Standard settings were used, except as noted. For assays, ΔΔCt assay was selected. Standard ramp was used, and the run length was standard (i.e., not a "fast" reaction). 6-FAM was the reporter, and the quencher was NFQ-MGB. The reaction volume was 20 ul, and a standard 40 cycle run was conducted, though no 50C step was used at the beginning as the samples lacked UNG. Thus, the run consisted of: 95°C 10 min, followed by 40 cycles of (95°C 15 s, 60°C 1 min). Data collection occurred on the 60°C step at each cycle.

DIANA mirPath and KEGG Analysis

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a free database (Kanehisa and Goto, 2000), which provides information about gene content of biological pathways, interrelationship of these genes, and involvement of pathways in pathogenesis of diseases. Data are supplied as tables and wiring diagrams of interactive networks. Analysis of KEGG pathways regulation by microRNA was performed using DIANA mirPath (v1.0 and v2.0), based on high-accuracy target prediction algorithms: DIANA-microT-CDS (Papadopoulos et al., 2009) and TargetScan (Garcia et al., 2011) with FDA correction (Benjamini and Hochberg, 1995). DIANA mirPath is a free web-based computational tool, developed by DNA Intelligent Analysis Laboratory (DIANA) that identifies biological pathways potentially altered by the expression of a single or multiple microRNAs (Papadopoulos et al., 2009; Vlachos et al., 2015). Pathways are sorted according to a *p* value (threshold 0.05) and corrected for a False Detection Rate (FDR). *p* value indicates extent of regulation of a particular pathway by specific microRNA(s). mirPath also provides the number and names of genes regulated by miRNAs in each KEGG pathway. We used mirPath-v1.0 and mirPath-v2.1 and, based on *p* value, the number of genes and the number of microRNAs involved, determined which pathways and diseases could be affected the most by the set of alcohol-regulated microRNAs.

Wright-Giemsa Staining

We used Wright-Giemsa staining (Sigma, MO) to visualize cellular components of saliva following the standard protocol. Briefly, around 0.5 ml of saliva was transferred onto a glass slide without bubbles and spread out to a thin, even layer using the edge of another glass slide. After air-drying for several minutes, the slide was stained with 1–2 ml of the Wright-Giemsa stain for two minutes and rinsed with deionized water. Excess water was blotted and the slide was air-dried. Images of stained cells were

captured using a brightfield mode on the Olympus FSX-100 microscope (Olympus, PA).

Cell Sorting and Flow Cytometry

Saliva was diluted 1:1 with magnetic affinity cell sorting (MACS) running buffer and filtered through a 5 μM membrane (Millipore, MA) by centrifugation at room temperature for 30 min at 300 rpm. The cells were collected from the membrane using FcR Blocking Reagent (Miltenyi Biotec, CA) and labeled with anti-CD45-conjugated microbeads (Miltenyi Biotec). After 15 min of incubation at 4°C, cells were resuspended in 50 μL of MACS buffer and separated through an autoMACS cell separator. The unlabeled cells (negative fraction) passed through while the magnetically labeled cells (positive fraction) were retained within the column eluted later from the column and stained with FITC conjugated anti-CD-45 antibodies (Miltenyi Biotec). After a ten-minute incubation in the dark at 4°C, the cells were washed and resuspended in a 200 μL MACS buffer for flow cytometry. Flow cytometry was performed using an Influx™ Mariner 209s high speed flow cytometer equipped with 488 nm 200 W Argon blue excitation laser and 70 μm nozzle tip (BD Biosciences, CA). The software used to collect data was BD Software. For analysis, the trigger threshold was based on forward scatter (FSC). Unstained cells were identified by FSC and side scatter (SSC). Initially, cells were sorted out by their size to remove large buccal cells. Size gating was determined by geometric mean of FSC value relative to known fluorescent standard bead size (6–7 μm). Next, the remaining cells were characterized as human leukocytes based on their size and simultaneous detection of 520 nm wavelength emitted by FITC-conjugated anti-CD45 antibodies recognizing CD45 antigen, characteristic for white blood cells. Data were analyzed using FlowJo software (BD Biosciences) and shown as 3D density scatter plots and histogram plots. The 3D density scatter plots show two-parameter data (*X* vs. *Y* axes) with the different colors representing the frequency of events falling at each position. The histogram plots show single-parameter data (*X* axis) with *Y* axis representing the number of events (counts) for this parameter.

Target Analysis

Four targets were selected using mirPath analysis of the KEGG (Kanehisa and Goto, 2000) adherens junction pathway in DIANA (Vlachos et al., 2012): catenin (CTNND1), a receptor tyrosine kinase (MET), a serine/threonine kinase (NLK) and a transcription factor (SNAi1). Targets were selected based upon the following criteria: 1) direct relevance to the pathway, 2) diverse cellular location and function, 3) prediction of being targeted by multiple miRNAs, 4) prediction of regulation by upregulated microRNAs only, 5) confirmation of possibility of regulation by microRNA *via* RNAhybrid (Rehmsmeier et al., 2004). Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was found to be a stable control under alcohol exposure in mammalian cells and was used as a reference (Maran et al., 2004; Rehmsmeier et al., 2004; Mason et al., 2012). Expression of each target was measured in at least 3 control and 3 alcohol samples.

Since each of these targets has several different alternative splice variants, based on bioinformatic and biological evidence in KEGG, NCBI and Ensembl, we designed qPCR primers and probes against the following transcripts: MET—ENST00000318493 and ENST00000397752, NLK—ENST00000407008 (**Supplementary Table S1**) using PrimerQuest software (IDT). PrimeTime primer and probe sets were selected to operate under the standard conditions for Taqman qRT-PCR assay, using 6-carboxyfluorescein (6-FAM) as the reporter. For detection of CTNND1, SNAi1 and GAPDH we used validated, pre-designed Taqman gene expression assays (Life Technologies)—**Supplementary Table S1**.

RNAhybrid (Rehmsmeier et al., 2004) was used to examine binding sites in the 3′UTRs of these transcripts as previously (Pietrzykowski et al., 2008). 3′UTRs of each transcript were obtained from Ensembl. Settings adjusted included: hits per target (10), energy cutoff (−19 kcal/mol), and helix constraint (2–8). Together these settings displayed only hits of −19 kcal/mol or lower, which had perfect matches to at least the seed region with a strong likelihood of being valid binding sites (**Supplementary Figure S1**).

Quantitative Real-Time PCR Analysis of Individual miRNA Targets

Reverse Transcription (RT) was conducted using the SuperScript VILO cDNA synthesis kit (Life Technologies) according to manufacturer's suggestions to create total cDNA from total RNA samples. Samples examined included 4 controls and 4 alcohol abusers for each target. Between 400ng and 1ug total RNA was used to generate cDNA for each sample. The thermocycler program used was 25°C for 10′ followed by 42°C for 120 min, 85°C for 5′, 4°C ∞. An extended time at 42°C was used as recommended in the protocol, to generate a higher quantity of template.

qPCR was conducted using the Taqman gene expression protocol and supplies on an ABI Step-one Plus machine (Life Technologies) using standard conditions and FAM as a reporter. Data was collected on the 60°C step at each cycle. No RT (no reverse transcriptase at RT step) and No template (No DNA added at qPCR step) were negative controls used in each run. The ΔΔCt method of analysis, comparing the results against GAPDH was used. Samples were internally compared against their own GAPDH, to result in accurate relative quantification.

Statistical Analysis

GraphPad (v6.0a) for Mac (GraphPad Software Inc., La Jolla, CA) and R package were used for analysis. The Mann-Whitney *U* test was performed to detect differentially expressed miRNAs between groups (alcohol abusers vs. non-abusing controls). Pearson correlation with two-tailed *p* values was used to examine correlation between qPCR results and the corresponding array results, as well as correlation between microRNA expression and the amount of consumed alcohol. Correlations were expressed as a correlation coefficient (r) with $p < 0.05$ set as significant.

TABLE 1 | Characterization of patients and their drinking behavior. On a few occasions patients did not provide exact number of years drinking (designated by "many*"), however the duration was always more than 10 years.

#	Age (years)	Gender	Ethnicity	Drinking period (years)	Drinking pattern	Alcohol intake (gEtOH/24 h)
1	39	M	Hispanic	20	Daily	137
2	56	M	African Am	20	Daily	82
3	44	M	Caucasian	15	Daily	82
4	55	M	Caucasian	25	Daily	164
5	40	M	African Am	15	Daily	137
6	47	F	African Am	20	Daily	96
7	61	M	Caucasian	45	Daily	274
8	45	F	Caucasian	28	Daily	110
9	44	M	Caucasian	many*	Daily	82
10	44	M	Hispanic	10	Daily	1,096
11	49	M	Asia/India	20	Daily	55
12	50	F	African Am	35	Daily	205
13	55	M	Caucasian	10	Daily	55
14	35	M	Hispanic	many*	Daily	82
15	56	M	Caucasian	20	Daily	96
16	41	F	Caucasian	23	Daily	55
17	52	M	Caucasian	35	Daily	164
18	49	M	African Am	30	Daily	110
19	46	M	Caucasian	many*	Daily	466
20	46	M	Hispanic	10	Daily	274
21	33	M	Hispanic	15	Daily	1,096
22	45	M	Caucasian	10	Daily	137
				Controls		
1	35	M	Hispanic	0	Occasionally	13
2	40	M	Caucasian	0	Occasionally	20
3	41	F	Caucasian	0	Occasionally	20
4	45	M	Caucasian	0	Occasionally	13
5	49	M	African Am	0	Occasionally	20
6	46	M	African Am	0	Occasionally	20
7	42	M	Caucasian	0	Occasionally	20
8	42	M	Caucasian	0	Occasionally	20
9	38	M	Caucasian	0	Occasionally	7
10	32	M	Caucasian	0	Occasionally	13
11	35	M	Caucasian	0	Occasionally	20
12	39	M	Caucasian	0	Occasionally	7
13	43	M	Caucasian	0	Occasionally	20
14	44	M	African Am	0	Occasionally	13
15	34	F	Caucasian	0	Occasionally	20

Receiver Operating Characteristic (ROC) Curve analysis was performed to determine sensitivity (true positive results) and specificity (true negative results) of the panel of 38 microRNAs, which expression was significantly changed in alcohol abusers compared to the control group, as a biomarker screen. The area under the ROC curve (AUC) quantifies the overall ability of the test to discriminate between individuals with the disease (i.e., alcohol abusers) and without the disease (i.e., controls). AUC values fall between 0.5 (no discrimination between two groups) and 1.0 (a perfect test with zero false positives and zero false negatives) with p value set at 0.05 indicating the significance of the test. ROC was used to compare the alcohol-specific microRNA panel with the control group, as well as each individual subject with the alcohol-specific microRNA panel to determine false positives and false negatives.

Principal Component Analysis (PCA) was performed using R package (Venables and Ripley, 2002), specifically "prcomp" function in R package "stats" (1) on the 38 miRNA expression dataset. Data were centered and scaled $[x/(x_{max} - x_{min})]$ per variable before analysis. Data are shown as a 3D plot.

RESULTS

Characterization of Patients

We wanted to establish whether profiling of microRNA in saliva could be a useful biomarker tool in distinguishing heavy alcohol drinkers from sporadic alcohol users. Thirty-eight patients at the primary health care center located in Trenton, NJ were considered for this study based on their chronic alcohol abuse or its lack. The characteristics of patients are summarized in **Table 1**. The alcohol-abusing patients were of both genders (82% males, 18% females), mixed ethnic backgrounds, and an average age of 47 years (+/− 7 years).

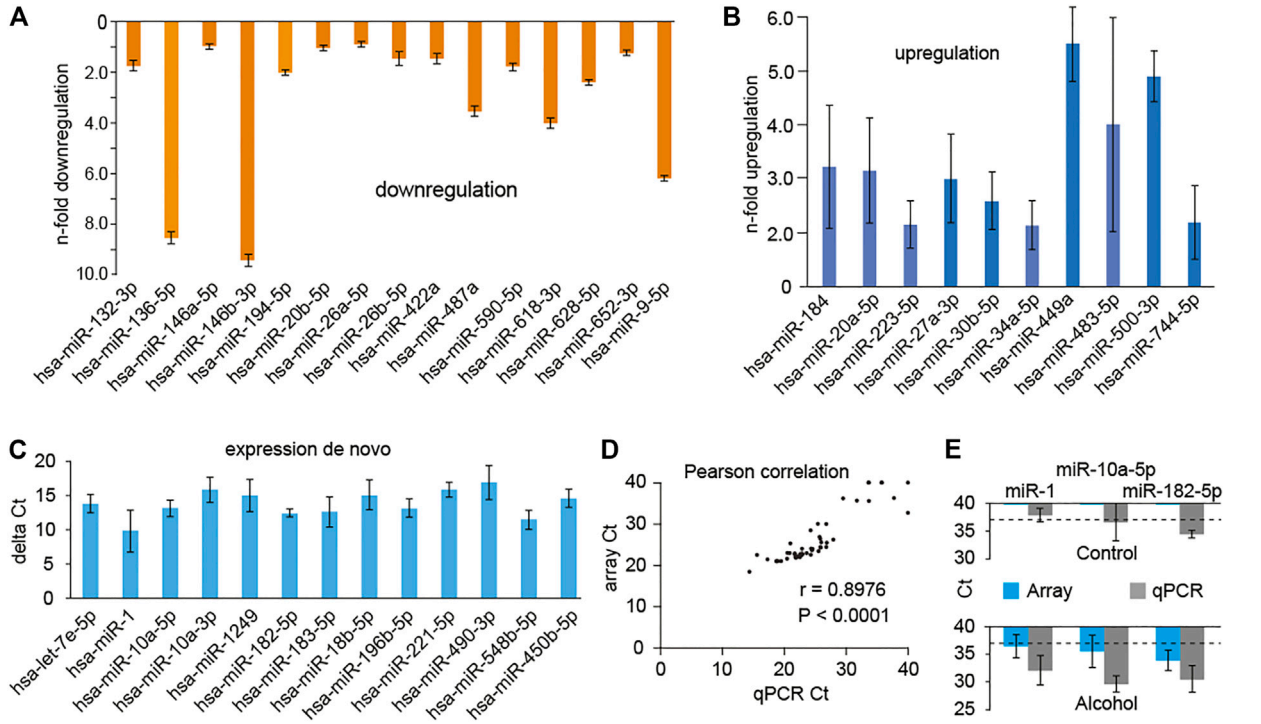


FIGURE 1 | microRNA species, for which expression is significantly changed in the saliva of alcohol abusers ($n = 22$) compared to non-abusing controls ($n = 15$). **(A)** microRNAs downregulated compared to the control group. Expression levels are expressed as n-fold change. **(B)** microRNAs upregulated compared to the control group. Expression levels are expressed as n-fold change. **(C)** microRNA species upregulated *de novo*, meaning these microRNAs were undetected in the control group, but present in the alcohol-abusing group. Lack of expression in controls prohibits comparison of expression levels of these microRNAs between both groups, therefore results are presented as ΔCt of the alcohol abuser group. **(D)** High ($r = 0.8976$) correlation between microRNA microarray results and individual microRNA quantitative PCR (qPCR) results. **(E)** Validation of *de novo* upregulation of microRNAs in alcohol abusers. Representative microRNAs indicate borderline expression levels in controls, while detectable in alcohol abusers by both, arrays and qPCR. Note: microRNAs are named according to conventional nomenclature in which mature microRNAs coming from the complementary strands of a precursor are called miR-xx-5p and -3p, respectively, not miR-xx and miR-xx*. Significance determined using Mann-Whitney *U* test with $p < 0.05$. Bars represent the mean±SEM (Standard Error of the Mean). U6 was used as a reference.

The patients had been drinking on a daily basis for at least 10 years before admission with the average drinking period of 21 years. Their current alcohol intake is shown in **Table 1**. The non-abusing controls are shown in the lower part of **Table 1**. They also represent both genders and mixed ethnicity, with an average age of 40 years (+/−5 years). In contrast to the alcohol-abusing group, the control patients were drinking sporadically, no more than 20 g alcohol per drinking session.

Changes in microRNA Expression in Saliva of Alcohol Abusers

Human-specific, low-density microRNA microarrays (v3.0, Applied Biosciences) were used to measure microRNA expression profiles in the saliva of alcohol non-abusers and abusers. First, we compared microRNA profiling results of our control group with microRNA profiling in saliva of healthy controls of other studies. U6 was used as a reference microRNA.

Comparison of microRNA expression levels between alcohol abusers and controls revealed 38 significantly ($p < 0.05$) changed microRNA species, which fell into three distinct groups (**Figure 1**): 1) microRNAs downregulated in alcohol abusers compared to non-abusers, 2) microRNAs upregulated in abusers, and 3) microRNAs undetected in non-abusers but present in abusers (expression *de novo*).

The downregulated group consists of 15 microRNAs (**Figure 1A**): miR-132-3p, miR-136-5p, miR-146a-5p, miR-146b-3p, miR-194-5p, miR-20b-5p, miR-26a-5p, miR-26b-5p, miR-422a, miR-487a, miR-590-5p, miR-618-3p, miR-628-5p, miR-652-3p, and miR-9-5p. The downregulation ranged from 1 to almost 10-fold. The upregulated group consisted of 10 microRNAs (**Figure 1B**): miR-184, miR-20a-5p, miR-223-5p, miR-27a-3p, miR-30b-5p, miR-34a-5p, miR-449a, miR-483-5p, miR-500-3p and miR-744-5p. The range of upregulation was from two to almost six-fold. The expression *de novo* group is shown in **Figure 1C**. Since these microRNAs were undetected in the controls, their expression level cannot be compared to the control level and expressed as a fold change. Therefore, we show expression of these microRNAs as ΔC_t, where values of expression have been normalized to a housekeeping microRNA (U6). The upregulated *de novo* group consisted of 13 microRNAs (**Figure 1C**): let-7e-5p, miR-1, miR-10a-5p, miR-10-3p, miR-1249, miR-182-5p, miR-183-5p, miR-18b-5p, miR-196b-5p, miR-221-5p, miR-490-3p, miR-548-5p and miR-450b-5p.

Microarray Validation by Real-Time RT-PCR Analysis

To validate results from microRNA microarray analysis, we performed qRT-PCR. We chose ten microRNAs from each category and measured their expression in both groups: non-abusers and abusers. We also included microRNAs for which expression didn't change or was undetectable in both groups. The miRNAs examined include hsa-miR-1, -10a-5p, -182-5p, -26a-5p, -27a-3p, -20a-5p, -29c-3p, -106a-5p, -9-5p, -618 and miR-U6 as a reference microRNA. All together we performed 60 individual qRT-PCR reactions in triplicate to validate microarray results. Overall correlation analysis indicated a high level of agreement between array and qRT-PCR data ($r = 0.8976$, $p < 0.0001$, **Figure 1D**, **Supplementary Table S2**) validating the microRNA microarray results.

Absence of microRNAs in the control subjects in the expression *de novo* group could be due to the low detection limit of microarrays. qRT-PCR is a more sensitive method able to detect a single copy of microRNA. Therefore, we used qRT-PCR to further test expression levels of some of the microRNAs in the expression *de novo* group using the same samples used for microarrays. Representative results are shown in **Figure 1E**. The qRT-PCR results confirmed that the level of expression of microRNAs in the control subjects of the expression *de novo* group is near or below detection limit, suggesting their physiological irrelevance in non-abusers. Furthermore, the qRT-PCR confirmed significant upregulation of these microRNAs in the abusers. Together, these results show significant deregulation of specific miRNAs in saliva of alcohol abusers, validating consideration of the use of microRNA profiling in the development of biomarkers of excessive alcohol consumption.

Functional Analysis

To investigate the functional significance of the combinatorial changes in the microRNA expression in alcohol abusers we performed biological pathway analysis using DIANA mirPath software and KEGG database combined with biological measure of key elements in one of these pathways. Top five biological pathways affected the most are presented in **Table 2** including number of genes in each pathway targeted by microRNAs, as well as most relevant upregulated and downregulated microRNA species.

Overall, the impact of upregulated microRNAs seems to be larger than downregulated microRNAs (**Table 2**, right column). As the primary role of microRNA is suppression of gene expression (Miranda, 2014), it is permissible to assume that the increased levels of microRNAs would likely cause overall downregulation of their targets and impairment of function of pathways, to which the targets belong. Thus, this bioinformatic analysis suggests that several crucial cellular pathways could be affected by changes in microRNA expression in alcohol abusers *via* suppression of some key elements of these pathways.

One of the top "hits" was adherens junction, which is a contact point between various types of cells, important in maintaining cellular function and tissue integrity. Indeed, miRNAs have been found to be crucial for the regulation and pathology of adherens junction (Cichon et al., 2014). Targeting of elements of the adherens junction pathway by alcohol-regulated microRNAs is shown as a wiring diagram in **Figure 2A**. Malfunction of adherens junctions has been indicated in the pathologies of the immune system (Wang et al., 2011; Kaymak et al., 2021) and the development of cancer (Vasioukhin, 2012; Bischoff et al., 2020). Since these diseases are aggravated by alcohol consumption (Sambuy, 2009; Meadows and Zhang, 2015), we decided to explore effects of microRNA de-regulation on the adherens junction pathway elements.

We focused on four targets located in different cellular compartments: 1) catenin (CTNND1)—a cytoplasmic protein associated with cadherin, plasmalemmal adherens junction receptor, 2) a plasma membrane receptor tyrosine kinase (MET), 3) a cytoplasmic kinase (NLK) and 4) a transcription factor (SNAi1) located in the cell nucleus (**Figure 2A**, blue circles). According to mirPath, transcripts of each of these genes were targeted by multiple microRNAs (**Figure 2B**), all of them upregulated. Since microRNAs work mainly as suppressors of gene expression, we assumed to observe an inverse correlation between miRNA expression changes and changes in the steady-state levels of their predicted targets. This was true for three of these targets: MET, SNAi1 and CTNND1 (**Figure 2C**) as determined by qPCR. However, the expression of the fourth tested target, NLK was unchanged (**Figure 2C**). In all cases GAPDH was used as a reference. Suppressive effect of microRNA on its mRNA target depends on the strength of thermodynamic interaction between both molecules. Closer examination of the strength of microRNA: mRNA target interaction by RNAhybrid (Rehmsmeier et al., 2004; Krüger and Rehmsmeier, 2006) revealed that the minimum free energy hybridization for any relevant microRNA with the NLK 3′UTR was above −20 kcal/mol (**Supplementary Figure S1**) suggesting a weak interaction (marked on **Figure 2B** as open ovals). In contrast, the minimum free energy of hybridization of all other downregulated targets with their microRNAs was below −20 kcal/mol (**Supplementary Figure S1**) indicating stronger interactions (**Figure 2B**—full ovals).

Altogether, modulation of cellular adherens junction pathway by alcohol *via* microRNA could have severe consequences on cell adherens by decreasing expression of important elements of the adherens junction pathway. Wiring diagrams of other pathways are shown in the **Supplementary Figure S2**.

Cellular Content of Saliva

Use of microRNA target prediction programs allowed us to understand better the impact of chronic alcohol abuse on cellular processes. To determine the main cellular source of human microRNA in the samples, we performed Giemsa staining (**Figure 3A**). We observed the presence of large cells of irregular shape with abundant cytoplasm—typical morphological features of squamous epithelial buccal mucosa cells (cheek cells, average size 50–70 μm diameter). We also detected much smaller cells (average size 9.4 μm diameter) with large nuclei and a rim of cytoplasm resembling

TABLE 2 | Top five biological pathways, which could be affected by an overall change in microRNAs expression in patients chronically abusing alcohol. Analysis was performed using DIANA mirPath and KEGG software with FDR correction and *p* value threshold of 0.05. Blue color depicts upregulated miRNA, orange—downregulated miRNA.

Biological Pathway	Pathway ID	# Genes	# Regulated Genes	p-value	microRNAs affecting at least 10% of genes (# regulated genes)
Adherens junction	hsa04520	88	43	1.86E-06	miR-1 (4), miR-9-5p (7), miR-182-5p (18), miR-548b-5p (7), miR-450b-5p (5), miR-449a (8), miR-34a-5p (8), miR-30b-5p (5), miR-26a-5p (6), miR-26b-5p (6), miR-20a-5p (7), miR-20b-5p (7), , miR-194-5p (4)
Endocytosis	hsa04144	139	199	6.74E-07	miR-182-5p (11), miR-18b-5p (10), miR-548b-5p (19), miR-450b-5p (13), miR-20a-5p (27), miR-30b-5p (14), miR-20b-5p (25), miR-9-5p (17)
Wnt signaling pathway	hsa04310	151	81	6.45E-06	miR-26a-5p (17), miR-182-5p (8), miR-183-5p (8), miR-26b-5p (16), miR-20a-5p (14), miR-20b-5p (14), miR-450b-5p (15), miR-548b-5p (13), miR-30b-5p (9), miR-34a-5p (9), miR-449a (8),
MAPK signaling pathway	hsa04010	259	127	1.64E-05	miR-1 (13), miR-20b-5p (22), miR-9-5p (20), miR-20a-5p (25), miR-183-5p (14), miR-548b-5p (16), miR-26a-5p (17), miR-26b-5p (17), miR-30b-5p (17), miR-450b-5p (16)
ErbB signaling	hsa04012	88	51	3.55E-05	miR-548b-5p (6), miR-450b-5p (8), miR-30b-5p (8), miR-20a-5p (9), miR-20b-5p (10), miR-26a-5p (7), miR-26b-5p (7), miR-194-5p (5)

leukocytes (**Figure 3A**). Indeed, immune cells constantly migrate to the oral cavity as the first line of defense against pathogens and they can be present in saliva (Schiott and Loe, 1970). To further determine the phenotype of these cells we performed flow cytometry first filtering out large cheek cells based on their size and then harvesting remaining cells using antibodies against CD45 (leukocyte common antigen) specific for human leukocytes. There were almost no cells left after filtering and harvesting (**Figure 3B**). In flow cytometry forward scatter (FSC) intensity is proportional to the diameter of the cell, therefore FSC signal can be used to discriminate cells by size. On FSC/Fluorescence 3D density scatter plot of unlabeled cells we observed specific clustering of the vast majority (92.1%) of the cells by macrophage-relevant size in one gating window, below the fluorescence threshold (**Figure 3C**). These cells, when labeled for presence of CD45 antigen, shifted above the fluorescence threshold and accumulated mostly (95%) as a single cluster of the cells of the same size as in (b) in one gating window of a scatter plot (**Figure 3D**). Homogeneity of cells and their specificity of expressing CD45 antigen can be also seen as a single peak on histograms, and the shift from below to above the threshold, as shown in **Figures 3E,F**, respectively. The shift of almost all cells from one gating window to another (from **Figures 3C,D**, and from **Figures 3E,F**), based on presence of CD45 antigen, indicated the presence of human leukocytes in saliva. Together, morphological and flow cytometry data indicate that the two major cell populations in saliva are squamous epithelial buccal mucosa cells and leukocytes.

microRNA Profiling Supports Alcohol Contribution to Carcinogenesis

Each time alcohol is consumed, it travels throughout the body and can affect other cell types in addition to cells present in the oral cavity. Thus, changes in microRNA profiles observed in cells present in the oral cavity could potentially provide a window into possible changes in microRNA expression in other cell types. We have used mirPath software and KEGG database to determine the potential contribution of alcohol-sensitive microRNAs to pathogenesis of diseases. The top hits of our analysis are various types of cancer, all originating mainly from epithelial cells (**Table 3**). Summary of the potential regulation of cellular pathways in carcinogenesis by alcohol-sensitive microRNAs is shown as a wiring diagram in **Figure 4** and in **Supplementary Figure S3**. Blue circles on **Figure 4** depict cellular processes abnormal in cancer, which expression of some elements has been indicated to be aberrant by the current study. Interestingly, some of alcohol-regulated microRNAs suggested to be involved in pathogenesis of cancer, are also regulators of inflammatory responses, regardless of the cancer type (**Table 3**, right column), corroborated studies establishing a link between inflammation and cancer (Miranda, 2014). In this study, a group of alcohol-regulated, inflammatory microRNAs involved in all of these types of cancers include six microRNAs: miR-9-5p, miR-20a-5p, miR-20b-5p, miR-27a-3p, miR-30b-5p, and miR-182-5p.

microRNAs as Potential Biomarkers of Alcohol Consumption

Our study indicates that all together 38 microRNAs are differentially expressed in a group of heavy alcohol consumers compared to the control group (**Figure 5**). We attempted to measure the correlation between change in the expression level of individual microRNAs and the amount of alcohol consumed. We observed that three microRNAs expressed moderate to weak correlation: miR-196b-5p ($r = -0.6672$, $p < 0.0127$), miR-26b ($r = -0.4273$, $p < 0.0473$) and miR-223 ($r = -0.4290$, $p < 0.0463$).

Next, we wondered whether the microRNA panel as a whole could adequately segregate alcohol abusers from control non-

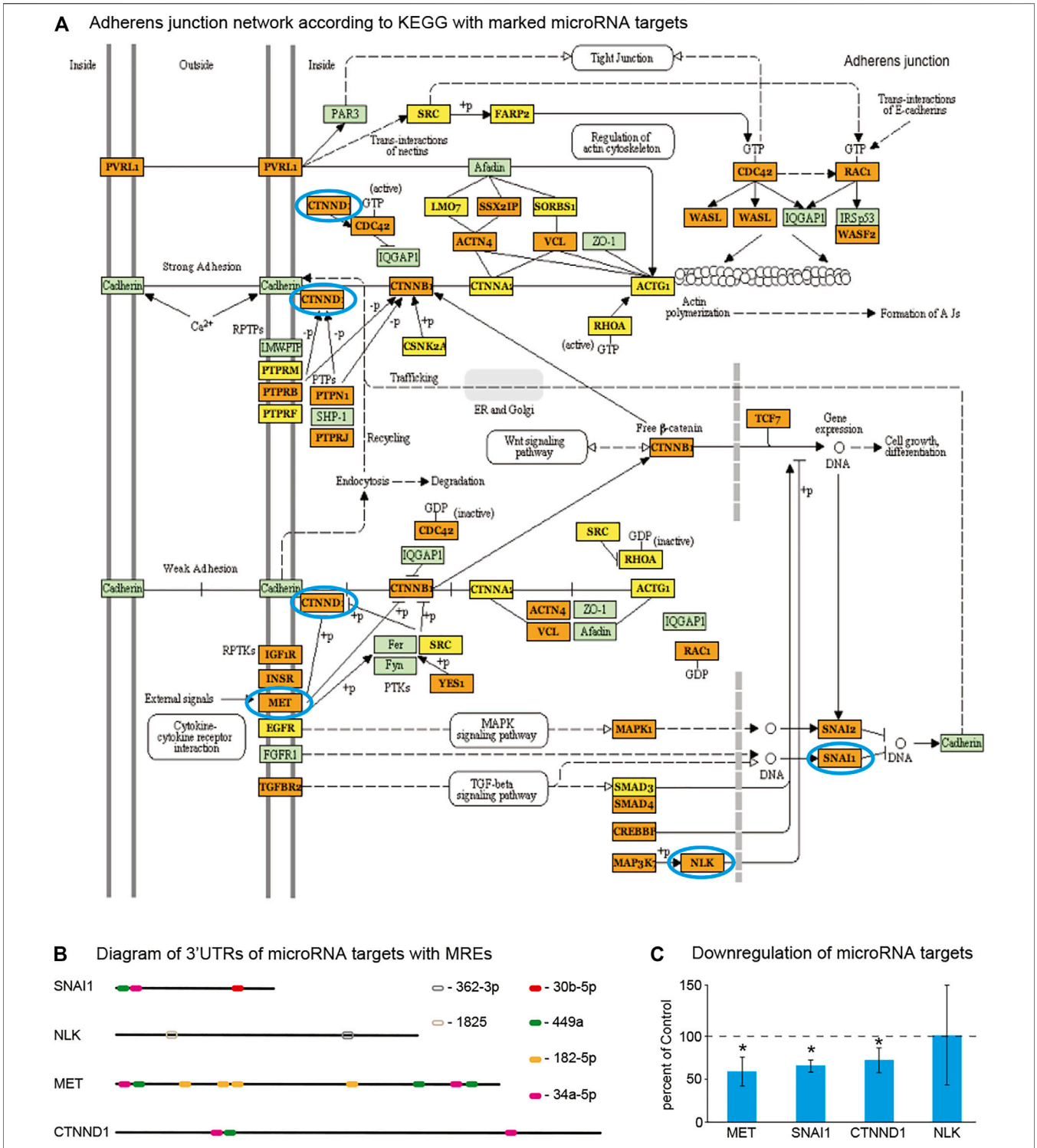


FIGURE 2 | A wiring diagram showing potential regulation of adherens junction pathway by microRNAs relevant in alcohol abuse. **(A)** The diagram shows relationships of several gene products important for the proper function of the adherens junction. It also shows links to other pathways, some of them potentially regulated by alcohol. Gene products in yellow boxes are affected by one microRNA, gene products in orange boxes are affected by more than one microRNA. Gene products in green boxes are unaffected. Gene products in blue circles were tested by qPCR. The wiring diagram was created using KEGG (Kyoto Encyclopedia of Genes and Genomes) and DIANA-mirPath software with FDR correction and *p*-value threshold of 0.05. **(B)** Diagram of 3′UTRs of four genes [gene products in blue circles in **(A)**] targeted by several alcohol-regulated microRNAs (see the legend) showing relative positions of individual microRNA responsive elements (MREs). Empty circles indicate hybridization energy between a microRNA and its target above −20 kcal/mol. **(C)** Decreased expression levels of four microRNA targets marked in **(A)** and depicted in **(B)** in alcohol samples. GABDH was used as a reference. $N \geq 3$. * - $p < 0.05$.

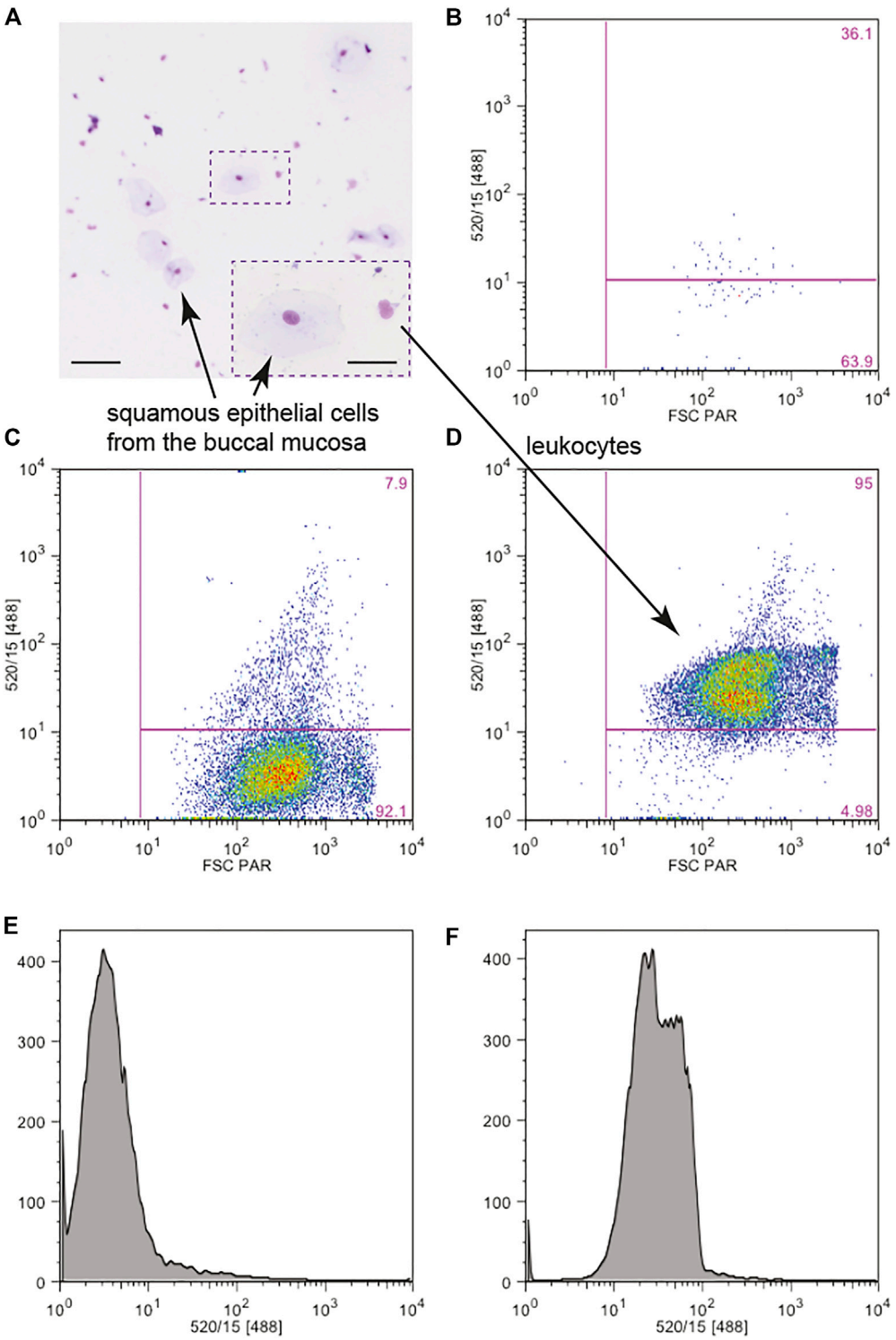


FIGURE 3 | Saliva contains epithelial buccal cells and leukocytes. **(A)** Wright-Giemsa staining of saliva indicates the presence of large buccal cells as well as smaller cells resembling leukocytes. The rectangular area demarcated by the dotted line is enlarged in the right, bottom corner. Lower magnification scale bar—100 um, higher magnification scale bar—30 um. Cell sorting and flow cytometry showed the presence of leukocytes in saliva. **(B)** The unlabeled cells were not detected. **(C)** A representative 3D density scatter plot shows that most of the cells, when unlabeled, clustered into a homogeneous group in one gating window below the fluorescent threshold based on size representing leukocytes. **(D)** Upon labeling with fluorescent CD45-specific antibodies almost all of cells from **(B)** shifted above the threshold and accumulated in one gating window indicating that these cells are human leukocytes. **(E)** A representative histogram showing the homogeneity of unlabeled cells based on size. **(F)** Upon labeling with fluorescent CD45-specific antibodies almost all of cells from **(E)** shifted above the threshold indicating that these cells are human leukocytes. Red lines in **(B–D)** delineate gating windows. FSC—forward scatter, 520/488—fluorescent emission/excitation wavelengths.

TABLE 3 | The overall change in microRNAs expression by alcohol abuse could lead to several cancers of epithelial origin. Analysis was performed using DIANA mirPath and KEGG software with FDR correction and a *p*-value threshold of 0.05. The top six cancers are shown. Inflammatory microRNAs are marked red.

Cancer	Origin	Pathway ID	# Genes	# Regulated Genes	p-value	microRNAs affecting at least 10% of genes (# regulated genes)
Glioma	Epithelium, Mesenchyme	hsa05214	65	39	1.17E-07	miR-20b-5p (6), miR-20a-5p (6), miR-30b-5p (6), miR-1 (5), miR-182-5p (5), miR-194-5p (4), miR-548b-5p (7), miR-26b-5p (5), miR-9-5p (4), miR-26a-5p (5), miR-450b-5p (6), miR-449a (6), miR-34a-5p (6)
Prostate cancer	Epithelium	hsa05215	89	52	6.45E-06	miR-20a-5p (10), miR-20b-5p (9), miR-182-5p (8), miR-9-5p (6), miR-30b-5p (7), miR-26b-5p (6), miR-1 (7), miR-450b-5p (5), miR-548b-5p (7), miR-449a (6), miR-34a-5p (6), miR-26a-5p (6)
Colorectal cancer	Epithelium	hsa05210	62	38	4.27E-05	miR-182-5p (4), miR-20a-5p (6), miR-9-5p (4), miR-30b-5p (6), miR-20b-5p (8), miR-1 (7), miR-449a (5), miR-34a-5p (6), miR-450b-5p (5)
Pancreatic cancer	Epithelium	hsa05212	70	37	8.23E-04	miR-20a (9), miR-20b (8), miR-548b-5p (8), miR-9 (4), miR-30b-5p (6), miR-449a (5), miR-34a-5p (6)
Small cell lung carcinoma	Epithelium	hsa05222	86	44	2.27E-04	miR-20b-5p (5), miR-20a-5p (7), miR-9-5p (6), miR-548b-5p (8), miR-30b-5p (6), miR-26b-5p (4), miR-26a-5p (4), , miR-34a-5p (5), miR-449a (5)
Melanoma	Epithelium	hsa05218	71	39	2.47E-03	miR-20b-5p (7), miR-20a-5p (7), miR-9-5p (4), miR-182-5p (6), miR-26b-5p (6), miR-1 (5), miR-30b (5), miR-26a-5p (6), miR-194 (4), miR-34a-5p (6), miR-449a (6), miR-450b-5p (4), miR-548b-5p (4)

abusers. First, we used the Principal Component Analysis (PCA), because PCA has been widely used to reduce high-dimensional data and to determine the similarities and dissimilarities of biological samples without losing too much information (van der Werf et al., 2006; de Haan et al., 2007; Simmons et al., 2015). In this study, PCA was applied to the 38 miRNAs expression data of 22 alcohol and 15 control individuals. The three PCs account for around 61% of total variance: PC1 explains 38.93% of the total variance in this dataset, PC2 explains 14.7% of the variance, and PC3 explains 7.16% of the variance (**Figure 6A**). The 3D plot of three principle components showed that most of alcohol and control individuals could be segregated into separate clouds with some overlap (**Figure 6A**).

Next, we used Receiver Operating Characteristic (ROC) Curve analysis to determine sensitivity and specificity of the microRNA panel, and to calculate false positive and false negative values. The area under the ROC curve (AUC) quantifies the overall ability of the test to discriminate between individuals with the disease (*i.e.,* alcohol abusers) and without the disease (*i.e.,* controls). An AUC of 1 would indicate a perfect test of 100% sensitivity and 100% specificity, while an AUC value of 0.5 would show no discrimination between two groups (**Figure 6B**, dotted line). The ROC used to compare the alcohol-specific microRNA panel with the control group had the AUC value of 0.7668 significantly above random sampling ($p < 0.0003$). We next used the ROC analysis to determine how well each individual enrolled in the study (either control or alcohol abuser) would be classified to the alcohol group by the microRNA alcohol panel. **Figure 6C** shows the probability of each individual's microRNA screen to differ from the microRNA alcohol panel. Two out of fifteen control individuals (13%) are false positives (**Figure 6C**, control, purple oval), while six out of twenty-two alcohol abusing individuals (27%) are false negatives (**Figure 6C**, alcohol, purple oval), showing rather encouraging results, and warranting pursuing further the use of microRNA profiling in saliva as a biomarker of alcohol abuse.

DISCUSSION

Alcohol Use Disorder (AUD) is a complex chronic disease affecting a large portion of our society. AUD is characterized by escalation of alcohol consumption over years, development of tolerance and addiction, and occurrence of relapse periods thwarting abstinence efforts. The pathological effects of alcohol on bodily organs vary, depending in part on the amount of consumed alcohol and the length of the excess of alcohol consumption. For example, chronic alcohol abuse is frequently associated with malfunction of the immune system, and can also lead to the development of several cancers (Meadows and Zhang, 2015).

Since AUD is a very serious societal issue, establishment of alcohol biomarkers is of great importance for clinical practice, public safety, criminal justice and research (Litten et al., 2010; Nanau and Neuman, 2015). A biomarker is defined as any substance present in an organism which can be measured and used to indicate specific functions of the body's systems in health and disease. Biomarkers can have an important diagnostic,

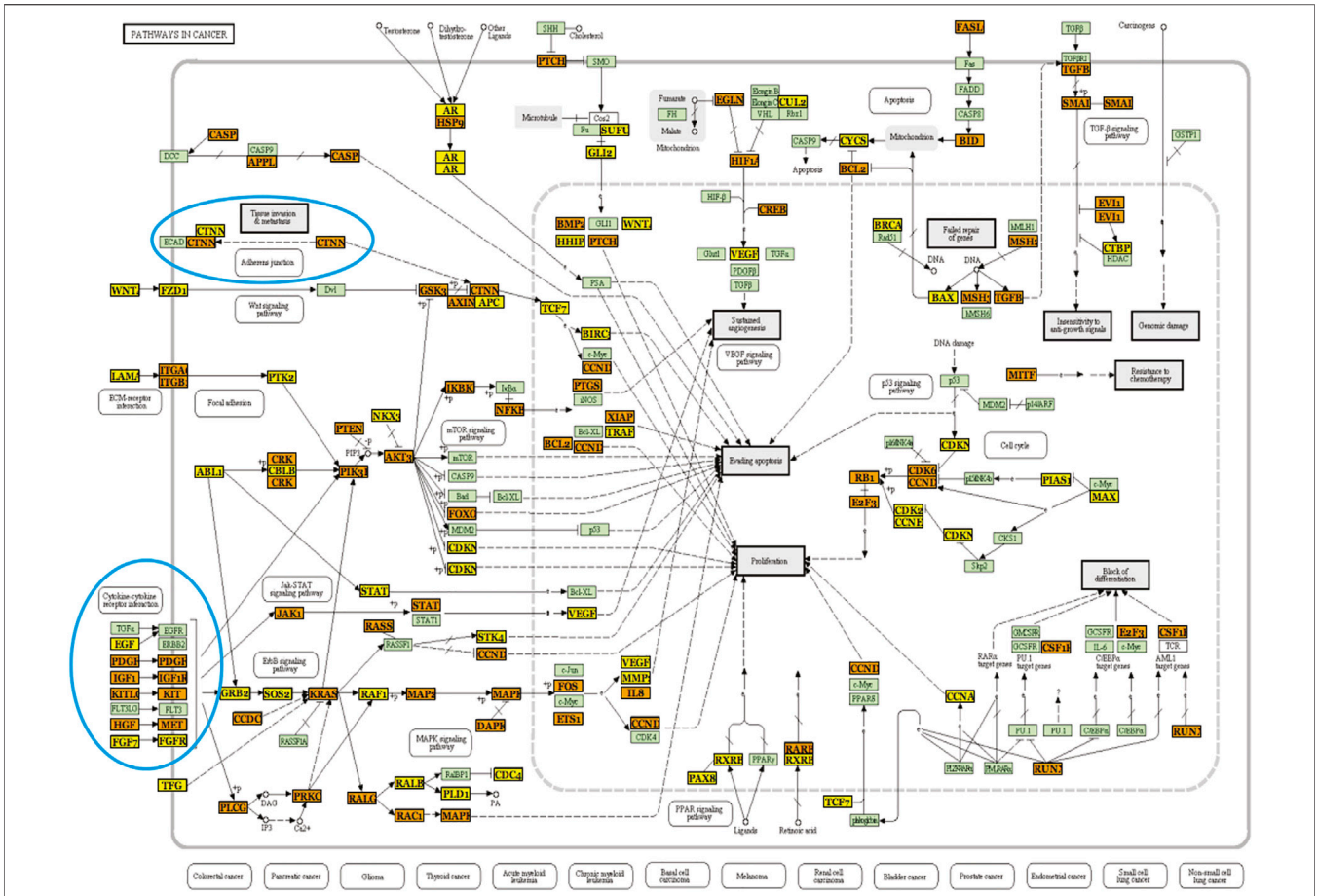


FIGURE 4 | A wiring diagram showing pathways in cancer and their regulation by microRNAs relevant in alcohol abuse. The diagram shows a network of several pathways important in the development of cancer. Gene products in yellow boxes are affected by one microRNA, gene products in orange boxes are affected by more than one microRNA. Gene products in green boxes are unaffected. Elements of pathways in blue circles were tested by qPCR (**Figure 2**). Wiring diagrams of specific cancers are available in the Supplementary Materials. The wiring diagram was created using KEGG (Kyoto Encyclopedia of Genes and Genomes) and DIANA-mirPath software with FDR correction and *p*-value threshold of 0.05.

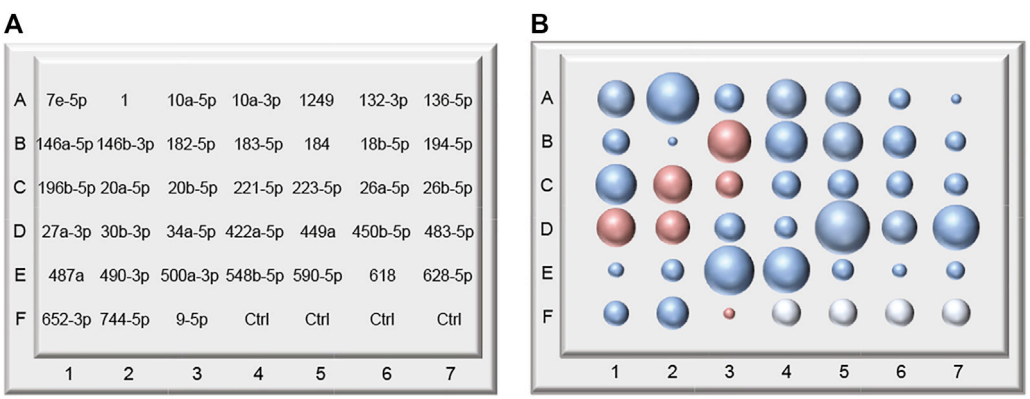


FIGURE 5 | A hypothetical model of an alcohol abuse microRNA biomarker chip. **(A)** The chip contains a panel of 38 microRNAs characteristic of alcohol abuse. Four spots are designated to controls (bottom, right corner). **(B)** Results characteristic for alcohol abuse. The size of a sphere corresponds to the change in microRNA expression. Upregulated microRNAs are shown as blue spheres larger than controls, downregulated spheres are shown as spheres smaller than controls. The red color depicts inflammatory microRNAs.

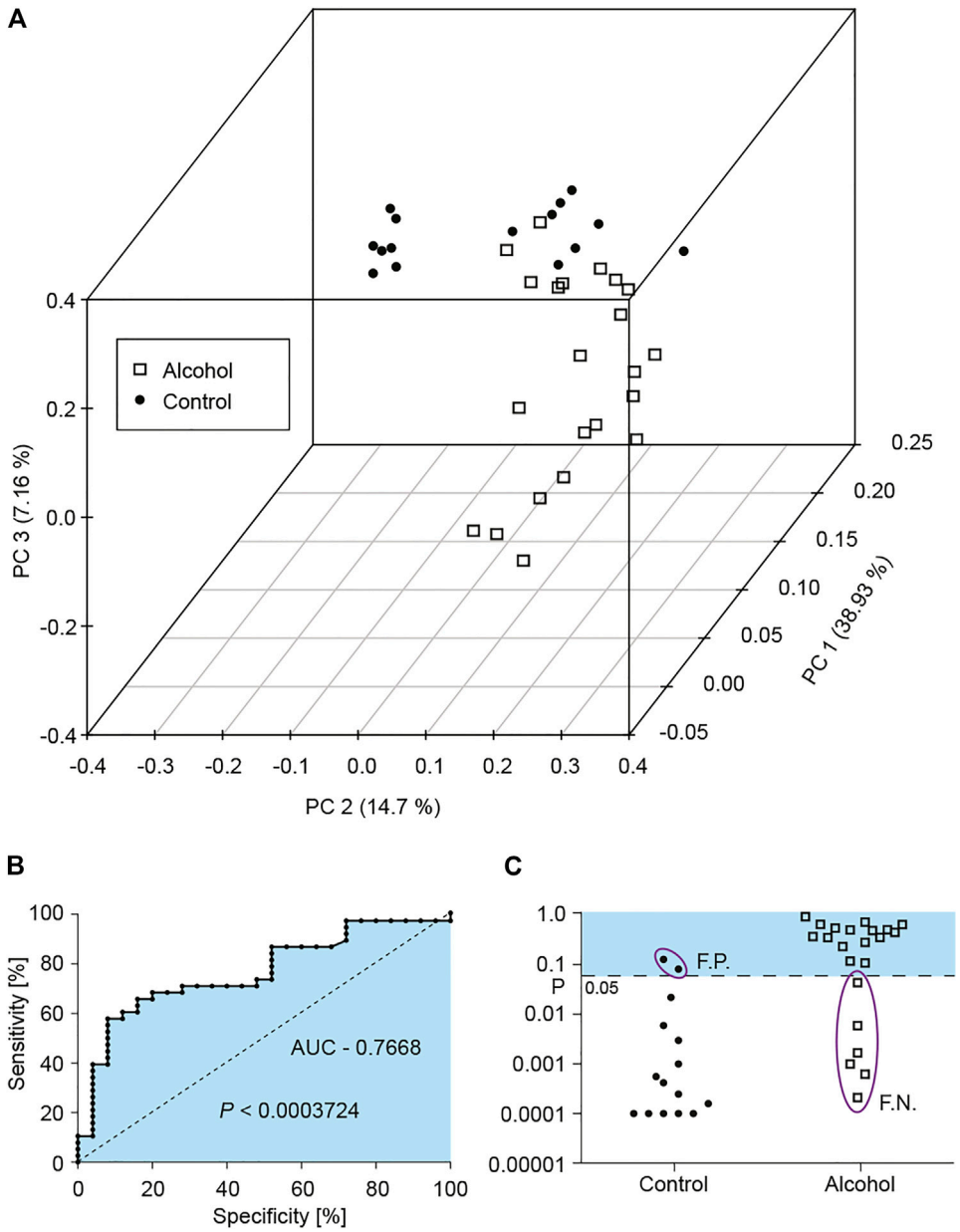


FIGURE 6 | Efficacy of the current microRNA panel to segregate alcohol abusers and non-abusing controls. **(A)** 3D graph of Principal Component Analysis (PCA, PC1—38.93%, PC2—14.7%, and PC3—7.16% of the total variance) shows that most alcohol and control individuals can be classified into separate clouds with some overlap. **(B)** A Receiver Operating Curve (ROC) of alcohol-regulated microRNA panel indicates the ability of the panel to distinguish between the alcohol group and the control group: the area under the ROC curve (AUC, blue) is significantly above random sampling: 0.7668 ($p < 0.0003724$). **(C)** The fit of each individual to an overall alcohol-regulated microRNA panel is shown as a p-value of departure of individual ROC curve from the panel's ROC with $p = 0.05$ used as a cut-off for significance (dotted line). If $p \geq 0.05$ the microRNA panel determines the individual as an alcohol abuser, if $p < 0.05$ the microRNA panel classifies the individual as a non-abuser. False positives and false negative results are circled.

prognostic and pharmacological value in the course of a development of an ailment. The field of biomarker development is rapidly growing (see Nanau and Neuman, 2015; Rosato et al., 2019). Development of biomarkers in AUD would help to determine an advancement of the disease, early detection of its complications, and/or efficacy of treatment.

Here, in our studies, we compared heavy, chronic abusers of alcohol with at least ten-year long history of alcohol abuse, with a matched non-alcohol abusing control group, to determine whether chronic excess of alcohol changes expression of miRNAs in saliva, and whether these changes have a potential to serve as a biomarker of alcohol abuse.

microRNAs are attractive biomarker candidates. Expression of microRNAs in saliva can change with a disease progression (Rosato et al., 2019; He et al., 2020), and they are more stable in saliva than mRNAs (Pritchard et al., 2012; Han et al., 2018) reducing variability and improving accuracy. Indeed, microRNAs in saliva have been used as or theorized as useful biomarkers in oral and pancreatobiliary tract cancer, as well as head and neck squamous cell carcinoma (He et al., 2020; Machida et al., 2016; Langevin et al., 2017). Simultaneous measurement of expression of several microRNAs (microRNA profiling) provides a more powerful diagnostic tool (Jones et al., 2012; Rosato et al., 2019; Mahnke et al., 2021), therefore we screened the expression of several hundred microRNAs to determine a unique microRNA signature of chronic alcohol exposure and its potential as a biomarker of alcohol abuse.

Due to our stringent inclusion/exclusion criteria the enrolled number of individuals (38) was rather small but sufficient to see significant differences between two studied groups. We excluded many individuals with a known history of using any drugs of abuse including nicotine, which is very frequently co-abused with alcohol. The main experimental variable between two groups studied here was alcohol consumption, and we feel confident that observed differences are most likely attributable to chronic alcohol exposure. Ten to fifteen samples have been shown through statistical analysis to be sufficient for the stable replication of findings in microarrays (Pavlidis et al., 2002), and to perform microRNA profiling in diagnosis of cancer (Jones et al., 2012), supporting the validity of our studies. Moreover, the observed 85.5% concordance with controls (n = 18) of Hanson et al. (2009), 91.3% concordance with controls (n = 12) of Park et al. (2009), and 100% concordance with controls (n = 20) of Patel et al. (2011) reports indicating that our control microRNA profiles can accurately represent microRNA expression in healthy individuals.

It is currently accepted that age, gender and ethnicity play a role in the development of AUD (Gupta et al., 2018; National Institutes of Health, 2021a) and in discovery of biomarkers (Tarallo et al., 2018; Rosato et al., 2019). However, the progress of human civilization now blurs ethnic, gender and age divisions, which thus become less and less relevant. Our data indicate that microRNA profiling in saliva could be a reliable biomarker of chronic alcohol consumption with rather strong predictive power, regardless of age, gender and ethnicity.

Saliva is a very promising source of biomarkers. Some commercially available kits (Genotek) ensure immediate stabilization of RNA after collection of saliva. The presence of bacteria in the oral cavity does not bias the results of microRNA profiling, as many oral bacteria are not disrupted during human microRNA isolation process due to thick bacterial walls, and the human microRNA microarrays are specific for human microRNAs. Moreover, the collected samples can be designated as non-biohazard material, since the samples are self- collected, immediately placed in sealed containers, and exposed to a SARS-CoV-2 virus inactivating agent, which causes a >99% reduction in infective virus (DNA Genotek, 2021). Nevertheless, we recommend taking further precautions during collection and shipping samples following local, national and international regulations, including the current edition of the International Air Transport Association (IATA) Dangerous Goods Regulations.

Continuation of this study with various AUD groups can likely help further hone the microRNA biomarker panels, as more miRNAs related to alcohol exposure are being identified in saliva (Rosato et al., 2019). microRNA biomarkers could be also combined with other, non-microRNA alcohol biomarkers. For example, alcohol metabolites are currently considered as alcohol exposure biomarkers. Two of the most promising alcohol metabolites biomarkers, are carbohydrate-deficient transferrin (CDT), and ethyl glucuronide (EtG) (Litten et al., 2010; Nanau and Neuman, 2015). Both are highly sensitive and specific to alcohol intake, however require collection of bio-hazardous blood, and the use of rather expensive detection techniques present only in specialized laboratories: high performance liquid chromatography (HPLC) and liquid chromatography-mass spectroscopy (LC-MS), respectively (Litten et al., 2010; Nanau and Neuman, 2015). microRNA detection on the other hand is based on PCR methodology, which became recently easily-available in many first-contact diagnostic centers due to the COVID-19 pandemic.

Genotyping can also complement the role of microRNA biomarkers in AUD. Characterization of an alcoholism-susceptible genotype can provide "static" information of potential vulnerability to the development of alcoholism, while microRNA profiling can permit "dynamic" monitoring of the disease status, escalation of alcohol consumption, development of tolerance, continuity of abstinence or occurrence of relapse.

Since microRNAs are very important regulators of gene expression, profiling of microRNA could also provide insight into molecular mechanisms of alcohol actions on cell biology. Thus, to understand biological consequences of microRNA dysregulation in saliva we first determined that the most prominent cellular source of miRNAs are buccal cells and leukocytes. Although we cannot exclude the contribution of exosomal microRNA to the profiling, as exosomal miRNAs are present in saliva (Hu et al., 2012; Wong 2015), many of them originate from oral cavity cells such as the salivary glands (Han et al., 2018; He et al., 2020), making them useful indicators of oral, if not general, health.

Analysis of miRNAs in relation to associated signaling pathways revealed insights into disease mechanisms. Almost all of the top cancers suggested by bioinformatic analysis in this study originate from epithelial tissues (**Table 3**). There is some evidence that alcohol could contribute to the pathogenesis of these cancers (Vagst et al., 2003; Gupta et al., 2010; Michaud et al., 2010; Fedirko et al., 2011; Morris et al., 2015). However, none of the enrolled patients in our study has been diagnosed yet with these cancers. This might be because the median age of diagnosis of these cancers falls into the sixth decade of life (Between 57 years old for glioma to 71 years old for pancreatic cancer) according to the National Cancer Institute (Howlander et al., 2012). It takes a few decades for a transformation of a normal cell into a detectable cancer according to the multistage clonal expansion carcinogenesis model (Luebeck et al., 2012).

During this period cancer remains undiagnosed (Luebeck et al., 2012), though improvements to modeling continue to shave down the delay (Avanzini et al., 2020). Interestingly, chronic alcohol consumption can lower the age of diagnosis from five (colorectal cancer, Zisman et al., 2006) to ten years (pancreatic cancer, Anderson et al., 2012). It is tempting to speculate that changes in microRNA expression could be a very early indication of de-regulation of biological processes, which could potentially lead to cancer in later years, thus making microRNA profiling an early biomarker of cancer development.

Comparison between our work and a salivary biomarker study in the literature for AUD (Rosato et al., 2019) found hsa-miR-10a-5p and miR-27a-5p as significantly upregulated in both studies. Alterations in miR-10a expression have been important in multiple types of cancers, likely through its role in the regulation of p53 (Ovcharenko et al., 2011; Bryant et al., 2012; Vu et al., 2021), a highly conserved, critical tumor suppressor gene. As a result, miR-10a has proven to be an important cancer biomarker (Vu et al., 2021). Given that alcohol abuse elevates the risk of many cancers (Fidler, 2003; International Agency for Research on Cancer, 2010; Wong et al., 2011; Zakhari et al., 2011; Miranda, 2014), it seems plausible that the dysregulation of miR-10a in alcohol abusers likewise leads to dysregulation of the p53 network in AUD individuals, underscoring how alcohol may act to elevate cancer risk. Likewise, miR-27a plays a role in tumorigenesis in multiple cancers (Li et al., 2019; Zhang et al., 2019; Cui et al., 2020).

We used a blend of bioinformatics (DIANA-mirPath, RNAhybrid) and wet-lab experiments (qPCR) to predict global impact of alcohol-regulated microRNAs on biological pathways and cellular processes. As recently suggested, targeting of transcripts by microRNA remains the same in the presence of alcohol (Miranda, 2012). We determined that several biological pathways could be affected, and we focused on the adherens junction pathway because adherens junctions are key for cell-cell interactions, act as contact points between cells and maintain tissue integrity. Changes in adherens junctions can cause changes in gene expression affecting cell adhesion and movement (Meadows and Zhang, 2015). Importantly, alcohol, by impairing adherens junction pathway, disrupts endothelial integrity increasing probability of bacterial infections (Wood et al., 2013) and metastatic cancer (Xu et al., 2012; Madoz-Gúrpide et al., 2015). The essential element of the adherens junction is transplasmalemmal cadherin-catenin complex, responsible directly for cell adhesion. We observed that alcohol-regulated microRNAs, could contribute to alcohol impairment of the adherens junction by targeting this complex in at least two ways. We detected significant downregulation of catenin transcripts, as well as transcripts encoding SNAil1, a transcription factor controlling expression of cadherin. Both types of molecules were targeted by several upregulated microRNAs.

The most downregulated element of the adherens junction and a target of alcohol-upregulated microRNAs was MET, a receptor tyrosine kinase, which plays a key role in c-Met signal transduction pathway. Activation of MET leads to multiple, diverse biological effects, while its de-regulation contributes to tumor progression and metastasis (Ma et al., 2003; Puccini et al., 2019). In has been shown that alcohol decreases expression of catenin in human neuronal stem cells (Vangipuram and Lyman, 2012) as well as its function in human bone marrow (Yeh et al., 2008), however increases catenin expression in the superior frontal cortex of chronic alcoholics (Al-Housseini et al., 2008), and in the hippocampal neurons of rats (Velázquez-Marrero et al., 2016). MET expression is not affected by alcohol in Chang (HeLa) liver cells [Ohira et al., 1995; Hall review discusses that Chang liver cells are HeLa cells (Hall, 2017)]. This is the first report indicating downregulated expression of catenin, SNAIL1 and MET in heavy alcohol drinkers. Future studies will be of importance to determine the exact contribution of de-regulation of these molecules by alcohol-regulated microRNAs to the disruption of cell adherence or increased tumor metastasis.

Interestingly, downregulated catenin is also a part of the Wnt signaling pathway, which was another mirPath top "hit". Other potentially de-regulated pathways included mitogen-activated protein kinase (MAPK), ErbB signaling and endocytosis. MAPK and ErbB signaling pathways are both part of the cellular responses to an array of stimuli, and are involved in transfer of signaling to the nucleus and regulation of gene expression in cellular proliferation, migration or differentiation. Endocytosis is a ubiquitous cellular mechanism allowing delivery of plasmalemmal elements and vesicular cargo into cellular milieu. Alcohol disrupts Wnt signaling (Lauing et al., 2012; Vangipuram and Lyman, 2012; Kapania et al., 2020), changes phosphorylation of p38 MAPK (Gu et al., 2013; Xu et al., 2016) and ErbB (Srivastava et al., 2011; Xu et al., 2016), and impairs endocytosis (Marin et al., 2010). Our results suggest that alcohol de-regulation of microRNAs could be, at least partially, involved in these perturbations.

Impairment of these pathways in epithelial cells can lead to important pathologies. Loss of adherence of the epithelial cells to the surrounding environment is a known step in the epithelial-mesenchymal transformation (EMT) towards malignancy. Indeed, chronic alcohol exposure increases occurrence of oral cancer (Pelucchi et al., 2011) originating from squamous epithelial buccal mucosa cells (Johnson et al., 2011; Leite et al., 2021).

Further functional studies are necessary to definitively indicate that de-regulation of biological pathways by alcohol is microRNA-dependent. Such studies should consider the validation of microRNA targets. We used several criteria to validate some of the microRNA targets in the adherens junction pathway. In one of the top pathways indicated by mirPath enrichment analysis, we have chosen several elements of the selected pathway based on their role in the pathway, cellular location, targeting by multiple microRNAs and regulation by microRNAs, which expression was changed in the same direction (all microRNA upregulated). Next, we determined the probability of microRNAs interaction with these targets by an independent program (RNAhybrid), considering presence of multiple splice variants and 3′UTR heterogeneity of selected targets. We observed that one of the targets indicated by mirPath had hybridization energy above −20 kcal/mol, and its levels were not changed as determined by qPCR. The other targets were significantly downregulated as predicted. These results indicate

that some genes in mirPath-selected pathways can be false positive.

Proper function of leukocytes also depends on cell motility and adherence. Impairment of these cellular properties by alcohol-regulated microRNAs in leukocytes could be one factor contributing to the known immunosuppressive effects of alcohol (Cook, 1998; Szabo, 1999; Messingham et al., 2002). microRNA regulation of the immune system also plays important role in regulation of inflammatory networks (O'Connell et al., 2012; Morris et al., 2015; Ureña-Peralta et al., 2018). Alcohol is known for its pro-inflammatory actions. Moreover, carcinogenic effects of alcohol are also thought to take place *via* modulation of inflammatory responses (Seitz and Stickel, 2007; Qin et al., 2008; Morris et al., 2015). We observed that out of 38 signature microRNAs, seven are involved in inflammation: miR-20a-5p, miR-20b-5p, miR-27a-3p, miR-30b-5p, miR-182-5p, miR-183-5p, and miR-9-5p. Specifically, miR-27a-3p and miR-30b-5p expression changes as a response to an inflammatory stimulus (Chang et al., 2012; Graff et al., 2012), while expression of miR-182-5p is altered in steatohepatitis (Dolganiuc et al., 2009). More recently, miR-182-5p and 183-5p, in the miR-183 cluster, were found to be dysregulated in mice exposed to chronic alcohol exposure (Osterndorff-Kahanek et al., 2018; Ureña-Peralta et al., 2018), a particularly relevant model for AUD patients. Both miR-182-5p and miR-183-5p regulate neuroinflammatory pathways involved in Toll-like receptor 4 (TLR-4) signaling, critical in innate immunity (Ureña-Peralta et al., 2018). Interestingly, miR-27a has been recently found to be upregulated in blood plasma from AUD individuals, underscoring its utility in signaling alcohol abuse (Saha et al., 2016).

Let-7, a miRNA in our study, has also been observed in inflammatory processes to act *via* TLR-7 regulation in alcoholism (Coleman et al., 2017; Crews et al., 2017). miR-155 and -199, while not observed in our studies, were implicated in studies of rat liver cells and endothelial cells in humans in response to chronic alcohol insult, and lead to inflammation in cirrhosis through endothelin-1 (ET-1) and hypoxia-inducible factor-1α (HIF-1α) (Yeligar et al., 2009). In a recent report of miRNA in immune cells from chronic inflammatory states (Salvi et al., 2019), several of the miRNA (miR-20b, -26a, -223) we found to be differentially regulated in our alcohol abusing population were also found and observed to be regulated in the same direction as in chronic inflammatory states. We found that let-7e was also regulated in the same direction, though the same target was not predicted for let-7e, but instead for the closely related let-7d, which likely overlaps many of the same targets. Most of these (all except miR-26a) were predicted by TargetScan 8.0 (McGeary et al., 2019) to have potential targets the same as known targets of miRNAs regulated in chronic inflammatory states primarily in immune cells as reported in Salvi et al. (2019). This suggests that these miRNAs are playing a similar role in the pathology of inflammation seen in chronic alcohol abuse.

miR-9 is a particularly interesting microRNA involved in alcohol actions and inflammation (Miranda, 2014). We have previously reported that acute alcohol (15 min exposure) causes upregulation of miR-9 (current nomenclature: miR-9-5p) and the development of molecular tolerance in neurons (Pietrzykowski et al., 2008). Alcohol exposure in fetal rats has been shown to upregulate miR-9, leading to dysregulation of pituitary dopamine D2 receptors (Gangisetty et al., 2017). A report by Sathyan et al. (2007) showed downregulation of miR-9 upon 24 h alcohol exposure in neurospheres. Our current results indicate that, in chronic alcohol abuse, miR-9 levels are downregulated in leukocytes and/or epithelial cells. This follows a usual pattern seen with addiction and tolerance; that is, abuse leads to a change in expression of a gene, but repeated or chronic abuse leads to tolerance and expression change in the opposite direction as the body tries to maintain homeostasis in the face of repeated spikes of the offending drug (Volkow and Morales, 2015; Cahill et al., 2016). Bazzoni et al. (2009), describe an inflammatory role of miR-9 in leukocytes. Together, these findings indicate that alcohol-sensitive microRNAs, like miR-9, could have a drastically different response depending on the length or frequency of alcohol exposure. Moreover, understanding temporal regulation of miR-9 and other inflammatory microRNAs by alcohol is of great interest in AUD, which is a chronic disease with an important inflammatory component (Kelley and Dantzer, 2011; Ureña-Peralta et al., 2018).

AUD affects chronic inflammation, at least partially, *via* increasing oxidative stress (Bala and Szabo, 2012). Alcohol consumption and metabolism leads to Reactive Oxygen Species (ROS) production (Rao, 2008; Szabo and Bala, 2010; Bala and Szabo, 2012), instrumental to pathological developments in Alcoholic Liver Disease (ALD) (Keshavarzian et al., 2009; Szabo and Bala, 2010). Several miRNAs, miR-27a-3p, miR-34a, miR-223, found to be differentially expressed in alcohol abusers in our study, were found to be differentially regulated in other physiological studies of oxidative stress (Bala and Szabo, 2012; Klieser et al., 2019). In particular, macrophages may be regulated by miRNAs in response to oxidative stress through the NF-κB pathway (Thulasingam et al., 2011).

Exposure to alcohol affects several inflammatory cascades including production of pro-inflammatory cytokines by many cells throughout the body (Qin et al., 2008). miRNAs can regulate pro-inflammatory cytokine pathways in chronic alcohol abuse (Coleman et al., 2017; Crews et al., 2017; Salvi et al., 2019). Several reports of individual microRNAs present on our panel describe microRNA regulation of pro-inflammatory cytokines, which can both attenuate or enhance cytokine production or secretion. Selected representatives (one per each of our three distinct microRNA groups) of such regulation include: secretion of pro-inflammatory cytokines is decreased by miR-20b targeting VEGF (Song C. et al., 2012), miR-20a targeting ASK1 (Philippe et al., 2013) and TLR4 (Li et al., 2020), and miR-183 targeting TGFa (Tao et al., 2021). On the other hand, secretion of pro-inflammatory cytokines is increased by miR-132-3p targeting TRIB1 (Niespolo et al.,. 2020), miR-223 targeting IKKa (Valmiki et al., 2020) or miR-183-5p targeting PPP2CA (Guo et al., 2020). These few examples underscore the importance of performing a comprehensive screen of multiple microRNAs to better

understand the functional outcomes of microRNA regulation of biological pathways.

Overall, our data help strengthen the pathogenic link between alcohol, inflammation and malignancies such as cancer, providing attractive plausible molecular mechanisms involving specific microRNAs. In view of the non-invasive nature of sample collection, suggestive involvement of de-regulated microRNAs in alcohol-related pathologies and potential use of these microRNAs as an alcohol biomarker, a continuation of this study on a larger group of patients is of great interest. Obtaining samples from patients at different stages of AUD could allow the establishment of fine-tuned correlation between disease states, specific microRNAs and their patterns of modulation. The current study helped to narrow down alcohol-relevant microRNAs in humans to 38 microRNA species using two end-point groups: sporadic drinkers (controls) and heavy drinkers. A power analysis is a possible method to reliably estimate the number of subjects (n) (Pavlidis, Li, and Noble, 2002; Ching et al., 2014). However, power analysis typically requires prior knowledge of the topic under bioinformatics investigation (Pavlidis, Li, and Noble, 2002). Our study, performed on two end-point groups, could provide this information and should aid in performing power analysis to define the number of subjects necessary to determine the usefulness of the microRNA panel as a biomarker in various alcohol-drinking groups. Our results suggest also that changes in the expressions of miRNAs in the saliva of alcohol abusers could be used to monitor the effectiveness of any therapeutic approaches, including medications and lifestyle.

In summary, this study underscores that measuring changes in the expression of microRNAs in saliva may have predictive value in determination of alcohol abuse and its consequences. We also demonstrate the potential feasibility and provide direction for establishing microRNA profiling in saliva as a sensitive and specific tool in determination of chronic, excessive alcohol consumption. Further studies are needed to establish microRNA profiling as a biomarker of various stages of the alcohol use disorders and its sequelae.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Capital Health Hospital System, and Rutgers State University of New Jersey. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

NB-P, YW, and AP were responsible for the study concept and design. NB-P performed the initial screen of patients and acquired human samples. YW, OA and NK prepared samples for and performed array experiments. EM performed the real-time PCR experiments. YW and NK performed staining experiments. YW and ML performed FACS experiments and analysis. EM, NK, SH and AP analyzed PCR results. YF and AP performed statistical analysis. NB-P, EM, YW, and AP performed overall data analysis, interpretation of findings and drafted the initial manuscript. EM and AP analyzed and drafted the revised manuscript. All authors critically reviewed content and approved the final version for publication.

FUNDING

This work was supported in part by NIH grants AA017920 and AA01748 to AP.

ACKNOWLEDGMENTS

We would like to thank Frank Natale for help with the flow cytometry experiments, Rutgers Core Facility for help with microarrays, and Austin Thekkumthala for assistance on the saliva images. Special thanks go to Rajesh Miranda and the Department of Neuroscience and Experimental Therapeutics, College of Medicine, Texas A and M University for invaluable help with reviewing the manuscript and support of the publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.804222/full#supplementary-material

REFERENCES

Al-Housseini, A. M., Sivanandam, T. M., Bradbury, E. L., Tannenberg, R. K., Dodd, P. R., and Gu, Q. (2008). Upregulation of β-Catenin Levels in Superior Frontal Cortex of Chronic Alcoholics. *Alcohol. Clin. Exp. Res.* 32, 1080–1090. doi:10.1111/j.1530-0277.2008.00670.x

Anderson, M. A., Zolotarevsky, E., Cooper, K. L., Sherman, S., Shats, O., Whitcomb, D. C., et al. (2012). Alcohol and Tobacco Lower the Age of Presentation in Sporadic Pancreatic Cancer in a Dose-dependent Manner: a Multicenter Study. *Am. J. Gastroenterol.* 107, 1730–1739. doi:10.1038/ajg.2012.288

Avanzini, S., Kurtz, D. M., Chabon, J. J., Moding, E. J., Hori, S. S., Gambhir, S. S., et al. (2020). A Mathematical Model of ctDNA Shedding Predicts Tumor Detection Size. *Sci. Adv.* 6 (50), eabc4308. doi:10.1126/sciadv.abc4308

Bala, S., and Szabo, G. (2012). MicroRNA Signature in Alcoholic Liver Disease. *Int. J. Hepatol.* 2012, 1–6. doi:10.1155/2012/498232

Bazzoni, F., Rossato, M., Fabbri, M., Gaudiosi, D., Mirolo, M., Mori, L., et al. (2009). Induction and Regulatory Function of miR-9 in Human Monocytes and

Neutrophils Exposed to Proinflammatory Signals. *Proc. Natl. Acad. Sci.* 106, 5282–5287. doi:10.1073/pnas.0810909106

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodological)* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bischoff, P., Kornhuber, M., Dunst, S., Zell, J., Fauler, B., Mielke, T., et al. (2020). Estrogens Determine Adherens Junction Organization and E-Cadherin Clustering in Breast Cancer Cells via Amphiregulin. *Iscience* 23 (11), 101683. doi:10.1016/j.isci.2020.101683

Bryant, A., Palma, C. A., Jayaswal, V., Yang, Y. W., Lutherborrow, M., and Ma, D. D. (2012). miR-10a Is Aberrantly Overexpressed in Nucleophosmin1 Mutated Acute Myeloid Leukaemia and its Suppression Induces Cell Death. *Mol. Cancer* 11, 8. doi:10.1186/1476-4598-11-8

Cahill, C. M., Walwyn, W., Taylor, A. M. W., Pradhan, A. A. A., and Evans, C. J. (2016). Allostatic Mechanisms of Opioid Tolerance beyond Desensitization and Downregulation. *Trends Pharmacol. Sci.* 37 (11), 963–976. doi:10.1016/j.tips.2016.08.002

Centers for Disease Control and Prevention (2020). Alcohol and Public Health: Alcohol-Related Disease Impact (ARDI). Annual Average for United States 2011–2015 Alcohol-Attributable Deaths Due to Excessive Alcohol Use, All Ages. Available at: https://nccd.cdc.gov/DPH_ARDI/Default/Default.aspx (Accessed July 13, 2021).

Chang, C.-C., Zhang, Q.-Y., Liu, Z., Clynes, R. A., Suciu-Foca, N., and Vlad, G. (2012). Downregulation of Inflammatory MicroRNAs by Ig-like Transcript 3 Is Essential for the Differentiation of Human CD8+ T Suppressor Cells. *J.I.* 188, 3042–3052. doi:10.4049/jimmunol.1102899

Ching, T., Huang, S., and Garmire, L. X. (2014). Power Analysis and Sample Size Estimation for RNA-Seq Differential Expression. *RNA* 20 (11), 1684–1696. doi:10.1261/rna.046011.114

Cichon, C., Sabharwal, H., Rüter, C., and Schmidt, M. A. (2014). MicroRNAs Regulate Tight junction Proteins and Modulate Epithelial/endothelial Barrier Functions. *Tissue Barriers* 2 (24), e944446. doi:10.4161/21688362.2014.944446

Coleman, L. G., Zou, J., and Crews, F. T. (2017). Microglial-derived miRNA Let-7 and HMGB1 Contribute to Ethanol-Induced Neurotoxicity via TLR7. *J. Neuroinflammation* 14 (1), 1422. doi:10.1186/s12974-017-0799-4

Cook, R. T. (1998). Alcohol Abuse, Alcoholism, and Damage to the Immune System-A Review. *Alcohol. Clin. Exp. Res.* 22, 1927–1942. doi:10.1111/j.1530-0277.1998.tb05900.x

Coussens, L. M., and Werb, Z. (2002). Inflammation and Cancer. *Nature* 420, 860–867. doi:10.1038/nature01322

Crews, F. T., Lawrimore, C. J., Walter, T. J., and Coleman, L. G., Jr. (2017). The Role of Neuroimmune Signaling in Alcoholism. *Neuropharmacology* 122 (122), 56–73. doi:10.1016/j.neuropharm.2017.01.031

Cui, M., Yao, X., Lin, Y., Zhang, D., Cui, R., and Zhang, X. (2020). Interactive Functions of microRNAs in the miR-23a-27a-24-2 Cluster and the Potential for Targeted Therapy in Cancer. *J. Cel Physiol.* 235 (1), 6–16. doi:10.1002/jcp.28958

de Haan, J. R., Wehrens, R., Bauerschmidt, S., Piek, E., Schaik, R. C. v., and Buydens, L. M. C. (2007). Interpretation of ANOVA Models for Microarray Data Using PCA. *Bioinformatics* 23, 184–190. doi:10.1093/bioinformatics/btl572

Dippold, R. P., Vadigepalli, R., Gonye, G. E., Patra, B., and Hoek, J. B. (2013). Chronic Ethanol Feeding Alters miRNA Expression Dynamics during Liver Regeneration. *Alcohol. Clin. Exp. Res.* 37 (Suppl. 1Suppl 1), E59–E69. doi:10.1111/j.1530-0277.2012.01852.x

DNA Genotek Inc (2021). Inactivation of SARS-CoV-2 in Samples Collected Using Oragene, ORAcollect and OMNIgene Products from DNA Genotek. Available at: https://res.cloudinary.com/crowdicity-eu-cld/image/upload/MK-01430_inactivation_d0amyf.

Dolganiuc, A., Petrasek, J., Kodys, K., Catalano, D., Mandrekar, P., Velayudham, A., et al. (2009). MicroRNA Expression Profile in Lieber-DeCarli Diet-Induced Alcoholic and Methionine Choline Deficient Diet-Induced Nonalcoholic Steatohepatitis Models in Mice. *Alcohol. Clin. Exp. Res.* 33, 1704–1710. doi:10.1111/j.1530-0277.2009.01007.x

Fedirko, V., Tramacere, I., Bagnardi, V., Rota, M., Scotti, L., Islami, F., et al. (2011). Alcohol Drinking and Colorectal Cancer Risk: an Overall and Dose-Response Meta-Analysis of Published Studies. *Ann. Oncol.* 22, 1958–1972. doi:10.1093/annonc/mdq653

Ferreira, M. P., and Willoughby, D. (2008). Alcohol Consumption: the Good, the Bad, and the Indifferent. *Appl. Physiol. Nutr. Metab.* 33, 12–20. doi:10.1139/H07-175

Fidler, I. J. (2003). The Pathogenesis of Cancer Metastasis: the 'seed and Soil' Hypothesis Revisited. *Nat. Rev. Cancer* 3, 453–458. doi:10.1038/nrc1098

Gambhir, R. S., Gupta, T., Kl, V., Teja, T. S., Kaur, D., and Gambhir, S. (2018). Salivary Parameters Among Chronic Alcoholics and Non-alcoholic Males in Ambala, Haryana - a Comparative Study. *Jdhodt* 9 (5), 377–380. doi:10.15406/jdhodt.2018.09.00412

Gangisetty, O., Jabbar, S., Wynne, O., and Sarkar, D. K. (2017). MicroRNA-9 Regulates Fetal Alcohol-Induced Changes in D2 Receptor to Promote Prolactin Production. *J. Endocrinol.* 235 (1), 1–14. doi:10.1530/JOE-17-0135

Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A., and Bartel, D. P. (2011). Weak Seed-Pairing Stability and High Target-Site Abundance Decrease the Proficiency of Lsy-6 and Other microRNAs. *Nat. Struct. Mol. Biol.* 18, 1139–1146. doi:10.1038/nsmb.2115

Gilad, S., Meiri, E., Yogev, Y., Benjamin, S., Lebanony, D., Yerushalmi, N., et al. (2008). Serum microRNAs Are Promising Novel Biomarkers. *PLoS One* 3, e3148. doi:10.1371/journal.pone.0003148

Graff, J. W., Dickson, A. M., Clay, G., McCaffrey, A. P., and Wilson, M. E. (2012). Identifying Functional microRNAs in Macrophages with Polarized Phenotypes. *J. Biol. Chem.* 287, 21816–21825. doi:10.1074/jbc.M111.327031

Grant, B. F., Harford, T. C., Muthén, B. O., Yi, H.-y., Hasin, D. S., and Stinson, F. S. (2007). DSM-IV Alcohol Dependence and Abuse: Further Evidence of Validity in the General Population☆. *Drug and Alcohol Dependence* 86, 154–166. doi:10.1016/j.drugalcdep.2006.05.019

Greten, F. R., and Grivennikov, S. I. (2019). Inflammation and Cancer: Triggers, Mechanisms, and Consequences. *Immunity* 51 (1), 27–41. doi:10.1016/j.immuni.2019.06.025

Gruenberg, E. M. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. Washington, D.C.: American Psychiatric Association.

Gu, L., Fink, A. M., Chowdhury, S. A. K., Geenen, D. L., and Piano, M. R. (2013). Cardiovascular Responses and Differential Changes in Mitogen-Activated Protein Kinases Following Repeated Episodes of Binge Drinking. *Alcohol Alcohol.* 48, 131–137. doi:10.1093/alcalc/ags090

Guo, J., Duan, Z., Zhang, C., Wang, W., He, H., Liu, Y., et al. (2020). Mouse 4T1 Breast Cancer Cell-Derived Exosomes Induce Proinflammatory Cytokine Production in Macrophages via miR-183. *J.I.* 205 (10), 2916–2925. doi:10.4049/jimmunol.1901104

Gupta, S., Wang, F., Holly, E. A., and Bracci, P. M. (2010). Risk of Pancreatic Cancer by Alcohol Dose, Duration, and Pattern of Consumption, Including Binge Drinking: a Population-Based Study. *Cancer Causes Control* 21, 1047–1059. doi:10.1007/s10552-010-9533-6

Hall, E. (2017). ICLAC Case Study: Chang Liver, Version 1.3. "Case Study: The Notorious Chang Liver". *ICLAC Sept* 28, 20171–20174.

Han, Y., Jia, L., Zheng, Y., and Li, W. (2018). Salivary Exosomes: Emerging Roles in Systemic Disease. *Int. J. Biol. Sci.* 14 (6), 633–643. doi:10.7150/ijbs.25018

Hanson, E. K., Lubenow, H., and Ballantyne, J. (2009). Identification of Forensically Relevant Body Fluids Using a Panel of Differentially Expressed microRNAs. *Anal. Biochem.* 387, 303–314. doi:10.1016/j.ab.2009.01.037

He, L., Ping, F., Fan, Z., Zhang, C., Deng, M., Cheng, B., et al. (2020). Salivary Exosomal miR-24-3p Serves as a Potential Detective Biomarker for Oral Squamous Cell Carcinoma Screening. *Biomed. Pharmacother.* 121, 109553. doi:10.1016/j.biopha.2019.109553

Howlader, N., Noone, A. M., Krapcho, M., Neyman, N., Aminou, R., Altekruse, S. F., et al. (Editors) (2012). *Data from: SEER Cancer Statistics Review, 1975-2009 (Vintage 2009 Populations)* (Bethesda: National Cancer Institute). Available at: https://seer.cancer.gov/archive/csr/1975_2009_pops09/.

Hu, G., Yao, H., Chaudhuri, A. D., Duan, M., Yelamanchili, S. V., Wen, H., et al. (2012). Exosome-mediated Shuttling of microRNA-29 Regulates HIV Tat and Morphine-Mediated Neuronal Dysfunction. *Cell Death Dis.* 3, e381. doi:10.1038/cddis.2012.114

International Agency for Research on Cancer (2010). "IARC Monographs on the Evaluation of Carcinogenic Risks to Humans," in *Alcohol Consumption and Ethyl Carbamate* (Lyon, France. Available at: http://monographs.iarc.fr/ENG/Monographs/vol96/mono96.pdf (Accessed October 24, 2021).

Johnson, N. W., Jayasekara, P., and Amarasinghe, A. A. H. K. (2011). Squamous Cell Carcinoma and Precursor Lesions of the Oral Cavity: Epidemiology and Aetiology. *Periodontol* 57, 19–37. doi:10.1111/j.1600-0757.2011.00401.x

Jones, C. I., Zabolotskaya, M. V., King, A. J., Stewart, H. J. S., Horne, G. A., Chevassut, T. J., et al. (2012). Identification of Circulating microRNAs as Diagnostic Biomarkers for Use in Multiple Myeloma. *Br. J. Cancer* 107, 1987–1996. doi:10.1038/bjc.2012.525

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nuc. Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Kapania, E. M., Reif, T. J., Tsumura, A., Eby, J. M., and Callaci, J. J. (2020). Alcohol-induced Wnt Signaling Inhibition during Bone Fracture Healing Is Normalized by Intermittent Parathyroid Hormone Treatment. *Anim. Models Exp. Med.* 3 (2), 200–207. doi:10.1002/ame2.12116

Kaymak, T., Hruz, P., and Niess, J. H. (2021). Immune System and Microbiome in the Esophagus: Implications for Understanding Inflammatory Diseases. *FEBS J.* doi:10.1111/febs.16103

Kelley, K. W., and Dantzer, R. (2011). Alcoholism and Inflammation: Neuroimmunology of Behavioral and Mood Disorders. *Brain Behav. Immun.* 25 (Suppl. 1), S13–S20. doi:10.1016/j.bbi.2010.12.013

Keshavarzian, A., Farhadi, A., Forsyth, C. B., Rangan, J., Jakate, S., Shaikh, M., et al. (2009). Evidence that Chronic Alcohol Exposure Promotes Intestinal Oxidative Stress, Intestinal Hyperpermeability and Endotoxemia Prior to Development of Alcoholic Steatohepatitis in Rats. *J. Hepatol.* 50 (3), 538–547. doi:10.1016/j.jhep.2008.10.028

Klieser, E., Mayr, C., Kiesslich, T., Wissniowski, T., Fazio, P. D., Neureiter, D., et al. (2019). The Crosstalk of miRNA and Oxidative Stress in the Liver: From Physiology to Pathology and Clinical Implications. *Ijms* 20 (21), 5266. doi:10.3390/ijms20215266

Kruger, J., and Rehmsmeier, M. (2006). RNAhybrid: microRNA Target Prediction Easy, Fast and Flexible. *Nucleic Acids Res.* 34, W451–W454. doi:10.1093/nar/gkl243

Kumar, S., Vijayan, M., Bhatti, J. S., and Reddy, P. H. (2017). MicroRNAs as Peripheral Biomarkers in Aging and Age-Related Diseases. *Prog. Mol. Biol. Transl. Sci.* 146, 47–94. doi:10.1016/bs.pmbts.2016.12.013

Langevin, S., Kuhnell, D., Parry, T., Biesiada, J., Huang, S., Wise-Draper, T., et al. (2017). Comprehensive microRNA-Sequencing of Exosomes Derived from Head and Neck Carcinoma Cells *In Vitro* Reveals Common Secretion Profiles and Potential Utility as Salivary Biomarkers. *Oncotarget* 8 (47), 82459–82474. doi:10.18632/oncotarget.19614

Lauing, K. L., Roper, P. M., Nauer, R. K., and Callaci, J. J. (2012). Acute Alcohol Exposure Impairs Fracture Healing and Deregulates β-Catenin Signaling in the Fracture Callus. *Alcohol. Clin. Exp. Res.* 36, 2095–2103. doi:10.1111/j.1530-0277.2012.01830.x

Leite, R. B., Marinho, A. C. O., Costa, B. L., Laranjeira, M. B. V., Araújo, K. D. T., and Cavalcanti, A. F. M. (2021). The Influence of Tobacco and Alcohol in Oral Cancer: Literature Review. *J. Bras. Patol. Med. Lab.* 57, 1–5. doi:10.5935/1676-2444.20210001

Lewohl, J. M., Nunez, Y. O., Dodd, P. R., Tiwari, G. R., Harris, R. A., and Mayfield, R. D. (2011). Up-regulation of microRNAs in Brain of Human Alcoholics. *Alcohol. Clin. Exp. Res.* 35, 1928–1937. doi:10.1111/j.1530-0277.2011.01544.x

Li, J., Wang, M., Song, L., Wang, X., Lai, W., and Jiang, S. (2020). Lnc RNA MALAT 1 Regulates Inflammatory Cytokine Production in Lipopolysaccharide-stimulated Human Gingival Fibroblasts through Sponging miR-20a and Activating TLR 4 Pathway. *J. Periodont Res.* 55 (2), 182–190. doi:10.1111/jre.12700

Li, X., Xu, M., Ding, L., and Tang, J. (2019). MiR-27a: A Novel Biomarker and Potential Therapeutic Target in Tumors. *J. Cancer* 10 (12), 2836–2848. doi:10.7150/jca.31361

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., et al. (2005). Microarray Analysis Shows that Some microRNAs Downregulate Large Numbers of Target mRNAs. *Nature* 433, 769–773. doi:10.1038/nature03315

Lim, Y., Beane-Ebel, J. E., Tanaka, Y., Ning, B., Husted, C. R., Henderson, D. C., et al. (2021). Exploration of Alcohol Use Disorder-Associated Brain miRNA-mRNA Regulatory Networks. *Transl. Psychiatry* 11 (1), 504. doi:10.1038/s41398-021-01635-w

Litten, R. Z., Bradley, A. M., and Moss, H. B. (2010). Alcohol Biomarkers in Applied Settings: Recent Advances and Future Research Opportunities. *Alcohol. Clin. Exp. Res.* 34, 955–967. doi:10.1111/j.1530-0277.2010.01170.x

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., et al. (2005). MicroRNA Expression Profiles Classify Human Cancers. *Nature* 435, 834–838. doi:10.1038/nature03702

Luebeck, E. G., Curtius, K., Jeon, J., and Hazelton, W. D. (2012). Impact of Tumor Progression on Cancer Incidence Curves. *Cancer Res.* 73, 1086–1096. doi:10.1158/0008-5472.CAN-12-2198

Ma, P. C., Maulik, G., Christensen, J., and Salgia, R. (2003). c-Met: Structure, Functions and Potential for Therapeutic Inhibition. *Cancer Metastasis Rev.* 22, 309–325. doi:10.1023/a:1023768811842

Machida, T., Tomofuji, T., Maruyama, T., Yoneda, T., Ekuni, D., Azuma, T., et al. (2016). miR-1246 and miR-4644 in Salivary Exosome as Potential Biomarkers for Pancreatobiliary Tract Cancer. *Oncol. Rep.* 36 (4), 2375–2381. doi:10.3892/or.2016.5021

Madoz-Gúrpide, J., Zazo, S., Chamizo, C., Casado, V., Caramés, C., Gavín, E., et al. (2015). Activation of MET Pathway Predicts Poor Outcome to Cetuximab in Patients with Recurrent or Metastatic Head and Neck Cancer. *J. Transl. Med.* 13, 282. doi:10.1186/s12967-015-0633-7

Mahnke, A. H., Sideridis, G. D., Salem, N. A., Tseng, A. M., Carter, R. C., Dodge, N. C., et al. (2021). Infant Circulating MicroRNAs as Biomarkers of Effect in Fetal Alcohol Spectrum Disorders. *Sci. Rep.* 11 (1), 1429. doi:10.1038/s41598-020-80734-y

Maran, A., Hefferan, T. E., Zhang, M., and Turner, R. T. (2004). Unanticipated Changes in Steady-State mRNA Levels for Glyceraldehyde-3-Phosphate Dehydrogenase in Rat Tibiae. *Calcified Tissue Int.* 74, 204–207. doi:10.1007/s00223-002-1098-2

Marín, M. P., Esteban-Pretel, G., Ponsoda, X., Romero, A. M., Ballestín, R., López, C., et al. (2010). Endocytosis in Cultured Neurons Is Altered by Chronic Alcohol Exposure. *Toxicol. Sci.* 115, 202–213. doi:10.1093/toxsci/kfq040

Mason, S., Anthony, B., Lai, X., Ringham, H. N., Wang, M., Witzmann, F. A., et al. (2012). Ethanol Exposure Alters Protein Expression in a Mouse Model of Fetal Alcohol Spectrum Disorders. *Int. J. Proteomics* 2012, 1–10. doi:10.1155/2012/867141

McGeary, S. E., Lin, K. S., Shi, C. Y., Pham, T. M., Bisaria, N., Kelley, G. M., et al. (2019). The Biochemical Basis of microRNA Targeting Efficacy. *Science* 366 (6472), eaav1741. doi:10.1126/science.aav1741

Meadows, G. G., and Zhang, H. (2015). Effects of Alcohol on Tumor Growth, Metastasis, Immune Response, and Host Survival. *Alcohol. Res.* 37 (2), 311–322.

Messingham, K. A. N., Faunce, D. E., and Kovacs, E. J. (2002). Review Article. *Alcohol* 28, 137–149. doi:10.1016/s0741-8329(02)00278-1

Michaud, D. S., Vrieling, A., Jiao, L., Mendelsohn, J. B., Steplowski, E., Lynch, S. M., et al. (2010). Alcohol Intake and Pancreatic Cancer: a Pooled Analysis from the Pancreatic Cancer Cohort Consortium (PanScan). *Cancer Causes Control* 21, 1213–1225. doi:10.1007/s10552-010-9548-z

Miranda, R. C. (2014). MicroRNAs and Ethanol Toxicity. *Int. Rev. Neurobiol.* 115, 245–284. doi:10.1016/B978-0-12-801311-3.00007-X

Miranda, R. C. (2012). MicroRNAs and Fetal Brain Development: Implications for Ethanol Teratology during the Second Trimester Period of Neurogenesis. *Front. Gene* 3, 77. doi:10.3389/fgene.2012.00077

Morris, N. L., Ippolito, J. A., Curtis, B. J., Chen, M. M., Friedman, S. L., Hines, I. N., et al. (2015). Alcohol and Inflammatory Responses: Summary of the 2013 Alcohol and Immunology Research Interest Group (AIRIG) Meeting. *Alcohol* 49 (1), 1–6. doi:10.1016/j.alcohol.2014.07.018

Nanau, R., and Neuman, M. (2015). Biomolecules and Biomarkers Used in Diagnosis of Alcohol Drinking and in Monitoring Therapeutic Interventions. *Biomolecules* 5 (3), 1339–1385. doi:10.3390/biom5031339

Natarajan, S., Pachunka, J., and Mott, J. (2015). Role of microRNAs in Alcohol-Induced Multi-Organ Injury. *Biomolecules* 5 (4), 3309–3338. doi:10.3390/biom5043309

National Institutes of Health (2021a). National Institute on Alcohol Abuse and Alcoholism (NIH-NIAAA). *Alcohol Facts Stat.* Updated June 2021. Available at: https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-facts-and-statistics.

National Institutes of Health (2021b). National Institute on Alcohol Abuse and Alcoholism (NIH-NIAAA). Alcohol Use Disorder: A Comparison Between DSM-IV and DSM–5. Updated April 2021. Available at: https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-use-disorder- comparison-between-dsm.

Niespolo, C., Johnston, J. M., Deshmukh, S. R., Satam, S., Shologu, Z., Villacanas, O., et al. (2020). Tribbles-1 Expression and its Function to Control Inflammatory Cytokines, Including Interleukin-8 Levels Are Regulated by miRNAs in Macrophages and Prostate Cancer Cells. *Front. Immunol.* 11 (11), 574046. doi:10.3389/fimmu.2020.574046

O'Connell, R. M., Rao, D. S., and Baltimore, D. (2012). microRNA Regulation of Inflammatory Responses. *Annu. Rev. Immunol.* 30, 295–312. doi:10.1146/annurev-immunol-020711-075013

Ohira, H., Obara, K., Kuroda, M., Takagi, T., Akatuka, H., Ochiai, H., et al. (1995). Effect of Ethanol on DNA Synthesis and C-Met Induction in Chang Liver Cells with Long-Term Culture. *Fukushima J. Med. Sci.* 41, 103–110.

Osterndorff-Kahanek, E. A., Tiwari, G. R., Lopez, M. F., Becker, H. C., Harris, R. A., and Mayfield, R. D. (2018). Long-term Ethanol Exposure: Temporal Pattern of microRNA Expression and Associated mRNA Gene Networks in Mouse Brain. *PLoS ONE* 13 (1), e0190841. doi:10.1371/journal.pone.0190841

Ovcharenko, D., Stölzel, F., Poitz, D., Fierro, F., Schaich, M., Neubauer, A., et al. (2011). miR-10a Overexpression Is Associated with NPM1 Mutations and MDM4 Downregulation in Intermediate-Risk Acute Myeloid Leukemia. *Exp. Hematol.* 39 (10), 1030–1042. doi:10.1016/j.exphem.2011.07.008

Papadopoulos, G. L., Alexiou, P., Maragkakis, M., Reczko, M., and Hatzigeorgiou, A. G. (2009). DIANA-mirPath: Integrating Human and Mouse microRNAs in Pathways. *Bioinformatics* 25, 1991–1993. doi:10.1093/bioinformatics/btp299

Park, N. J., Zhou, H., Elashoff, D., Henson, B. S., Kastratovic, D. A., Abemayor, E., et al. (2009). Salivary microRNA: Discovery, Characterization, and Clinical Utility for Oral Cancer Detection. *Clin. Cancer Res.* 15, 5473–5477. doi:10.1158/1078-0432.CCR-09-0736

Patel, R. S., Jakymiw, A., Yao, B., Pauley, B. A., Carcamo, W. C., Katz, J., et al. (2011). High Resolution of microRNA Signatures in Human Whole Saliva. *Arch. Oral Biol.* 56, 1506–1513. doi:10.1016/j.archoralbio.2011.05.015

Pavlidis, P., Li, Q., and Noble, W. S. (2003). The Effect of Replication on Gene Expression Microarray Experiments. *Bioinformatics* 19, 1620–1627. doi:10.1093/bioinformatics/btg227

Pelucchi, C., Tramacere, I., Boffetta, P., Negri, E., and Vecchia, C. L. (2011). Alcohol Consumption and Cancer Risk. *Nutr. Cancer* 63, 983–990. doi:10.1080/01635581.2011.596642

Peterson, S. M., Thompson, J. A., Ufkin, M. L., Sathyanarayana, P., Liaw, L., and Congdon, C. B. (2014). Common Features of microRNA Target Prediction Tools. *Front. Genet.* 5, 23. doi:10.3389/fgene.2014.00023

Philippe, L., Alsaleh, G., Pichot, A., Ostermann, E., Zuber, G., Frisch, B., et al. (2013). MiR-20a Regulates ASK1 Expression and TLR4-dependent Cytokine Release in Rheumatoid Fibroblast-like Synoviocytes. *Ann. Rheum. Dis.* 72 (6), 1071–1079. doi:10.1136/annrheumdis-2012-201654

Pietrzykowski, A. Z., Friesen, R. M., Martin, G. E., Puig, S. I., Nowak, C. L., Wynne, P. M., et al. (2008). Posttranscriptional Regulation of BK Channel Splice Variant Stability by miR-9 Underlies Neuroadaptation to Alcohol. *Neuron* 59, 274–287. doi:10.1016/j.neuron.2008.05.032

Pritchard, C. C., Cheng, H. H., and Tewari, M. (2012). MicroRNA Profiling: Approaches and Considerations. *Nat. Rev. Genet.* 13, 358–369. doi:10.1038/nrg3198

Puccini, A., Marín-Ramos, N. I., Bergamo, F., Schirripa, M., Lonardi, S., Lenz, H.-J., et al. (2019). Safety and Tolerability of C-MET Inhibitors in Cancer. *Drug Saf.* 42 (2), 211–233. doi:10.1007/s40264-018-0780-x

Qin, L., He, J., Hanes, R. N., Pluzarev, O., Hong, J.-S., and Crews, F. T. (2008). Increased Systemic and Brain Cytokine Production and Neuroinflammation by Endotoxin Following Ethanol Treatment. *J. Neuroinflammation* 5, 10. doi:10.1186/1742-2094-5-10

Rao, R. K. (2008). "Acetaldehyde-induced Barrier Disruption and Paracellular Permeability in Caco-2 Cell Monolayer," in *Alcohol. Methods in Molecular Biology*. Editor L. E. Nagy (New York, NY: Humana Press), 447, 171–183. doi:10.1007/978-1-59745-242-7_13

Rehmsmeier, M., Steffen, P., Höchsmann, M., and Giegerich, R. (2004). Fast and Effective Prediction of microRNA/target Duplexes. *RNA* 10, 1507–1517. doi:10.1261/rna.5248604

Rindom Schiott, C., and Loe, H. (1970). The Origin and Variation in Number of Leukocytes in the Human Saliva. *J. Periodontal Res.* 5, 36–41. doi:10.1111/j.1600-0765.1970.tb01835.x

Romani, C., Salviato, E., Paderno, A., Zanotti, L., Ravaggi, A., Deganello, A., et al. (2021). Genome-wide Study of Salivary miRNAs Identifies miR-423-5p as Promising Diagnostic and Prognostic Biomarker in Oral Squamous Cell Carcinoma. *Theranostics* 11 (6), 2987–2999. doi:10.7150/thno.45157

Rosato, A. J., Chen, X., Tanaka, Y., Farrer, L. A., Kranzler, H. R., Nunez, Y. Z., et al. (2019). Salivary microRNAs Identified by Small RNA Sequencing and Machine Learning as Potential Biomarkers of Alcohol Dependence. *Epigenomics* 11 (7), 739–749. doi:10.2217/epi-2018-0177

Saha, B., Momen-Heravi, F., Kodys, K., and Szabo, G. (2016). MicroRNA Cargo of Extracellular Vesicles from Alcohol-Exposed Monocytes Signals Naive Monocytes to Differentiate into M2 Macrophages. *J. Biol. Chem.* 291 (1), 149–159. doi:10.1074/jbc.M115.694133

Salvi, V., Gianello, V., Tiberio, L., Sozzani, S., and Bosisio, D. (2019). Cytokine Targeting by miRNAs in Autoimmune Diseases. *Front. Immunol.* 10, 15. doi:10.3389/fimmu.2019.00015

Sambuy, Y. (2009). A Sideways Glance. Alcoholic Breakdown of Barriers: How Ethanol Can Initiate a Landslide towards Disease. *Genes Nutr.* 4, 77–81. doi:10.1007/s12263-009-0118-5

Sathyan, P., Golden, H. B., and Miranda, R. C. (2007). Competing Interactions between Micro-RNAs Determine Neural Progenitor Survival and Proliferation after Ethanol Exposure: Evidence from an *Ex Vivo* Model of the Fetal Cerebral Cortical Neuroepithelium. *J. Neurosci.* 27, 8546–8557. doi:10.1523/JNEUROSCI.1269-07.2007

Seitz, H. K., and Stickel, F. (2007). Molecular Mechanisms of Alcohol-Mediated Carcinogenesis. *Nat. Rev. Cancer* 7, 599–612. doi:10.1038/nrc2191

Shafi, G., Aliya, N., and Munshi, A. (2010). MicroRNA Signatures in Neurological Disorders. *Can. J. Neurol. Sci.* 37, 177–185. doi:10.1017/s0317167100009902

Shu, J., Silva, B. V. R. e., Gao, T., Xu, Z., and Cui, J. (2017). Dynamic and Modularized MicroRNA Regulation and its Implication in Human Cancers. *Sci. Rep.* 7 (1), 13356. doi:10.1038/s41598-017-13470-5

Silva, S. S., Lopes, C., Teixeira, A. L., Sousa, M. J. C. d., and Medeiros, R. (2015). Forensic miRNA: Potential Biomarker for Body Fluids? *Forensic Sci. Int. Genet.* 14, 1–10. doi:10.1016/j.fsigen.2014.09.002

Simmons, S., Peng, J., Bienkowska, J., and Berger, B. (2015). Discovering what Dimensionality Reduction Really Tells Us about RNA-Seq Data. *J. Comput. Biol.* 22 (8), 715–728. doi:10.1089/cmb.2015.0085

Song, C., Ma, H., Yao, C., Tao, X., and Gan, H. (2012a). Alveolar Macrophage-Derived Vascular Endothelial Growth Factor Contributes to Allergic Airway Inflammation in a Mouse Asthma Model. *Scand. J. Immunol.* 75, 599–605. doi:10.1111/j.1365-3083.2012.02693.x

Srivastava, V. K., Hiney, J. K., and Dees, W. L. (2011). Prepubertal Ethanol Exposure Alters Hypothalamic Transforming Growth Factor-α and erbB1 Receptor Signaling in the Female Rat. *Alcohol* 45, 173–181. doi:10.1016/j.alcohol.2010.08.014

Stickel, F., and Dubuquoy, L. (2016). MicroRNA in Alcoholic Hepatitis: Implications for Pathophysiology and Treatment. *Gut* 65, 1400–1401. doi:10.1136/gutjnl-2016-312101

Szabo, G., and Bala, S. (2010). Alcoholic Liver Disease and the Gut-Liver axis. *Wjg* 16 (11), 1321–1329. doi:10.3748/wjg.v16.i11.1321

Szabo, G. (1999). Consequences of Alcohol Consumption on Host Defence. *Alcohol Alcohol* 34, 830–841. doi:10.1093/alcalc/34.6.830

Tang, Y., Banan, A., Forsyth, C. B., Fields, J. Z., Lau, C. K., Zhang, L. J., et al. (2008). Effect of Alcohol on miR-212 Expression in Intestinal Epithelial Cells and its Potential Role in Alcoholic Liver Disease. *Alcohol. Clin. Exp. Res.* 32, 355–364. doi:10.1111/j.1530-0277.2007.00584.x

Tao, Z., Zhou, Y., Zeng, B., Yang, X., and Su, M. (2021). MicroRNA-183 Attenuates Osteoarthritic Pain by Inhibiting the TGFα-Mediated CCL2/CCR2 Signalling axis. *Bone Jt. Res.* 10 (8), 548–557. doi:10.1302/2046-3758.108.BJR-2019-0308.R2

Tarallo, A., Carissimo, A., Gatto, F., Nusco, E., Toscano, A., Musumeci, O., et al. (2019). microRNAs as Biomarkers in Pompe Disease. *Genet. Med.* 21, 591–600. doi:10.1038/s41436-018-0103-8

Thulasingam, S., Massilamany, C., Gangaplara, A., Dai, H., Yarbaeva, S., Subramaniam, S., et al. (2011). miR-27b*, an Oxidative Stress-Responsive microRNA Modulates Nuclear Factor-kB Pathway in RAW 264.7 Cells. *Mol. Cel Biochem* 352 (1-2), 181–188. doi:10.1007/s11010-011-0752-2

Ureña-Peralta, J. R., Alfonso-Loeches, S., Cuesta-Diaz, C. M., García-García, F., and Guerri, C. (2018). Deep Sequencing and miRNA Profiles in Alcohol-

Induced Neuroinflammation and the TLR4 Response in Mice Cerebral Cortex. *Sci. Rep.* 8 (1), 15913. doi:10.1038/s41598-018-34277-y

Vagts, A. J., He, D.-Y., Yaka, R., and Ron, D. (2003). Cellular Adaptation to Chronic Ethanol Results in Altered Compartmentalization and Function of the Scaffolding Protein RACK1. *Alcohol. Clin. Exp. Res.* 27, 1599–1605. doi:10.1097/01.ALC.0000089957.63597.A4

Valmiki, S., Ahuja, V., Puri, N., and Paul, J. (2020). miR-125b and miR-223 Contribute to Inflammation by Targeting the Key Molecules of NFκB Pathway. *Front. Med.* 6, 313. doi:10.3389/fmed.2019.00313

van der Werf, M. J., Pieterse, B., van Luijk, N., Schuren, F., van der Werff-van der Vat, B., Overkamp, K., et al. (2006). Multivariate Analysis of Microarray Data by Principal Component Discriminant Analysis: Prioritizing Relevant Transcripts Linked to the Degradation of Different Carbohydrates in Pseudomonas Putida S12. *Microbio* 152, 257–272. doi:10.1099/mic.0.28278-0

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., et al. (2002). Accurate Normalization of Real-Time Quantitative RT-PCR Data by Geometric Averaging of Multiple Internal Control Genes. *Genome Biol.* 3, 0034.1. doi:10.1186/gb-2002-3-7-research0034

Vangipuram, S. D., and Lyman, W. D. (2012). Ethanol Affects Differentiation-Related Pathways and Suppresses Wnt Signaling Protein Expression in Human Neural Stem Cells. *Alcohol. Clin. Exp. Res.* 36, 788–797. doi:10.1111/j.1530-0277.2011.01682.x

Vasioukhin, V. (2012). Adherens Junctions and Cancer. *Sub-cellular Biochem.* 60, 379–414. doi:10.1007/978-94-007-4186-7_16

Velazquez-Marrero, C., Burgos, A., Garcia, J. O., Palacio, S., Marrero, H. G., Bernardo, A., et al. (2016). Alcohol Regulates BK Surface Expression via Wnt/-Catenin Signaling. *J. Neurosci.* 36 (41), 10625–10639. doi:10.1523/JNEUROSCI.0491-16.2016

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer-Verlag.

Vlachos, I. S., Kostoulas, N., Vergoulis, T., Georgakilas, G., Reczko, M., Maragkakis, M., et al. (2012). DIANA miRPath v.2.0: Investigating the Combinatorial Effect of microRNAs in Pathways. *Nuc. Acids Res.* 40, W498–W504. doi:10.1093/nar/gks494

Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., et al. (2015). DIANA-miRPath v3.0: Deciphering microRNA Function with Experimental Support. *Nucleic Acids Res.* 43 (W1), W460–W466. doi:10.1093/nar/gkv403

Volkow, N. D., and Morales, M. (2015). The Brain on Drugs: from Reward to Addiction. *Cell* 162 (4), 712–725. doi:10.1016/j.cell.2015.07.046

Vu, T. T., Stölzel, F., Wang, K. W., Röllig, C., Tursky, M. L., Molloy, T. J., et al. (2021). miR-10a as a Therapeutic Target and Predictive Biomarker for MDM2 Inhibition in Acute Myeloid Leukemia. *Leukemia* 35 (7), 1933–1948. doi:10.1038/s41375-020-01095-z

Wang, H., Taguchi, Y. H., and Liu, X. (2021). Editorial: miRNAs and Neurological Diseases. *Front. Neurol.* 12, 662373. doi:10.3389/fneur.2021.662373

Wang, J., Chen, J., and Sen, S. (2016). MicroRNA as Biomarkers and Diagnostics. *J. Cel. Physiol.* 231 (1), 25–30. doi:10.1002/jcp.25056

Wang, Y.-l., Malik, A. B., Sun, Y., Hu, S., Reynolds, A. B., Minshall, R. D., et al. (2011). Innate Immune Function of the Adherens junction Protein P120-Catenin in Endothelial Response to Endotoxin. *J.I.* 186, 3180–3187. doi:10.4049/jimmunol.1001252

Wong, A. W., Paulson, Q. X., Hong, J., Stubbins, R. E., Poh, K., Schrader, E., et al. (2011). Alcohol Promotes Breast Cancer Cell Invasion by Regulating the Nm23-ITGA5 Pathway. *J. Exp. Clin. Cancer Res.* 30, 75. doi:10.1186/1756-9966-30-75

Wong, D. T. W. (2015). Salivary Extracellular Noncoding RNA: Emerging Biomarkers for Molecular Diagnostics. *Clin. Ther.* 37 (3), 540–551. doi:10.1016/j.clinthera.2015.02.017

Wood, S., Pithadia, R., Pithadia, R., Rehman, T., Zhang, L., Plichta, J., et al. (2013). Chronic Alcohol Exposure Renders Epithelial Cells Vulnerable to Bacterial Infection. *PLoS One* 8, e54646. doi:10.1371/journal.pone.0054646

Xu, M., Chen, G., Fu, W., Liao, M., Frank, J. A., Bower, K. A., et al. (2012). Ethanol Disrupts Vascular Endothelial Barrier: Implication in Cancer Metastasis. *Toxicol. Sci.* 127, 42–53. doi:10.1093/toxsci/kfs087

Xu, M., Ren, Z., Wang, X., Comer, A., Frank, J. A., Ke, Z.-j., et al. (2016). ErbB2 and P38*γ* MAPK Mediate Alcohol-Induced Increase in Breast Cancer Stem Cells and Metastasis. *Mol. Cancer* 15, 52. doi:10.1186/s12943-016-0532-4

Yang, Y., Chen, H., Ding, N., Wang, S., Duan, Z., Birnbaum, Y., et al. (2018). Expression Profiling of Circular RNAs and Micrornas in Heart Tissue of Mice with Alcoholic Cardiomyopathy. *Cell Physiol. Biochem.* 46 (6), 2284–2296. doi:10.1159/000489596

Yeh, C.-H., Chang, J.-K., Wang, Y.-H., Ho, M.-L., and Wang, G.-J. (2008). Ethanol May Suppress Wnt/β-Catenin Signaling on Human Bone Marrow Stroma Cells. *Clin. Orthop. Relat. Res.* 466, 1047–1053. doi:10.1007/s11999-008-0171-1

Yeligar, S., Tsukamoto, H., and Kalra, V. K. (2009). Ethanol-Induced Expression of ET-1 and ET-BR in Liver Sinusoidal Endothelial Cells and Human Endothelial Cells Involves Hypoxia-Inducible Factor-1α and MicroRNA-199. *J. Immunol.* 183, 5232–5243. doi:10.4049/jimmunol.0901084

Zakhari, S., Vasiliou, V., and Guo, Q. M. (2011). *Alcohol and Cancer*. New York: Springer.

Zhang, J., Cao, Z., Yang, G., You, L., Zhang, T., and Zhao, Y. (2019). MicroRNA-27a (miR-27a) in Solid Tumors: A Review Based on Mechanisms and Clinical Observations. *Front. Oncol.* 9, 893. doi:10.3389/fonc.2019.00893

Zisman, A. L., Nickolov, A., Brand, R. E., Gorchow, A., and Roy, H. K. (2006). Associations between the Age at Diagnosis and Location of Colorectal Cancer and the Use of Alcohol and Tobacco. *Arch. Intern. Med.* 166, 629–634. doi:10.1001/archinte.166.6.629

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Mead, Boulghassoul-Pietrzykowska, Wang, Anees, Kinstlinger, Lee, Hamza, Feng and Pietrzykowski. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 20 January 2022
doi: 10.3389/fgene.2021.794348

An Update Evolving View of Copy Number Variations in Autoimmune Diseases

*Rong-hua Song, Chao-qun Gao, Jing Zhao and Jin-an Zhang**

Department of Endocrinology and Rheumatology, Shanghai University of Medicine and Health Sciences Affiliated Zhoupu Hospital, Shanghai, China

Autoimmune diseases (AIDs) usually share possible common mechanisms, i.e., a defect in the immune tolerance exists due to diverse causes from central and peripheral tolerance mechanisms. Some genetic variations including copy number variations (CNVs) are known to link to several AIDs and are of importance in the susceptibility to AIDs and the potential therapeutic responses to medicines. As an important source of genetic variants, DNA CNVs have been shown to be very common in AIDs, implying these AIDs may possess possible common mechanisms. In addition, some CNVs are differently distributed in various diseases in different ethnic populations, suggesting that AIDs may have their own different phenotypes and different genetic and/or environmental backgrounds among diverse populations. Due to the continuous advancement in genotyping technology, such as high-throughput whole-genome sequencing method, more susceptible variants have been found. Moreover, further replication studies should be conducted to confirm the results of studies with different ethnic cohorts and independent populations. In this review, we aim to summarize the most relevant data that emerged in the past few decades on the relationship of CNVs and AIDs and gain some new insights into the issue.

Keywords: copy number variations, single-nucleotide polymorphism, autoimmune disease, autoimmune thyroid disease, systemic lupus erythematosus

OPEN ACCESS

Edited by:
William C. Cho,
QEH, Hong Kong SAR, China

Reviewed by:
Yang Zhou,
Huazhong Agricultural University, China
George Wiggins,
University of Otago, New Zealand

****Correspondence:***
Jin-an Zhang
zhangjinan@hotmail.com

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: *13 October 2021*
Accepted: *06 December 2021*
Published: *20 January 2022*

Citation:
Song R-h, Gao C-q, Zhao J and Zhang J-a (2022) An Update Evolving View of Copy Number Variations in Autoimmune Diseases. Front. Genet. 12:794348. doi: 10.3389/fgene.2021.794348

INTRODUCTION

Copy number variations (CNVs), as a main type of structure variation (SV) caused by genomic rearrangement, mainly include deletion and duplication of sub-microscopic but large genomic segments ranging from 1 kb to 3 Mb (Redon et al., 2006). Single-nucleotide polymorphisms (SNPs) have been recognized to be involved in many autoimmune diseases (AIDs) (Song et al., 2021a; Song et al., 2021b; Jiang et al., 2021); however, CNVs containing more nucleotide content per genome than SNPs are responsible for a large proportion of human genetic variation and show an importance in genetic diversity and evolution (Redon et al., 2006; Cleynen et al., 2016). Thus, more attention has been paid to the research of CNVs in diseases. Nowadays, the genome-wide assays for CNV study include array-based comparative genomic hybridization (aCGH), SNP genotyping microarrays, next-generation sequencing, and long-read sequencing techniques (Hehir-Kwa et al., 2018). There are several main categories in the molecular mechanisms during the process of CNV formation, DNA recombination, rearrangement, and error replication. Besides, CNVs also have several types, such as insertions, deletions, inversions, and translocations (Feuk et al., 2006; Human Genome Structural Variation Working Group et al., 2007). Numerous reports imply that CNV is one of the main genetic factors underlying human diseases, including AIDs (Hauptmann et al., 1974; Tomer

TABLE 1 | Copy number variant loci or genes related to autoimmune diseases.

CNV-related genes or regions	Autoimmune diseases/ syndromes	Populations (sample size)	CNV detection methods	Most common copies in healthy normal controls	Risk-associated CNVs (*p*-values)	Results	References
Beta-defensin gene	Psoriasis	Dutch and German population (179 Dutch patients and 272 controls; 319 German patients and 305 controls)	High-throughput paralogue ratio test (PRT)	40% controls were four copies	Higher gnomic copy number (p = 0.01 for Dutch, p = 7.8E–5 for German)	Consistent	Hollox et al. (2008)
BMP8A	AS	Iranian (450 patients and 450 healthy controls)	TaqMan real-time polymerase chain reaction (PCR)	Two copies	No significant association	–	Shahba et al. (2018)
C3 and *C5*	BD	Han Chinese population (1,064 patients and 2,174 controls)	Real-time PCR	Two copies of C3 and C5	More than two copies of C3 (p = 5.5E–3) and C5 (p = 1.1E–8)	–	Xu et al. (2015)
C4 and its two isotypes, *C4A* and *C4B*	SLE	Caucasians (Columbus): 1,241 European Americans (Yang et al. 2007) and Brazilian (427 patients and 301 controls)	TaqI southern blots (Yang et al. 2007); real-time PCR (Pereira et al. 2019)	Four copies of C4 genes (majority)	Lower copy number of C4 and C4A (p = 0.00002; Yang et al. 2007); low total copy number of C4 (p < 0.001), C4A (p < 0.001), and C4B (p = 0.03)	Consistent	Hauptmann et al. (1974), Yang et al. (2007), and Pereira et al. (2019)
		221 Caucasians (North American)	TaqMan-based real-time PCR and Southern blotting	Not reported	Higher copy number of C4B associated with hypertension and effective response to statin therapy in childhood-onset SLE patients (p = 0.016) and higher diastolic blood pressure (p = 0.015)	–	Mulvihill et al. (2019)
	CD	Belgian population (770 patients and 345 controls)	Array comparative genomic hybridization	Not reported	Lower C4L (p = 7.68E–3) and higher C4S copies (p = 6.29E–3)	–	Cleynen et al. (2016)
	BD	Han Chinese population (905 patients and 1,238 controls)	Real-time PCR	Four copies of C4 genes, 2 copies of C4A genes	More than two copies of C4A (p = 1.65E–7)	–	Hou et al. (2013)
	T1DM	American population (cohort 1: 50 patients and 57 controls; cohort 2: 110 patients)	TaqMan quantitative PCR	Four copies of C4 genes; two copies of C4A gene	Fewer copies of C4A (p = 1.76E–5) and HERV-K (C4) (p = 4.59E–7)	–	Mason et al. (2014)
	GD	Chinese population (624 patients and 160 healthy individuals)	Quantitative real-time polymerase chain reaction	Four copies of C4 genes; two copies of C4A and C4B genes	Four copies of C4 (p = 0.001); copies of C4A (p = 0.008) and C4B (p = 2.42E-5) in GD	–	Liu et al. (2011)
CCL3L1	SLE	Caucasians (San Antonio), 1,084 subjects (469 cases of SLE and 615 matched controls)	59-RACE and reverse transcription-PCR (RT-PCR)	Two copies	Lower than or greater than two copies (p = 0.032)	–	Mamtani et al. (2008)
	RA	Caucasian (1,136 patients and 1,470 controls; McKinneyet al.2008), Tunisians and French (100 French patients and 200 controls; 166 Tunisian patients and 102 controls; Ben Kilani et al. 2016), United Kingdom population	Reverse transcriptase (RT)-PCR (McKinney et al. 2008), droplet digital PCR (ddPCR) (Ben Kilani et al. 2016), PRT methodology (Carpenter et al. 2011)	Two copies	Higher than two in the New Zealand cohort (p = 0.009) but not the United Kingdom cohort; no association in the French study, a protective effect of five copies in the Tunisian population (p value not	Controversial	McKinney et al. (2008), Ben Kilani et al. (2016), and Carpenter et al. (2011)

(Continued on following page)

TABLE 1 | (*Continued*) Copy number variant loci or genes related to autoimmune diseases.

CNV-related genes or regions	Autoimmune diseases/ syndromes	Populations (sample size)	CNV detection methods	Most common copies in healthy normal controls	Risk-associated CNVs (*p*-values)	Results	References
		(657 patients and 192 controls), and United Kingdom (274 patients and 276 controls)			reported (Ben Kilani et al. 2016), no association (Carpenter et al. 2011)		
	CD and psoriasis	United Kingdom (657 CD patients, 202 psoriasis patients, and 276 controls)	PRT methodology	Not reported	No association	–	Carpenter et al. (2011)
	AS	Algerian (81 patients and 119 controls)	Digital droplet PCR (ddPCR)	Two copies	No association	–	Dahmani et al. (2019)
	T1DM	Caucasian population (252 patients and 1,470 controls; McKinney et al. 2008) and 2,000 patients and 3,000 controls (Wellcome Trust Case Control Consortium et al. 2010)	Reverse transcriptase (RT)-PCR (McKinney et al. 2008) and the Agilent Comparative Genomic Hybridization (CGH) platform (Wellcome Trust Case Control Consortium et al. 2010)	Two copies	A copy number higher than 2 ($p = 0.064$) in the Caucasian population (McKinney et al. 2008); no association in the study (Wellcome Trust Case Control Consortium et al. 2010)	Consistent	McKinney et al. (2008) and Wellcome Trust Case Control Consortium et al. (2010)
CD40, *PTPN22*, *and CTLA-4*	GD	Caucasian (191 patients and 192 controls)	Quantitative-PCR (Q-PCR) assays	Two copies	No copy variation in the CD40, CTLA-4, and PTPN22 gene number variation in GD	–	Huber et al. (2011)
CFH, *CFHR1*, *KIAA0125*, *UGT2B15*, *UGT2B17*, *TRY6*, and *CCL3L1*	GD	Chinese Han population (144 patients and 144 controls)	TaqMan quantitative polymerase chain reaction (TaqMan qPCR)	Two copies	No association	–	Song et al. (2017)
DEFA1	BD	Korean population (55 patients and 35 controls)	A duplex TaqMan® real-time PCR assay	Most samples (31.1%) had a CN of five	A high CN associated with intestinal involvement in BD patients ($p = 0.005$)	–	Ahn et al. (2012)
DEFB4	BD	Korean population (197 patients and 197 controls)	A novel comparative multiplex polymerase chain reaction (PCR): paralogue ratio test (PRT)	Median copy number was four	Lower copy number, but no statistical difference ($p = 0.245$)	–	Park et al. (2011)
DEFA1A3	CD	Danes (240 patients)	Combined real-time quantitative PCR and pyrosequencing	Mean copy number was 6.7	Higher copy number ($p < 0.001$)	–	Jespersgaard et al. (2011)
DEFB103	AS	Chinese population (406 patients and 401 controls)	A multiplex fluorescence competitive polymerase chain reaction (PCR)	Ranged from 2 to 6	No association	–	Cai et al. (2015)
FAS, caspase8, caspase3, and *BCL2*	BD	Han Chinese population (1,014 patients and 2,076 controls)	TaqMan copy number assays and real-time PCR	Diploid 2 copy number carriers	High FAS copy number (>2) ($p = 1.05E–3$ in the first-stage study, $p = 3.35E–8$ in the replication and combined study)	–	Yu et al. (2015)
FCGR3A	SLE	Taiwanese population (846 patients with SLE and 1,420 healthy control subjects)	Custom TaqMan CNV real-time quantitative polymerase chain reaction (PCR) assays	Two copies	A low FCGR3A copy number (<2) ($p = 5.06E–4$) and a high (>2) FCGR3A copy number ($p = 0.003$)	–	Chen et al. (2014)

(Continued on following page)

TABLE 1 | (*Continued*) Copy number variant loci or genes related to autoimmune diseases.

CNV-related genes or regions	Autoimmune diseases/ syndromes	Populations (sample size)	CNV detection methods	Most common copies in healthy normal controls	Risk-associated CNVs (*p*-values)	Results	References
	RA	Taiwanese population (948 patients with RA and 1,420 healthy control subjects)	Custom TaqMan CNV real-time quantitative polymerase chain reaction (PCR) assays	Two copies	A low copy number (p = 5.83E–4)	–	Chen et al. (2014)
	AS	Algerian (81 patients and 119 controls; Dahmani et al.2019), Chinese population (402 patients and 399 controls; Wang et al.2016)	Digital droplet PCR (ddPCR) (Dahmani et al. 2019) and AccuCopy™ method (Wang et al. 2016)	Two copies	Less than two copies (<2) (p = 0.0001; Dahmani et al. 2019), a low copy number (p < 0.001; Wang et al. 2016)	Consistent	Dahmani et al. (2019) and Wang et al. (2016)
FCGR3B	SLE	United Kingdom Caucasians (171 patients and 176 controls; Willcocks et al. 2008), Spanish ancestry (146 patients and 409 controls; Mamtani et al. 2010), Afro-Caribbean (134 patients and 589 controls; Molokhia et al. 2011), Taiwanese (846 patients and 1,420 controls; Chen et al. 2014), and Brazilian population (135 unrelated SLE patients and 200 healthy unrelated subjects; Barbosa et al. 2018)	qPCR (Willcocks et al. 2008), real-time PCR (Mamtani et al. 2010), paralogue ratio test (PRT) assay (Molokhia et al. 2011), custom TaqMan CNV real-time quantitative polymerase chain reaction (PCR) assays (Chen et al. 2014), and quantitative real-time PCR (Barbosa et al. 2018)	Two copies	A lower (<2) copy number in United Kingdom Caucasians (p = 0.027; Willcocks et al. 2008), copy number <2 or >2 in cases of Spanish ancestry (p = 0.001 and 0.013, respectively; Mamtani et al. 2010), Afro-Caribbean (p = 0.02; Molokhia et al. 2011), Taiwanese (p = 0.0032; Chen et al. 2014), and Brazilian population (p = 1.66E–3; Barbosa et al. 2018)	Consistent	Willcocks et al. (2008), Mamtani et al. (2010), Molokhia et al. (2011), Chen et al. (2014), and Barbosa et al. (2018)
	RA	Spanish ancestry (158 patients and 409 controls; Mamtani et al. 2010), South Australia (197 patients and 162 controls; Graf et al. 2012), Taiwanese (948 patients and 1,420 controls; Chen et al. 2014), British population (480 patients; Rahbari et al. 2017), French population (Bai Kilani et al. 2019)	Real-time PCR (Mamtani et al. 2010), custom TaqMan® CN assay (Graf et al. 2012), custom TaqMan CNV real-time quantitative polymerase chain reaction (PCR) assays (Chen et al. 2014), a PRT/REDVR approach (Rahbari et al. 2017), and droplet digital PCR (Bai Kilani et al. 2019)	Two copies	No association in cases of Spanish ancestry (Mamtani et al. 2010), lower and higher copy number in South Australia (p = 0.017; Graf et al. 2012), no association in Taiwanese (Chen et al. 2014), deletion in British population (p = 2.9E–3; Rahbari et al. 2017), and without null allele (one–three copy numbers) in French population (Bai Kilani et al. 2019)	Controversial	Mamtani et al. (2010), Graf et al. (2012), Chen et al. (2014), Rahbari et al. (2017), and Bai Kilani et al. (2019)
	UC	Japanese population (752 patients and 2,062 controls)	TaqMan assay	Not reported	Abnormal copies (p = 0.02)	–	Asano et al. (2013)
	Psoriasis	Han Chinese population (343 patients and 574 controls)	TaqMan® copy number assays	Not reported	A higher copy number (p < 0.02)	–	Wu et al. (2014)
	BD	Iran (187 patients and 178 controls)	Quantitative real-time PCR	Two copies	No association		Black et al. (2012)
	AS	Algerian (81 patients and 119 controls; Dahmani et al.	Digital droplet PCR (ddPCR) (Dahmani	Two copies	No association in Algerian population	Controversial	

(Continued on following page)

TABLE 1 | (*Continued*) Copy number variant loci or genes related to autoimmune diseases.

CNV-related genes or regions	Autoimmune diseases/ syndromes	Populations (sample size)	CNV detection methods	Most common copies in healthy normal controls	Risk-associated CNVs (*p*-values)	Results	References
		2019) and Chinese population (402 patients and 399 controls; Wang et al. 2016)	et al. 2019) and AccuCopy™ method (Wang et al. 2016)		(Dahmani et al. 2019) and a low (≤3) FCGR3B copy number (p = 0.001; Wang et al. 2016)		Dahmani et al. (2019) and Wang et al. (2016)
	pSS	Spanish ancestry (61 patients and 409 controls), Australian (174 patients and 162 controls; Nossent et al. 2012), and Swedish and Norwegian population (124 patients and 139 controls; Haldorsen et al. 2013)	Real-time PCR (Mamtani et al. 2010), a quantitative real-time polymerase chain reaction assay (Nossent et al. 2012), TaqMan copy number assay (Haldorsen et al. 2013)	Two copies	Copy number <2 or >2 in cases of Spanish ancestry (p = 0.074 and 0.048, respectively; Mamtani et al. 2010), low FCGR3B CN (<2 copies) in Australian population (p = 0.016; Nossent et al. 2012), and no association in Swedish and Norwegian population Haldorsen et al. (2013)	Controversial	Mamtani et al. (2010), Nossent et al. (2012), and Haldorsen et al. (2013)
GPC5, *B9D2,* and *ASB11*	AITD	Chinese Han population (158 patients and 181 controls)	Chromosome microarray on the Affymetrix CytoScan™ HD platform, then identified by RTPCR	Not reported	The frequency of CNV loss for GPC5, B9D2, and ASB11 genes was higher in AITD (p < 0.05)	–	Guan et al. (2020)
GSTM1	RA	Swedish (2,426 cases and 1,257 controls) and Tunisian population (165 cases and 102 controls)	TaqMan copy number assays (Lundstrom et al. 2011) and digital droplet PCR (ddPCR) (Achour et al. 2018)	51.8% for 0 copies and 40.1% for 1 copy	No association in Swedish population (Lundstrom et al. 2011) and lack of association (Achour et al. 2018)	Consistent	Lundstrom et al. (2011) and Achour et al. (2018)
HBD-2	CD	German	Genome-wide DNA copy number profiling by array-based comparative genomic hybridization and quantitative polymerase-chain reaction analysis	Median of 4 (range 2–10) copies	The copy number distribution shifted to lower numbers (p = 0.002)	–	Fellermann et al. (2006)
HSP90 and its two major isoforms	SLE	Han Chinese population (419 patients and 538 controls)	A custom-by-design Multiplex AccuCopy™ method	Two copies	Abnormal copies of HSP90AB1 (p = 0.02)	–	Zhang et al. (2019)
KIR3DL1 and *KIR3DS1*	T1DM	White European ancestry (6,744 cases and 5,362 controls)	A hybrid qPCR/SNP array	Two copies of KIR3DL1 and 0 copy of KIR3DS1	No evidence of association	–	Pontikos et al. (2014)
LCE3B and *LCE3C*	Psoriasis	The Dutch population (1,039 cases and 759 controls)	Array comparative genomic hybridization and a polymerase chain reaction: TaqMan SNP genotyping assay	Not reported	Absence (p < 0.0001)		Bergboer et al. (2012)
IL17F, *IL23A*, *IL17A*, and *IL23R*	BD	Han Chinese population (1,036 patients and 2,050 controls)	TaqMan real-time polymerase chain reaction assay	Not reported	More than two copies of IL17F (p = 4.17E–8) and IL23A (p = 2.86E–11) associated with BD and no association	–	Hou et al. (2015)

(Continued on following page)

TABLE 1 | (*Continued*) Copy number variant loci or genes related to autoimmune diseases.

CNV-related genes or regions	Autoimmune diseases/ syndromes	Populations (sample size)	CNV detection methods	Most common copies in healthy normal controls	Risk-associated CNVs (*p*-values)	Results	References
					between CNV of IL17A and IL23R		
IL-22 gene exon1	Psoriasis	Estonian (290 patients and 263 controls)	Quantitative RT-PCR	Not reported	Abnormal copies associated with psoriasis severity ($p < 0.0001$)	–	Prans et al. (2013)
miR-143, *miR-146a*, *miR-9-3*, *miR-205*, *miR-301a*, and *miR-23a*	AS	Chinese Han population (768 patients and 660 controls)	TaqMan PCR	Two copies	Low copy numbers of miR-143 ($p = 1.126E–7$), miR-146a ($p = 3.716E–8$), miR-9-3 ($p = 2.566E–5$), and miR-205 ($p = 7.187E–6$) and high copy numbers of miR-301a ($p = 3.725E–5$) and miR-23a ($p = 8.033E–9$) AAU+, AS+; additionally, a low copy number of miR-146a ($p = 0.001$) and a high copy number of miR-23a and miR-205 ($p = 0.002$) in AAU+, AS–	–	Yang et al. (2017)
miR-23a, miR-146a, and *miR-301a*	BD	Han Chinese population (377 patients and 2,291 controls)	TaqMan PCR	Two copies	No association	–	Hou et al. (2016)
NCF1	RA	The middle and southern parts of Sweden (494 patients and 480 controls)	A multiplex qPCR assay	Two copies	RA less likely to have an increased copy number ($p = 0.037$)	–	Olsson et al. (2012)
SIRPB1 and *TMEM91*	AITD	Chinese Han population (15 patients and 15 controls)	Chromosome microarray	Not reported	The frequency of CNV gain for SIRPB1 was higher in AITD ($p = 0.001$) and no association of TMEM91 CNV and AITD	–	Jin et al. (2018)
T-Bet, *GATA-3*, *RORC*, and *FOXP3*	AS	Chinese Han population (676 patients with AAU, including 298 patients with AAU+, AS+; 378 patients with AAU+, AS–; and 596 unrelated healthy controls)	Real-time PCR	Two copies	A high copy number (CN) of T-bet in AAU+, AS– and AAU+, AS+ ($p = 4.3E–5$ and 1.2E–8, respectively), a high CN of GATA-3 in AAU+, AS+ ($p = 1.8E–7$), a higher frequency of CN of FOXP3 in female AAU+, AS+ and female AAU+, AS– ($p = 0.005$ and 0.004, respectively), and no association between RORC CNVs and AS	–	Bai et al. (2016)
TBX21, *GATA3*, *Rorc*, and *Foxp3*	BD	Han Chinese population (1,048 patients and 2,236 controls)	TaqMan real-time PCR	Two copies of these four genes	High Rorc CNV in BD ($p = 8.99E–8$) and low Foxp3 CNV	–	Liao et al. (2015)

(Continued on following page)

TABLE 1 | (*Continued*) Copy number variant loci or genes related to autoimmune diseases.

CNV-related genes or regions	Autoimmune diseases/ syndromes	Populations (sample size)	CNV detection methods	Most common copies in healthy normal controls	Risk-associated CNVs (*p*-values)	Results	References
					(p = 1.92E–5) in female BD		
TLR7	BD	Chinese Han population (400 patients and 600 controls)	Real-time PCR	One copy for male and two copies for female	A high copy number of TLR7 (p = 0.021 for males and p = 0.048 for females)	–	Fang et al. (2015)
	AS	Chinese Han population (649 patients and 628 controls)	AccuCopy™ method	Not reported	Lower copy number (=1), especially in males (p = 0.009 for TLR7_1 fragment and p = 0.01 for TLR7_2 fragment)	–	Wang et al. (2018)
	GD and GO	Chinese population (196 controls and 484 GD patients, including 203 patients with GO)	Real-time polymerase chain reaction (PCR)	Not reported	A protective effect of lower than normal CNV for TLR7 (CNV <2 for females and CNV <1 for males) but no statistical significance and no association in GO	–	Liao et al. (2014)
TSHR	GD and GO	Chinese population (196 controls and 484 GD patients, including 203 patients with GO)	Real-time polymerase chain reaction (PCR)	Two copies	Copy number <2 or >2 in GD, not in GO (p = 0.01)	–	Liao et al. (2014)
UGT2B17	AS	Newfoundland (298 patients and 299 controls)	Built-in DNA analytics aberration detection method-2 (ADM-2) algorithm	Two copies	The frequency of two copies higher in cases (p < 0.05)	–	Uddin et al. (2013)

CNVs, copy number variations; SLE, systemic lupus erythematosus; RA, rheumatoid arthritis; IBD, inflammatory bowel disease; BD, Behcet's disease, AS, ankylosing spondylitis (AS); pSS, primary Sjogren's syndrome; T1DM, type 1 diabetes mellitus; AITD, autoimmune thyroid disease

and Davies, 2003; Iafrate et al., 2004; Mack et al., 2004; Redon et al., 2006; Yang et al., 2007; Yim et al., 2010). The mechanisms underlying the involvement of CNVs in clinical phenotypes are mainly gene disruption and rearrangement (McCarroll and Altshuler, 2007; Yim et al., 2010). Further deep studies on CNVs have shed new light on human genome structure, genetic variations between individuals, and genetic pathogenic factors of human AIDs.

AIDs usually share possible common mechanisms, i.e., a defect in the immune tolerance exists due to diverse causes from central and peripheral tolerance mechanisms. Although majority of human genetic variations do not contribute to overt diseases (Tomer and Davies, 2003), some genetic variations including SNPs, nucleotide insertions/deletions, structural variations, and CNVs are known to link to several AIDs during the past few decades and are of importance in the susceptibility to AIDs and the potential therapeutic responses to medicines (Iafrate et al., 2004; Redon et al., 2006). Studies have revealed that some SNPs are related to AIDs and could be genetic mechanisms underlying the development of AIDs (Song et al., 2021a; Song et al., 2021b; Jiang et al., 2021). Structural variations including complex rearrangement of segments with sizes of thousands to millions of base pairs have been recognized as a rich source of genetic diversity. CNVs as a crucial source of genomic diversity caused by the rearrangement of genome are ubiquitously presented in human genome and may affect the susceptibility to many diseases (Iafrate et al., 2004). During the past years, thousands of gene CNVs have been reported (Redon et al., 2006). However, most studies focus on their relationship with tumors and chronic diseases (Jia et al., 2011; Saadati et al., 2016; Voll et al., 2017). Moreover, CNVs are very common in genomic regions encoding immune-related genes, which are closely related to the etiology of AIDs. Thus, they potentially impact polygenic autoimmunity and may lead to the imbalance of the autoimmune system and the development of some AIDs. Additionally, some common CNVs have also been reported to be correlated to several specific AIDs (Mamtani et al., 2008; Mamtani et al., 2010; Liu et al., 2011). Yim et al. (2015) reviewed studies on the clinical implications of copy number variations in autoimmune disorders in detail. However, the relationship between CNVs and the pathogenesis of AIDs has not been fully revealed. Furthermore, in recent years, as new

technologies develop, researches on CNVs have made new discoveries and progress. Herein, based on the emergence of numerous studies on the relationship between CNVs and AIDs in the past decades, we reviewed all the related studies and summarized their findings in order to provide new ideas for future explorations and to uncover the mechanisms underlying AIDs (**Table 1**).

SYSTEMIC LUPUS ERYTHEMATOSUS

Systemic lupus erythematosus (SLE), with a prevalence rate of eight–nine times higher in females than in males during childbearing age, is a typical systemic autoimmune inflammatory disease with a strong genetic susceptibility and is characterized by the production of autoantibodies and the existence of chronic inflammation. There is a wide range of clinical features in SLE patients, such as discoid lesions, nephritis, arthritis, and malar rash (Barbosa et al., 2018). So far, the exact genetic physiology of SLE remains an open question. Complement is well known to be involved in many immune-mediated diseases. Among them, complement component C4 is a pivotal effector of the immune system. There are two common isoforms of C4: C4A and C4B, and 95% of C4A and 54% of C4B contain an endogenous retroviral sequence in their ninth intron, HERV-K (C4) (Wouters et al., 2009). HERV-K (C4) can cause the antisense transcription of C4 as it is oriented opposite to C4 (Mack et al., 2004). According to the presence or absence of HERV-K (C4), there are two different size varieties of isotypes for both C4A and C4B: C4L (long) and C4S (short) (Mack et al., 2004; Wu et al., 2008). Hauptmann et al. (1974) first found the deficiency of *C4* copy number in SLE patients. Later, Yang et al. (2007) uncovered that SLE patients had lower copy number of total *C4* and *C4A* genes than healthy volunteers. Pereira et al. (2019) observed that the risk of developing SLE was 2.62 times higher in subjects with low total *C4* copy number and 3.59 times higher in subjects with low *C4A* copy number. These consistent results imply that deletion or deficiency of *C4* or *C4* isotypes will increase the risk for SLE. The potential mechanism is like this, the decrease of *C4* CNVs will cause C4 deficiency, bring the impairment of autoantigene clearance and the negative selection of auto-reactive B cells, and then favor the onset of SLE (Pereira et al., 2019). In addition, there was another study aiming at the relationship between *C4* CNVs and the drug response to treatment of SLE. The study of Mulvihill et al. (2019) found that higher copy number of complement *C4B* and elevated serum complement levels were associated with hypertension and effective response to statin therapy in childhood-onset SLE patients.

C-C chemokine ligand 3 like-1 (CCL3L1) is a potent ligand for the HIV coreceptor, and C-C chemokine receptor 5 (CCR5) is an important factor in immune response (Gonzalez et al., 2005). Mamtani et al. (2008) found that the CNVs of *CCL3L1–CCR5* were strong predictors for the overall risk of SLE and high autoantibody titers and lupus nephritis, and subjects with lupus nephritis differentially recruit leukocytes.

Receptors for the Fc portion of IgG are involved in the handling and clearance of immune complex and in the regulation of B cell activation during SLE development (Niederer et al., 2010). Willcocks et al. (2008) have found that low copy number of *Fc gamma receptor 3B* (*FCGR3B*), which is correlated with protein expression and immune complex uptake, was associated with SLE, implying that the association of this gene CNVs with SLE may influence protein expression and function and further confer risk for the predisposition of AIDs. Molokhia et al. (2011) identified that low copy number of *FCGR3B* was associated with the risk of SLE in the Afro-Caribbean population, but not in the African ancestry population. Chen et al. (2014) further showed that low copy numbers (less than two copies) of *FCGR3A* or *FCGR3B* were significantly associated with SLE, and high copy numbers (more than two copies) of *FCGR3A* were also related to SLE onset in the Taiwanese population. Another team, Barbosa et al. (2018), detected CNV at whole-genome level using a case–control design and showed that increased FCGR3B/ADAM3A copy number was a protective factor against SLE development. In addition, they, for the first time, uncovered heterozygous deletions overlapping the CFHR4, CFHR5, and HLA-DPB2 genes in SLE patients. Notably, different genetic manifestations attributing to different backgrounds may present different trends of gene CNV association with diseases.

Heat shock proteins 90 (HSP90) is a pivotal modulator of multiple innate and adaptive inflammatory processes (Tamura et al., 2012). HSP90 has two major cytosolic isoforms, HSP90AA1 and HSP90AB1. Zhang et al. (2019) showed that HSP90AB1 CNV was correlated with SLE in the Chinese Han population, especially in females, implying that HSP90AB1 CNV is involved in the pathomechanism and development of SLE.

RHEUMATOID ARTHRITIS

Rheumatoid arthritis (RA) is characterized by a massive tissue infiltration of inflammatory cells and affects approximately 1% of the adult population worldwide. Clinically, it mainly causes the chronic inflammation of synovial joints, which will result in the progressive destruction of the cartilage and bone (Achour et al., 2018). Many lines of data have implied the association of RA with genetic variations including SNPs and CNVs of several immune-related genes (McKinney et al., 2008; Graf et al., 2012; Olsson et al., 2012).

It is intriguing whether *FCGR3B* CNV is involved in RA. Graf et al. (2012) found a significant association between low *FCGR3B* copy number and RA (Rahbari et al., 2017). Chen et al. (2014) revealed a significant association of RA with low copy number of *FCGR3A*, but not *FCGR3B* in a Chinese cohort. Then, Rahbari et al. (2017) verified that RA patients from the UK had decreased copy number of *FCGR3B*. More recently, Ben Kilani et al. (2019) reported that genotypes without null allele of *FCGR3B* gene (copy numbers range from 1 to 3) were significantly associated with RA. In addition, increased *FCGR3B* copy number was only found in RA, and deletion of *FCGR3B* may have a protective effect on RA. The discrepancies on the correlation of RA with *FCGR3B* CNVs

may partly be due to different genetic backgrounds, and this requires further investigation in different ethnic populations.

CNVs of *CCL3L1* are potentially associated with RA. McKinney et al. (2008) uncovered that a higher than two copy number of *CCL3L1* gene was a risk factor for RA in the New Zealand population (Wouters et al., 2009). Mohamed et al. showed that *CCL3L1* copy number varies from 0 to 4 in the French population and from 0 to 7 in the Tunisian population. In addition, five copies of *CCL3L1* gene protect people from RA development in the Tunisian population (Ben Kilani et al., 2016). The results showing ethnic heterogeneity of *CCL3L1* among different populations suggest that *CCL3L1* is a susceptible factor for RA, and these differences of results may be due to genetic background and/or environmental differences existing among various populations and the different technical detection methods for genotyping multi-allelic CNVs. In the future, more reliable technologies are needed to explore the relationship between *CCL3L1* CNVs and RA in more populations to identify the results.

Inflammation is the central trait of RA, accompanied with the production of reactive oxygen species (ROS), which induces oxidative damages to cellular molecules including DNA and lipids, causing diverse cytotoxic products (Fahmi et al., 2002). The *neutrophil cytosolic factor 1* (*NCF1*) gene encodes one of five sub-units of the NADPH oxidase (NOX2) complex, which could produce ROS in the immune cells, including antigen presenting cells (APCs), phagocytes, etc. Olsson et al. (2012) found that RA patients are less likely to have an elevated copy number of *NCF1* compared to controls, which may suggest that a higher copy number of *NCF1* could be a protective factor for RA. Glutathione S-transferase (GST), a multifunctional enzyme that is required for the detoxification mechanism against ROS products, exerts a vital protective antioxidant function in cells against ROS aggression (Blackburn et al., 2006). In 2011, a Swedish study found that more than one copy of *glutathione S-transferase M1* (*GSTM1*), a member of the GST family, seems to be a risk factor for autoantibody-positive RA in non-smoking females of age older than 60 years, and GSTM1 acts as a protective factor in ACPA-negative smoking men, suggesting that the copy number of the *GSTM1* gene is correlated with the development and severity of RA (Lundström et al., 2011). In 2016, it was also found that *GSTM1* is deleted in Tunisian anti-cyclic citrullinated peptide (anti-CCP)-positive RA patients, although a genetic association of *GSTM1* CNV with predisposition to RA was not detected (Achour et al., 2018). The above data indicate that *GSTM1* CNVs do not influence the susceptibility to RA, but may have an effect on its severity because deletion of *GSTM1* could increase the risk of anti-CCP-positive RA. These interesting results showed that the CNVs of inflammatory-related genes may be involved in the development of RA through influencing the production of ROS and oxidative damages to cellular molecules.

INFLAMMATORY BOWEL DISEASE

Inflammatory bowel disease (IBD) is a kind of organ-specific inflammatory disease and has two main clinical subtypes, ulcerative colitis (UC) and Crohn's disease (CD). The occurrence of CD is 300 per 100,000 people in the population with European ancestry and increasing in other ethnic population. CD affects the gastrointestinal tract and has such symptoms, like diarrhea, abdominal pain, and aberrant weight loss (Cleynen et al., 2016). Studies have demonstrated the associations between *C4* gene and several AIDs (Hauptmann et al., 1974; Hou et al., 2013), and the team led by Cleynen et al. (2016) showed that CD cases tended to have lower *C4L* and higher *C4S* copies. They also found that serum C4 protein level was not significantly different between CD patients and controls, but CD patients with higher *C4* copy number may have higher serum C4 concentration (Cleynen et al., 2016). These results suggest that more *C4* copy number may lead to higher *C4* expression and there may be a dose–efficiency correlation between *C4* copy number and protein expression in CD patients. Asano et al. (2013) studied polymorphisms of *FCGR* genes in the Japanese population and found that *FCGR3B* copy number was related to susceptibility of UC.

Defensins are endogenous antimicrobial peptides to protect the intestinal mucosa against bacterial invasion. Fellermann et al. (2006) showed that healthy volunteers as well as UC patients have 2 to 10 copies of the *human beta-defensin 2* (*HBD-2*) gene with the median of 4 copies. However, patients with colonic CD have lower *HBD-2* copy than the controls. In addition, they also found that less than four copies of *HBD-2* gene were correlated with diminished mucosal HBD-2 mRNA expression (Fellermann et al., 2006). The *DEFA1A3* gene encodes alpha-defensins 1–3. Jespersgaard et al. (2011) found that a higher *DEFA1A3* copy number was related to CD, especially to colonic CD.

PSORIASIS

Psoriasis is a serious inflammatory disease of the skin, scalp, nails, and joints and has a prevalence of about 2% in the populations of developed countries (McKinney et al., 2008). Multiple studies have identified a strong genetic component in the development of psoriasis and demonstrated the relationship between CNV of some genes and psoriasis. Wu et al. (2014) found that Chinese patients with psoriasis vulgaris had a higher copy number (more than two copies) of *FCGR3B* compared to controls through a case–control study. Carpenter et al. (2011) found no association between *CCL3L1* copy number and psoriasis. Hollox et al. studied psoriasis in Dutch and German populations and found significant associations between higher genomic copy number for beta-defensin genes (DEFB4, SPAG11, DEFB103, DEFB104, DEFB105, DEFB106, and DEFB107) and the risk of psoriasis in both of these cohorts.

IL-22, which belongs to IL-10 cytokine family, has a significant proliferative effect on different cell lines and a role of immune regulation. Prans et al. (2013) showed that the copy number variation in exon 1 but not exon 5 of IL-22 gene was significantly correlated with the severity of psoriasis.

With the emergence of new technologies, more loci with CNVs have been identified to be associated with psoriasis. Bergboer et al. (2012) utilized a pooling approach, genome-

wide CNV analysis, and array comparative genomic hybridization to detect CNV variability in psoriasis and found that the absence of the late cornified envelope (*LCE*) gene cluster members *LCE3B* and *LCE3C* (LCE3C-LCE3B-del) was significantly associated with the predisposition of psoriasis in populations from Netherlands.

BEHCET'S DISEASE

Behcet's disease (BD) is an immune-mediated systemic inflammatory disorder involving non-granulomatous uveitis, recurrent oral and genital ulcers, as well as skin lesions (Yang et al., 2008a). It has been verified to be associated with the *HLA-B51* gene (Xu et al., 2015). Further studies have provided new insight into the pathogenesis of BD and attracted more attentions on whether complement is involved in BD. Hou et al. (2013) investigated the relationship of CNV of *C4A* and *C4B* with BD and found that the frequency of more than two copies of *C4A* was significantly increased in BD patients, and *C4A* CNV was an independent risk factor for BD. Moreover, *C4A* expression was significantly increased in BD patients with high *C4A* copy number than with low *C4A* copy number (Hou et al., 2013). Xu et al. (2015) found that BD patients have increased frequencies of more than two *C3* copies, and C5 CNV was associated with BD. Furthermore, interleukin-17 (IL-17) and IFN-gamma expressions were upregulated in BD patients with high *C3* copy number, but not in BD patients with high *C5* copy number (Xu et al., 2015). The results imply that there are indeed CNVs in complement-related genes, and the CNVs in these genes may be involved in the development of BD.

There were also some studies targeting the relationship between *FCGR3B* CNVs and BD. Black et al. (2012) found that CNV of the *FCGR3B* gene was associated with the risk of BD in the Iranian population. The risk of BD was decreased by 40% in people with less than two copies of *FCGR3B* and by 25% in people with more than two copies of *FCGR3B*, although these tendencies were not statistically significant. They concluded that no association exists between high or low copy number of *FCGR3B* and BD or its clinical features (Black et al., 2012). Of course, further studies are need to identify this result in other populations.

Toll-like receptors (TLRs), along with RIG-I-like receptor and NOD-like receptors, belong to the family of pattern recognition receptors and contain 10 functional members in human beings, namely, TLR1–10 (Medzhitov et al., 1997). TLRs have been known to play important functional roles in several AIDs and inflammatory disorders (Chang et al., 2004; Chang et al., 2006). Fang et al. (2015) uncovered that more than 98% of people tested have two copies of all *TLRs* except for *TLR7*. In addition, they found that compared to healthy controls, male BD patients had an increased frequency of more than one copy of TLR7, and female BD patients had an increased frequency of more than two copies of this gene. This research suggested that a high *TLR7* copy number may contribute to the pathogenesis of BD (Fang et al., 2015). Therefore, from this research, we included that *TLRs* also play a potential role in BD, and the mechanism underlying this still needs to be clarified.

Many interleukins play vital roles in BD development. Hou et al. (2015) showed that BD patients have increased frequencies of more than two copies of *IL-17F* and *IL-23A*, and after stratified by sex, the association just exists in male BD patients. IL-17F protein expression is positively correlated with its gene copy number, and higher *IL-17F* copies are associated with enhanced proliferation of peripheral blood mononuclear cells (PBMC) (Hou et al., 2015). This suggests that not only SNPs but also CNVs of interleukins are involved in the pathogenesis of BD. Liao et al. (2015) investigated the association of some transcript factors with BD and showed that high CNV of related orphan receptor (*RORC*) was associated with BD susceptibility, and low *Foxp3* CNV was correlated with female BD. In addition, individuals with high *RORC* copy number seem to have relatively high mRNA levels of RORC, IL-1β, and IL-6, but not Foxp3 (Liao et al., 2015).

The disturbed apoptosis has been reported to be involved in BD development. Yu et al. (2015)investigated whether CNVs of apoptosis-related genes, including *FAS*, *caspase8*, *caspase3*, and *BCL2*, were associated with BD in the Chinese population and showed that BD patients had an increased frequency of high *FAS* copy number, and BD patients with more than two copies of *FAS* had an increased mRNA expression of *FAS* in anti-CD3/CD28 antibody-stimulated $CD4^+$ T cells. Their results provided important evidence that high *FAS* copy number is involved in the pathogenesis of BD (Yu et al., 2015).

It is well-known that BD may be triggered by infectious agents in some genetically susceptible people. *DEFB4* CNV can affect the level of human beta-defensin 2, which is an inducible antimicrobial peptide. Park et al. (2011) found that *DEFB4* copy number was lower in BD samples than in controls without statistical significance, and *DEFB4* copy number was not associated with the clinical characteristics of BD. This suggests that *DEFB4* CNVs confer no risk for the susceptibility of BD (Park et al., 2011). In 2012, Ahn et al. (2012) found that 31.1% of samples had five copies of *DEFA1* with a mean of 5.4 ± 0.2. Although the distribution of *DEFA1* copy number is not different between BD patients and the controls, high *DEFA1* copy number is related to the intestinal involvement in BD, suggesting that a high *DEFA1* copy number may be associated with the development of intestinal involvement in BD (Ahn et al., 2012). Hitherto, the genes related to infectious agents may be involved in BD through changing the variation of copy number, and further researches are needed to be carried out to identify this.

There was another study that investigated the relationship between microRNA CNVs and BD, like miR-23a, miR-146a, and miR-301a. As a result, no association of CNVs of the above-mentioned miRNAs was observed in BD patients (Hou et al., 2016). However, whether other microRNA CNV is associated with BD still needs further studies to explore and clarify.

ANKYLOSING SPONDYLITIS

Ankylosing spondylitis (AS) is an inflammatory AID causing spondyloarthritis of the spine and sacroiliac joints and prevalent

mainly in men with a ratio of 10:1 at the age of 20–30 years old in respect to women (Chimenti et al., 2021). Because the exact pathology of AS is still unclear, there are incoming data about the relationship between CNVs of these genes and AS, including *CCL3L1*, *FCGR3A*, *FCGR3B*, *TLR7*, *UGT2B17*, *BMP8A*, and so on. Wang et al. (2016) found that AS patients had low copies (≦3) of *FCGR3A* and *FCGR3B* in the Chinese population, implying that a lower copy number of these two genes confers risk for the susceptibility of AS. The study of Dahmani et al. (2019) found that the proportion of AS patients with less than two copies of *FCGR3A* was higher in the Algerian population and that less than two copies of *FCGR3A* was only associated with HLA-B27-negative AS patients, suggesting that *FCGR3A* deletion has an independent effect on AS regarding HLA-B27 status. Their results also showed that *CCL3L1* and *FCGR3B* CNVs may not be involved in the predisposition of AS in the Algerian population (Dahmani et al., 2019). Wang et al. (2018) found that one copy of *TLR7* was related to AS in the Chinese population after Bonferroni correction and adjustment of age and sex, and less than one copy of *TLR7* confers risk for AS susceptibility in male patients, but is a protective factor in female AS patients. Uddin et al. (2013) conducted a genome-wide CNV analysis and found that *UGT2B17* copy number was increased in a large AS multiplex family. The *UGT2B17* gene encodes an enzyme that metabolizes steroid hormones such as testosterone and selected xenobiotics (Xue et al., 2008). *UGT2B17* copy number has been shown to be related to bone mineral density and involved in the pathogenesis of osteoporosis (Yang et al., 2008b). It is known that AS in patients is often accompanied with osteoporosis (Vosse et al., 2009). This may, to some extent, explain the underlying mechanism that gain in *UGT2B17* copy number could increase the risk for AS. Bone morphogenic protein 8A (BMP8A) plays multiple functions in the formation of a bone. Shahba et al. (2018) reported that the expression of BMP8A in PBMCs is decreased in AS patients, and *BMP8A* CNVs do not influence its transcription in PBMCs and are not associated with AS susceptibility in the Iranian population. Cai et al. (2015) found that the copy number of defensing-related gene *DEFB103* was in the range of two to six in both AS patients and controls, and it was not associated with AS. More studies in different populations are needed to further identify the relationship between these gene CNVs and BD.

$CD4^+$ T cells play pivotal roles in many AIDs, but whether CNVs of transcription factor genes in $CD4^+$ T cells are involved in AS remains poorly defined. Bai et al. (2016) investigated whether CNV of transcription factor genes in $CD4^+$ T cells including T-bet, GATA binding protein 3 (*GATA-3*), *RORC*, and fork-head box protein 3 (*FOXP3*) are associated with acute anterior uveitis (AAU) in the presence or absence of AS. They found that a higher *T-bet* copy number is more common in AAU+, AS+ and AAU+, AS− cases compared with healthy controls. Additionally, the frequency of AAU+, AS+ patients with high *GATA-3* copy is higher, and the proportion of female AAU+, AS+ patients with high *FOXP3* copy number is also higher than that of other populations, but the copy number of *RORC* is not correlated with AAU+, AS+ or AAU+, AS− patients (Bai et al., 2016).

People are also interested in the relationship between CNVs of various microRNAs (miRNAs). Yang et al. (2017) studied the association between CNVs of miRNAs and AS and found that the frequencies of AAU+, AS+ patients with low copy numbers of miR-143, miR-146a, miR-9-3, and miR-205 as well as high copy numbers of miR-301a and miR-23a all increased, and the frequencies of patients with AAU+, AS− with low copy number of miR-146a and high copy numbers of miR-23a and miR-205 are significantly different. In addition, they found that miR-9-3 mRNA expression is significantly decreased in AAU+, AS+ patients and positively correlated with its copy number (Yang et al., 2017).

PRIMARY SJOGREN'S SYNDROME

Primary Sjogren's syndrome (pSS) is characterized by the presence of circulating autoantibody (anti-Ro/SSA and anti-La/SSB), as well as the involvement of the exocrine glands (salivary and lacrimal gland), joint, and muscle (Nossent et al., 2012; Haldorsen et al., 2013). There were few studies on the issue of *FCGR3B* CNVs and *pSS*. Mamtani et al. (2010) showed that the median *FCGR3B* gene copy is two in the cohort of Spanish ancestry. The risk of pSS would increase if people carry less or more than two copies of *FCGR3B* (Mamtani et al., 2010). Nossent et al. (2012) conducted a case–control study and found that less than two copies of *FCGR3B* can confer risk for pSS in the Australian population, and low FCGR3B copy number is associated with the levels of rheumatoid factor (RF) titers and serum IgG, but not with anti-Ro ± La autoantibodies. They further identified that *FCGR3R* CNV is a genetic susceptibility factor for pSS (Liao et al., 2015). However, Haldorsen et al. (2013) showed no association of *FCGR3B* CNV with pSS in the Norway and Switzerland populations. To clarify, these controversial results need more studies in more populations with different ethnicities and regions.

TYPE 1 DIABETES MELLITUS

Type 1 diabetes mellitus (T1DM) is characterized by β cell destruction in the pancreas and the production of antibodies against β cells, with a high prevalence of 1 in 350 teenagers in the UK (Bluestone et al., 2010). There is a series of symptoms in T1DM, such as polydipsia, polyuria, polyphagia, and weight loss. *C4* is a gene of the highly variable complement pathway situated ~500 kb from DRB1 and DQB1 and strongly associated with diverse AIDs (Hauptmann et al., 1974; Hou et al., 2013; Cleynen et al., 2016), so some scientists also carried out numerous researches to uncover whether *C4* is associated with T1DM development. Kingery et al. (2012) found that higher *C4A* copy number tends to be correlated with the protection of residual β-cell function in new-onset T1DM patients, while lower *C4B* copy number is related to the end of disease remission at 9 months post diagnosis. Mason et al. (2014) explored the relationship between *C4* CNV and T1DM and found that individuals with T1DM have significantly fewer

copies of HERV-K (C4), one notable component of *C4*. About the relationship between *CCL3L1*, people also did a lot of work. McKinney et al. (2008) revealed an association between *CCL3L1* CNVs and T1DM in the Caucasian population from New Zealand, but the association was not statistically significant with *P* of 0.064. Then, the Wellcome Trust Case Control Consortium found no association between *CCL3L1* CNV and T1DM (Wellcome Trust Case Control Consortium et al., 2010). These work are consistent and implies that *CCL3L1* CNVs may be not associated with T1DM. *FCGR3B* CNVs are involved in the pathogenesis of several AIDs, such as SLE and RA, because its role in the clearance of immune complexes is impaired in these disease settings (Willcocks et al., 2008; Graf et al., 2012). Almal and Padh (2015) found that *FCGR3B* copy number in the Indian population varies significantly from zero to two per diploid genome in other populations, which helps us to understand the potential role of *FCGR3B* CNV and its association with AIDs in the Indian population.

Killer immunoglobulin-like receptors (KIRs) reside on the surface of natural killer cells to bind to their corresponding human leukocyte antigen (HLA) class I ligands. It is noted that KIRs are vital candidates for HLA-associated AIDs, including T1DM. Although Pontikos et al. (2014) did not find a relation of *KIR3DL1/3DS1* copy number to T1DM in the white European population, Grayson et al. (2010) utilized a more powerful genome-wide CNV analysis and found 39 CNVs either enriched or depleted in T1DM patients, including a deletion on chromosome 6p21, near an HLA-DQ allele. Their results indicated that both enrichment and depletion of these genes are high risk factors for developing T1DM, and genetic variants such as CNVs may contribute to the development of islet autoimmunity in T1DM (Grayson et al., 2010).

AUTOIMMUNE THYROID DISEASE

Autoimmune thyroid disease (AITD), which mainly includes Graves' disease (GD), Hashimoto's thyroiditis (HT), and Graves' ophthalmopathy (GO), is a kind of organ-specific AID with a prevalence of 5% of the overall population (Jin et al., 2018). GD is characterized by hyperthyroidism caused by positive autoantibodies against thyroid-stimulating hormone receptors (TSHR), and HT is often characterized by positive anti-thyroid peroxidase antibody (TPOAb), anti-thyroglobulin antibody (TgAb), and hypothyroidism. There have been many studies focusing upon the relationship between the CNVs of immune-related genes and the development of AITD, and lots of interesting results have been reported. The two earliest studies on CNVs and AITD were both published in 2011. One study aimed to explore the association between CNVs of immune regulatory genes and AITD and found no CNVs of *CD40* and *CTLA4* genes in GD cases and few *PTPN22* CNVs in two GD individuals (Huber et al., 2011). Liu et al. (2011) found that CNVs of *C4*, *C4A*, and *C4B* contribute to the predisposition of GD, but not GO. Liao et al. (2014) revealed that *TSHR* CNVs harbor the etiology of GD, but not GO. Besides, they also observed that only female GD patients have fewer *TLR7* copies, and there is no significant association between sex and *TLR7* CNVs (Liao et al., 2014). GD is a disease predominantly in females (Manji et al., 2006); therefore, they believed that *TLR7* CNVs affect the pathogenesis of female GD due to the gender-dependent immune response (Liao et al., 2014). In 2017, we conducted microarray to explore the profile of genes with CNVs in GD and found some genes with copy number gain. In addition, seven of these genes including *CFH*, *CFHR1*, *KIAA0125*, *UGT2B15*, *UGT2B17*, *TRY6*, and *CCL3L1* were chosen to further validate these findings in an expanded cohort. The results showed no correlation between CNVs of these genes and GD (Song et al., 2017). Jin et al. (2018) assessed CNVs of two immune-related genes *SIRPB1* and *TMEM91* in AITD and found that the distributions of *SIRPB1* copy number were different between AITD patients and the controls, implying *SIRPB1* is a risk factor for AITD. Guan et al. (2020) showed that the distribution of copy numbers of cell growth-related genes glypican-5 (*GPC5*), B9 domain-containing 2 (*B9D2*), as well as ankyrin repeat and suppressor of cytokine signaling (SOCS) box-containing protein 11 (*ASB11*) is different in AITD patients and the controls, and *GPC5* CNVs are risk factors for AITD. However, they did not find any association between their CNVs and the occurrence of AITD (Guan et al., 2020). The different relationship between CNVs of these genes and different sub-types of AITD implies the diverse genetic mechanisms underlying AITD. Overall, researchers have found that CNVs of several thyroid-susceptible genes are correlated with the development of AITD. Studies on the correlation between CNVs and AITD susceptibility inevitably deepen our understanding of the pathogenetic mechanism of AITD and further promote molecular diagnosis and therapies of AITD. These studies were mainly done in Chinese population; in the future, we will need further studies in more populations to identify these results.

FUTURE OUTLOOK

To further gain new and deep insights on the genetic mechanism of AIDs, we reviewed the association between different AIDs and CNVs of some genes with potential pivotal roles in the development of AIDs. This review provides an update evolving the view of copy number variations in AIDs. For the first, several CNVs are very common in diverse AIDs, implying these AIDs may potentially possess a similar genetic pathomechanism. Therefore, more association studies should be done on some other diseases when a certain link is identified between some CNVs and a specific AID. For the second, some CNVs are differently distributed in various diseases in different ethnic populations, suggesting that AIDs have their own different phenotypes and different genetic and/or environmental backgrounds among diverse populations. Herein, more researches aiming to uncover the relationship between environmental factors and diseases and the influences of environmental elements on immunity should be encouraged. For the third, with the continuous advancement in genotyping technology such as the high-throughput whole-genome sequencing method, more susceptible variants will be found.

Thus, further replication studies should be conducted to confirm the results of studies with different ethnic cohorts and independent populations.

For that, CNVs may be an important pathogenesis of AIDs; CNVs will also become an effective way to study the molecular mechanism of AIDs; and we can find molecular markers for genetic diagnosis or judgment of prognosis of this kind of disease. At the same time, it opens a new page for the research of AIDs and is becoming a new research hotspot. We believe that genetic diagnosis or judgment of prognosis based on CNVs will cover more AID spectrum and benefit a wider population. It is believed that with the high-throughput genome-wide CNV scanning platform and the new development of statistical calculation method, GWAS based on CNVs, a new genetic susceptibility marker, will become a powerful tool to study the genetic susceptibility of AIDs, just like the traditional GWAS based on SNPs and their haplotypes. These two complementary genetic markers will help us to understand the molecular mechanism, identify susceptible genes, and understand the relationship between genetic variations and disease phenotype of complex diseases, such as AIDs, which is of great significance. Simultaneously, to our hope, more functional experiments and more replication studies should be done and collected, and an entire autoimmune CNVs database should be set up, which can be searched easily and help us to understand the pathogenesis of AIDs much better. GWAS has already pointed toward genetic susceptibility loci and potential mechanisms of pathogenicity. Chromosomal microarrays have added little further information. It therefore seems unlikely that whole-genome sequencing alone will answer the necessary questions. Rather, genomic DNA approaches will likely need to be combined with other measures, such as RNA sequencing and/or proteomics to solve many of the remaining questions at hand.

Although CNVs describe the pathogenesis of AIDs from a new perspective, it is still early to explain the occurrence and development of complex diseases only by CNVs at the genomic level because the molecular mechanism of complex diseases at the chromosomal level is not completely clear. Due to the existence of multiple mechanisms of CNVs and the different effects of CNVs on molecular phenotype and gene expression, therefore, the interpretation of clinical significance and genetic mode of CNVs must be done more carefully, and it should be based on the comprehensive assessment of genomic variation.

AUTHOR CONTRIBUTIONS

R-HS designed the study. C-QG and JZ extracted the data. R-HS analyzed the data and wrote the first draft of the report. J-AZ revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The present work has received funding from the National Natural Science Foundation of China (Grant No. 81800696 and 81873636), the Shanghai University of Medicine and Health Sciences hundreds of Talented Teachers Project (No. B3-0200-20-311008-30), the Science and Technology Development Fund of Pudong New District Minsheng Scientific Research (Medical and Health) Project (No. PKJ2018-Y39). The grant number of the Talent Youth Cultivation Plan of Pudong New District is No. PWRq2020-11, then the grant number of the Shanghai Medical Key Specialty is No. ZK2019C09.

ACKNOWLEDGMENTS

We thank all authors who took the original studies here.

REFERENCES

Achour, Y., Ben Kilani, M. S., Ben Hamad, M., Marzouk, S., Mahfoudh, N., Bahloul, Z., et al. (2018). Measurement of Absolute Copy Number Variation of Glutathione S-Transferase M1 Gene by Digital Droplet PCR and Association Analysis in Tunisian Rheumatoid Arthritis Population. *J. Clin. Lab. Anal.* 32 (3), e22300. doi:10.1002/jcla.22300

Ahn, J. K., Cha, H.-S., Lee, J., Jeon, C. H., and Koh, E.-M. (2012). Correlation of DEFA1 Gene Copy Number Variation with Intestinal Involvement in Behcet's Disease. *J. Korean Med. Sci.* 27 (1), 107–109. doi:10.3346/jkms.2012.27.1.107

Almal, S. H., and Padh, H. (2015). Frequency Distribution of Autoimmunity associatedFCGR3Bgene Copy Number in Indian Population. *Int. J. Immunogenet.* 42 (1), 26–30. doi:10.1111/iji.12165

Asano, K., Matsumoto, T., Umeno, J., Hirano, A., Esaki, M., Hosono, N., et al. (2013). Impact of Allele Copy Number of Polymorphisms in FCGR3A and FCGR3B Genes on Susceptibility to Ulcerative Colitis. *Inflamm. Bowel Dis.* 19 (10), 2061–2068. doi:10.1097/mib.0b013e318298118e

Bai, L., Liu, Y., Hou, S., Liao, D., Kijlstra, A., and Yang, P. (2016). Association of T-Bet, GATA-3, RORC, and FOXP3 Copy Number Variations with Acute Anterior Uveitis with or without Ankylosing Spondylitis in Chinese Han. *Invest. Ophthalmol. Vis. Sci.* 57 (4), 1847–1852. doi:10.1167/iovs.15-17960

Barbosa, F. B., Simioni, M., Wiezel, C. E. V., Torres, F. R., Molck, M. C., Bonilla, M. M., et al. (2018). Copy Number Variation in the Susceptibility to Systemic Lupus Erythematosus. *PLoS One* 13 (11), e0206683. doi:10.1371/journal.pone.0206683

Ben Kilani, M. S., Achour, Y., Perea, J., Cornelis, F., Bardin, T., Chaudru, V., et al. (2016). Characterization of Copy Number Variants for CCL3L1 Gene in Rheumatoid Arthritis for French Trio Families and Tunisian Cases and Controls. *Clin. Rheumatol.* 35 (8), 1917–1922. doi:10.1007/s10067-015-3156-y

Ben Kilani, M. S., Cornélis, F., Olaso, R., Chaudru, V., and Petit-Teixeira, E. (2019). Investigation of Candidate Gene Copy Number Identifies FCGR3B as a Potential Biomarker for Rheumatoid Arthritis. *Clin. Exp. Rheumatol.* 37 (6), 923–928.

Bergboer, J. G. M., Umićević-Mirkov, M., Fransen, J., den Heijer, M., Franke, B., van Riel, P. L. C. M., et al. (2012). A Replication Study of the Association between Rheumatoid Arthritis and Deletion of the Late Cornified Envelope Genes LCE3B and LCE3C. *PLoS One* 7 (2), e32045. doi:10.1371/journal.pone.0032045

Black, R., Lester, S., Dunstan, E., Shahram, F., Nadji, A., Bayat, N., et al. (2012). Fc-Gamma Receptor 3B Copy Number Variation Is Not a Risk Factor for Behçet's Disease. *Int. J. Rheumatol.* 2012, 167096. doi:10.1155/2012/167096

Blackburn, A. C., Matthaei, K. I., Lim, C., Taylor, M. C., Cappello, J. Y., Hayes, J. D., et al. (2006). Deficiency of Glutathione Transferase Zeta Causes Oxidative

Stress and Activation of Antioxidant Response Pathways. *Mol. Pharmacol.* 69 (2), 650–657. doi:10.1124/mol.105.018911

Bluestone, J. A., Herold, K., and Eisenbarth, G. (2010). Genetics, Pathogenesis and Clinical Interventions in Type 1 Diabetes. *Nature* 464, 1293–1300. doi:10.1038/nature08933

Cai, G., Xia, Q., Fan, D., Li, X., Ding, N., Hu, Y., et al. (2015). Association between DEFB103 Gene Copy Number Variation and Ankylosing Spondylitis: a Case-Control Study. *Tissue Antigens* 86 (3), 195–198. doi:10.1111/tan.12630

Carpenter, D., Walker, S., Prescott, N., Schalkwijk, J., and Armour, J. A. (2011). Accuracy and Differential Bias in Copy Number Measurement of CCL3L1 in Association Studies with Three Auto-Immune Disorders. *BMC Genomics* 12, 418. doi:10.1186/1471-2164-12-418

Chang, J. H., McCluskey, P., and Wakefield, D. (2004). Expression of Toll-like Receptor 4 and its Associated Lipopolysaccharide Receptor Complex by Resident Antigen-Presenting Cells in the Human Uvea. *Invest. Ophthalmol. Vis. Sci.* 45 (6), 1871–1878. doi:10.1167/iovs.03-1113

Chang, J. H., McCluskey, P. J., and Wakefield, D. (2006). Toll-like Receptors in Ocular Immunity and the Immunopathogenesis of Inflammatory Eye Disease. *Br. J. Ophthalmol.* 90 (1), 103–108. doi:10.1136/bjo.2005.072686

Chen, J. Y., Wang, C. M., Chang, S. W., Cheng, C. H., Wu, Y. J. J., Lin, J. C., et al. (2014). Association of FCGR3A and FCGR3B Copy Number Variations with Systemic Lupus Erythematosus and Rheumatoid Arthritis in Taiwanese Patients. *Arthritis Rheumatol.* 66 (11), 3113–3121. doi:10.1002/art.38813

Chimenti, M. S., Perricone, C., D'Antonio, A., Ferraioli, M., Conigliaro, P., Triggianese, P., et al. (2021). Genetics, Epigenetics, and Gender Impact in Axial-Spondyloarthritis Susceptibility: An Update on Genetic Polymorphisms and Their Sex Related Associations. *Front. Genet.* 12, 671976. doi:10.3389/fgene.2021.671976

Cleynen, I., Konings, P., Robberecht, C., Laukens, D., Amininejad, L., Théâtre, E., et al. (2016). Genome-Wide Copy Number Variation Scan Identifies Complement Component C4 as Novel Susceptibility Gene for Crohn's Disease. *Inflamm. Bowel Dis.* 22 (3), 505–515. doi:10.1097/mib.0000000000000623

Dahmani, C. A., Benzaoui, A., Amroun, H., Zemani-Fodil, F., Petit-Teixeira, E., and Boudjema, A. (2019). Association Study of Copy Number Variants in CCL3L1, FCGR3A and FCGR3B Genes with Risk of Ankylosing Spondylitis in a West Algerian Population. *Int. J. Immunogenet.* 46 (6), 437–443. doi:10.1111/iji.12454

Fahmi, H., Pelletier, J. P., and Martel-Pelletier, J. (2002). PPARgamma Ligands as Modulators of Inflammatory and Catabolic Responses in Arthritis. An Overview. *J. Rheumatol.* 29 (1), 3–14.

Fang, J., Chen, L., Tang, J., Hou, S., Liao, D., Ye, Z., et al. (2015). Association between Copy Number Variations of TLR7 and Ocular Behcet's Disease in a Chinese Han Population. *Invest. Ophthalmol. Vis. Sci.* 56 (3), 1517–1523. doi:10.1167/iovs.14-15030

Fellermann, K., Stange, D. E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C. L., et al. (2006). A Chromosome 8 Gene-Cluster Polymorphism with Low Human Beta-Defensin 2 Gene Copy Number Predisposes to Crohn Disease of the colon. *Am. J. Hum. Genet.* 79 (3), 439–448. doi:10.1086/505915

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural Variation in the Human Genome. *Nat. Rev. Genet.* 7, 85–97. doi:10.1038/nrg1767

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., et al. (2005). The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. *Science* 307 (5714), 1434–1440. doi:10.1126/science.1101160

Graf, S. W., Lester, S., Nossent, J. C., Hill, C. L., Proudman, S. M., Lee, A., et al. (2012). Low Copy Number of the FCGR3B Gene and Rheumatoid Arthritis: a Case-Control Study and Meta-Analysis. *Arthritis Res. Ther.* 14 (1), R28. doi:10.1186/ar3731

Grayson, B. L., Smith, M. E., Thomas, J. W., Wang, L., Dexheimer, P., Jeffrey, J., et al. (2010). Genome-wide Analysis of Copy Number Variation in Type 1 Diabetes. *Plos one* 5 (11), e15393. doi:10.1371/journal.pone.0015393

Guan, Y., Liu, L., Jia, Q., Jin, X., Pang, Y., Meng, F., et al. (2020). The Role of Cell Growth-Related Gene Copy Number Variation in Autoimmune Thyroid Disease. *Biol. Trace Elem. Res.* 195 (2), 409–416. doi:10.1007/s12011-019-01880-7

Haldorsen, K., Appel, S., Le Hellard, S., Bruland, O., Brun, J. G., Omdal, R., et al. (2013). No Association of Primary Sjögren's Syndrome with Fcγ Receptor Gene Variants. *Genes Immun.* 14 (4), 234–237. doi:10.1038/gene.2013.12

Hauptmann, G., Grosshans, E., and Heid, E. (1974). Lupus Erythematosus Syndrome and Complete Deficiency of the Fourth Component of Complement. *Boll Ist Sieroter Milan* 53 (1), 228.

Hehir-Kwa, J. Y., Tops, B. B. J., and Kemmeren, P. (2018). The Clinical Implementation of Copy Number Detection in the Age of Next-Generation Sequencing. *Expert Rev. Mol. Diagn.* 18 (10), 907–915. doi:10.1080/14737159.2018.1523723

Hollox, E. J., Huffmeier, U., Zeeuwen, P. L. J. M., Palla, R., Lascorz, J., Rodijk-Olthuis, D., et al. (2008). Psoriasis Is Associated with Increased β-defensin Genomic Copy Number. *Nat. Genet.* 40 (1), 23–25. doi:10.1038/ng.2007.48

Hou, S., Qi, J., Liao, D., Zhang, Q., Fang, J., Zhou, Y., et al. (2013). Copy Number Variations of Complement Component C4 Are Associated with Behçet's Disease but Not with Ankylosing Spondylitis Associated with Acute Anterior Uveitis. *Arthritis Rheum.* 65 (11), 2963–2970. doi:10.1002/art.38116

Hou, S., Liao, D., Zhang, J., Fang, J., Chen, L., Qi, J., et al. (2015). Genetic Variations of IL17F and IL23A Show Associations with Behçet's Disease and Vogt-Koyanagi-Harada Syndrome. *Ophthalmology* 122 (3), 518–523. doi:10.1016/j.ophtha.2014.09.025

Hou, S., Ye, Z., Liao, D., Bai, L., Liu, Y., Zhang, J., et al. (2016). miR-23a, miR-146a and miR-301a Confer Predisposition to Vogt-Koyanagi-Harada Syndrome but Not to Behcet's Disease. *Sci. Rep.* 6, 20057. doi:10.1038/srep20057

Huber, A. K., Concepcion, E. S., Gandhi, A., Menconi, F., Smith, E. P., Keddache, M., et al. (2011). Analysis of Immune Regulatory Genes' Copy Number Variants in Graves' Disease. *Thyroid* 21 (1), 69–74. doi:10.1089/thy.2010.0262

Human Genome Structural Variation Working GroupEichler, E. E., Nickerson, D. A., Altshuler, D., Bowcock, A. M., Brooks, L. D., et al. (2007). Completing the Map of Human Genetic Variation. *Nature* 447, 161–165. doi:10.1038/447161a

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., et al. (2004). Detection of Large-Scale Variation in the Human Genome. *Nat. Genet.* 36 (9), 949–951. doi:10.1038/ng1416

Jespersgaard, C., Fode, P., Dybdahl, M., Vind, I., Nielsen, O. H., Csillag, C., et al. (2011). Alpha-defensin DEFA1A3 Gene Copy Number Elevation in Danish Crohn's Disease Patients. *Dig. Dis. Sci.* 56 (12), 3517–3524. doi:10.1007/s10620-011-1794-8

Jia, D., Wei, L., Guo, W., Zha, R., Bao, M., Chen, Z., et al. (2011). Genome-wide Copy Number Analyses Identified Novel Cancer Genes in Hepatocellular Carcinoma. *Hepatology* 54 (4), 1227–1236. doi:10.1002/hep.24495

Jiang, Y., Mao, X., Song, R., Mu, K., Yang, Y., and Zhang, J.-a. (2021). Psoriasis Susceptibility 1 Candidate 1 (PSORS1C1) Polymorphism Is Associated with Autoimmune Thyroid Disease in a Chinese Han Population. *Immunol. Invest.*, 1–10. doi:10.1080/08820139.2021.1923734

Jin, X., Guan, Y., Shen, H., Pang, Y., Liu, L., Jia, Q., et al. (2018). Copy Number Variation of Immune-Related Genes and Their Association with Iodine in Adults with Autoimmune Thyroid Diseases. *Int. J. Endocrinol.* 2018, 1705478. doi:10.1155/2018/1705478

Kingery, S. E., Wu, Y. L., Zhou, B., Hoffman, R. P., and Yu, C. Y. (2012). Gene CNVs and Protein Levels of Complement C4A and C4B as Novel Biomarkers for Partial Disease Remissions in New-Onset Type 1 Diabetes Patients. *Pediatr. Diabetes* 13 (5), 408–418. doi:10.1111/j.1399-5448.2011.00836.x

Liao, W.-L., Wan, L., Wang, T.-Y., Chen, C.-C., Tse, S.-S., Lu, C.-H., et al. (2014). Association of TLR7 and TSHR Copy Number Variation with Graves' Disease and Graves' Ophthalmopathy in Chinese Population in Taiwan. *BMC Ophthalmol.* 14, 15. doi:10.1186/1471-2415-14-15

Liao, D., Hou, S., Zhang, J., Fang, J., Liu, Y., Bai, L., et al. (2015). Copy Number Variants and Genetic Polymorphisms in TBX21, GATA3, Rorc, Foxp3 and Susceptibility to Behcet's Disease and Vogt-Koyanagi-Harada Syndrome. *Sci. Rep.* 5, 9511. doi:10.1038/srep09511

Liu, Y.-H., Wan, L., Chang, C.-T., Liao, W.-L., Chen, W.-C., Tsai, Y., et al. (2011). Association between Copy Number Variation of Complement Component C4 and Graves' Disease. *J. Biomed. Sci.* 18, 71. doi:10.1186/1423-0127-18-71

Lundström, E., Hartshorne, T., Li, K., Lindblad, S., Wick, M. C., Bengtsson, C., et al. (2011). Effects of GSTM1 in Rheumatoid Arthritis; Results from the Swedish EIRA Study. *PLoS One* 6 (3), e17880. doi:10.1371/journal.pone.0017880

Mack, M., Bender, K., and Schneider, P. M. (2004). Detection of Retroviral Antisense Transcripts and Promoter Activity of the HERV-K(C4) Insertion in the MHC Class III Region. *Immunogenetics* 56, 321–332. doi:10.1007/s00251-004-0705-y

Mamtani, M., Rovin, B., Brey, R., Camargo, J. F., Kulkarni, H., Herrera, M., et al. (2008). CCL3L1 Gene-Containing Segmental Duplications and Polymorphisms in CCR5 Affect Risk of Systemic Lupus Erythaematosus. *Ann. Rheum. Dis.* 67 (8), 1076–1083. doi:10.1136/ard.2007.078048

Mamtani, M., Anaya, J.-M., He, W., and Ahuja, S. K. (2010). Association of Copy Number Variation in the FCGR3B Gene with Risk of Autoimmune Diseases. *Genes Immun.* 11 (2), 155–160. doi:10.1038/gene.2009.71

Manji, N., Carr-Smith, J. D., Boelaert, K., Allahabadia, A., Armitage, M., Chatterjee, V. K., et al. (2006). Influences of Age, Gender, Smoking, and Family History on Autoimmune Thyroid Disease Phenotype. *J. Clin. Endocrinol. Metab.* 91 (12), 4873–4880. doi:10.1210/jc.2006-1402

Mason, M. J., Speake, C., Gersuk, V. H., Nguyen, Q. A., O'Brien, K. K., Odegard, J. M., et al. (2014). Low HERV-K(C4) Copy Number Is Associated with Type 1 Diabetes. *Diabetes* 63 (5), 1789–1795. doi:10.2337/db13-1382

McCarroll, S. A., and Altshuler, D. M. (2007). Copy-number Variation and Association Studies of Human Disease. *Nat. Genet.* 39 (7 Suppl. l), S37–S42. doi:10.1038/ng2080

McKinney, C., Merriman, M. E., Chapman, P. T., Gow, P. J., Harrison, A. A., Highton, J., et al. (2008). Evidence for an Influence of Chemokine Ligand 3-like 1 (CCL3L1) Gene Copy Number on Susceptibility to Rheumatoid Arthritis. *Ann. Rheum. Dis.* 67 (3), 409–413. doi:10.1136/ard.2007.075028

Medzhitov, R., Preston-Hurlburt, P., and Janeway, C. A., Jr. (1997). A Human Homologue of the Drosophila Toll Protein Signals Activation of Adaptive Immunity. *Nature* 388 (6640), 394–397. doi:10.1038/41131

Molokhia, M., Fanciulli, M., Petretto, E., Patrick, A. L., McKeigue, P., Roberts, A. L., et al. (2011). FCGR3B Copy Number Variation Is Associated with Systemic Lupus Erythematosus Risk in Afro-Caribbeans. *Rheumatology* 50 (7), 1206–1210. doi:10.1093/rheumatology/keq456

Mulvihill, E., Ardoin, S., Thompson, S. D., Zhou, B., Yu, G. R., King, E., et al. (2019). Elevated Serum Complement Levels and Higher Gene Copy Number of Complement C4B Are Associated with Hypertension and Effective Response to Statin Therapy in Childhood-Onset Systemic Lupus Erythematosus (SLE). *Lupus Sci. Med.* 6 (1), e000333. doi:10.1136/lupus-2019-000333

Niederer, H. A., Clatworthy, M. R., Willcocks, L. C., and Smith, K. G. C. (2010). FcγRIIB, FcγRIIIB, and Systemic Lupus Erythematosus. *Ann. N. Y Acad. Sci.* 1183, 69–88. doi:10.1111/j.1749-6632.2009.05132.x

Nossent, J. C., Rischmueller, M., and Lester, S. (2012). Low Copy Number of the Fc-γ Receptor 3B Gene FCGR3B Is a Risk Factor for Primary Sjögren's Syndrome. *J. Rheumatol.* 39 (11), 2142–2147. doi:10.3899/jrheum.120294

Olsson, L. M., Nerstedt, A., Lindqvist, A.-K., Johansson, Å. C. M., Medstrand, P., Olofsson, P., et al. (2012). Copy Number Variation of the GeneNCF1Is Associated with Rheumatoid Arthritis. *Antioxid. Redox Signal.* 16 (1), 71–78. doi:10.1089/ars.2011.4013

Park, J. J., Oh, B. R., Kim, J. Y., Park, J. A., Kim, C., Lee, Y. J., et al. (2011). Copy Number Variation of β-defensin Genes in Behçet's Disease. *Clin. Exp. Rheumatol.* 29 (4 Suppl. 67), S20–S23.

Pereira, K. M. C., Perazzio, S., Faria, A. G. A., Moreira, E. S., Santos, V. C., Grecco, M., et al. (2019). Impact of C4, C4A and C4B Gene Copy Number Variation in the Susceptibility, Phenotype and Progression of Systemic Lupus Erythematosus. *Adv. Rheumatol.* 59 (1), 36. doi:10.1186/s42358-019-0076-6

Pontikos, N., Smyth, D. J., Schuilenburg, H., Howson, J. M., Walker, N. M., Burren, O. S., et al. (2014). A Hybrid qPCR/SNP Array Approach Allows Cost Efficient Assessment of KIR Gene Copy Numbers in Large Samples. *BMC Genomics* 15, 274. doi:10.1186/1471-2164-15-274

Prans, E., Kingo, K., Traks, T., Silm, H., Vasar, E., and Kõks, S. (2013). Copy Number Variations in IL22 Gene Are Associated with Psoriasis Vulgaris. *Hum. Immunol.* 74 (6), 792–795. doi:10.1016/j.humimm.2013.01.006

Rahbari, R., Zuccherato, L. W., Tischler, G., Chihota, B., Ozturk, H., Saleem, S., et al. (2017). Understanding the Genomic Structure of Copy-Number Variation of the Low-Affinity Fcγ Receptor Region Allows Confirmation of the Association ofFCGR3BDeletion with Rheumatoid Arthritis. *Hum. Mutat.* 38 (4), 390–399. doi:10.1002/humu.23159

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global Variation in Copy Number in the Human Genome. *Nature* 444 (7118), 444–454. doi:10.1038/nature05329

Saadati, H. R., Wittig, M., Helbig, I., Häsler, R., Anderson, C. A., Mathew, C. G., et al. (2016). Genome-wide Rare Copy Number Variation Screening in Ulcerative Colitis Identifies Potential Susceptibility Loci. *BMC Med. Genet.* 17, 26. doi:10.1186/s12881-016-0289-z

Shahba, S., Jafari Shakib, R., Jamshidi, A., Vojdanian, M., Akhtari, M., Aslani, S., et al. (2018). Association Study of Copy Number Variation in BMP8A Gene with the Risk of Ankylosing Spondylitis in Iranian Population. *J. Cel Biochem.* doi:10.1002/jcb.28120

Song, R.-h., Shao, X.-q., Li, L., Wang, W., and Zhang, J.-a. (2017). Copy Number Variations Exploration of Multiple Genes in Graves' Disease. *Medicine (Baltimore)* 96 (4), e5866. doi:10.1097/md.0000000000005866

Song, R.-h., Zhao, J., Gao, C.-q., Qin, Q., and Zhang, J.-a. (2021). Inclusion of ALKBH5 as a Candidate Gene for the Susceptibility of Autoimmune Thyroid Disease. *Adv. Med. Sci.* 66 (2), 351–358. doi:10.1016/j.advms.2021.07.006

Song, R.-h., Liu, X.-r., Gao, C.-q., Du, P., and Zhang, J.-a. (2021). METTL3 Gene Polymorphisms Contribute to Susceptibility to Autoimmune Thyroid Disease. *Endocrine* 72 (2), 495–504. doi:10.1007/s12020-020-02503-1

Tamura, Y., Torigoe, T., Kutomi, G., Hirata, K., and Sato, N. (2012). New Paradigm for Intrinsic Function of Heat Shock Proteins as Endogenous Ligands in Inflammation and Innate Immunity. *Curr. Mol. Med.* 12 (9), 1198–1206. doi:10.2174/156652412803306710

Tomer, Y., and Davies, T. F. (2003). Searching for the Autoimmune Thyroid Disease Susceptibility Genes: from Gene Mapping to Gene Function. *Endocr. Rev.* 24 (5), 694–717. doi:10.1210/er.2002-0030

Uddin, M., Maksymowych, W. P., Inman, R., Gladman, D., Munn, A., Yazdani, R., et al. (2013). UGT2B17 Copy Number Gain in a Large Ankylosing Spondylitis Multiplex Family. *BMC Genet.* 14, 67. doi:10.1186/1471-2156-14-67

Voll, S. L., Boot, E., Butcher, N. J., Cooper, S., Heung, T., Chow, E. W. C., et al. (2017). Obesity in Adults with 22q11.2 Deletion Syndrome. *Genet. Med.* 19 (2), 204–208. doi:10.1038/gim.2016.98

Vosse, D., Landewe, R., van der Heijde, D., van der Linden, S., van Staa, T.-P., and Geusens, P. (2009). Ankylosing Spondylitis and the Risk of Fracture: Results from a Large Primary Care-Based Nested Case-Control Study. *Ann. Rheum. Dis.* 68 (12), 1839–1842. doi:10.1136/ard.2008.100503

Wang, L., Yang, X., Cai, G., Xin, L., Xia, Q., Zhang, X., et al. (2016). Association Study of Copy Number Variants in FCGR3A and FCGR3B Gene with Risk of Ankylosing Spondylitis in a Chinese Population. *Rheumatol. Int.* 36 (3), 437–442. doi:10.1007/s00296-015-3384-0

Wang, M., Xu, S., Zhang, X., Chen, M., Han, R., Hu, X., et al. (2018). Association of TLR7 Gene Copy Number Variations with Ankylosing Spondylitis in a Chinese Population: a Case Control Study. *Clin. Exp. Rheumatol.* 36 (5), 814–819.

Wellcome Trust Case Control ConsortiumCraddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., et al. (2010). Genome-wide Association Study of CNVs in 16,000 Cases of Eight Common Diseases and 3,000 Shared Controls. *Nature* 464 (7289), 713–720. doi:10.1038/nature08979

Willcocks, L. C., Lyons, P. A., Clatworthy, M. R., Robinson, J. I., Yang, W., Newland, S. A., et al. (2008). Copy Number of FCGR3B, Which Is Associated with Systemic Lupus Erythematosus, Correlates with Protein Expression and Immune Complex Uptake. *J. Exp. Med.* 205 (7), 1573–1582. doi:10.1084/jem.20072413

Wouters, D., van Schouwenburg, P., van der Horst, A., de Boer, M., Schooneman, D., Kuijpers, T. W., et al. (2009). High-throughput Analysis of the C4 Polymorphism by a Combination of MLPA and Isotype-specific ELISA's. *Mol. Immunol.* 46, 592–600. doi:10.1016/j.molimm.2008.07.028

Wu, Y. L., Yang, Y., Chung, E. K., Zhou, B., Kitzmiller, K. J., Savelli, S. L., et al. (2008). Phenotypes, Genotypes and Disease Susceptibility Associated with Gene Copy Number Variations: Complement C4 CNVs in European American Healthy Subjects and Those with Systemic Lupus Erythematosus. *Cytogenet. Genome Res.* 123, 131–141. doi:10.1159/000184700

Wu, Y., Zhang, Z., Tao, L., Chen, G., Liu, F., Wang, T., et al. (2014). A High Copy Number of FCGR3B Is Associated with Psoriasis Vulgaris in Han Chinese. *Dermatology* 229 (2), 70–75. doi:10.1159/000360160

Xu, D., Hou, S., Zhang, J., Jiang, Y., Kijlstra, A., and Yang, P. (2015). Copy Number Variations and Gene Polymorphisms of Complement Components in Ocular Behcet's Disease and Vogt-Koyanagi-Harada Syndrome. *Sci. Rep.* 5, 12989. doi:10.1038/srep12989

Xue, Y., Sun, D., Daly, A., Yang, F., Zhou, X., Zhao, M., et al. (2008). Adaptive Evolution of UGT2B17 Copy-Number Variation. *Am. J. Hum. Genet.* 83 (3), 337–346. doi:10.1016/j.ajhg.2008.08.004

Yang, Y., Chung, E. K., Wu, Y. L., Savelli, S. L., Nagaraja, H. N., Zhou, B., et al. (2007). Gene Copy-Number Variation and Associated Polymorphisms of Complement Component C4 in Human Systemic Lupus Erythematosus (SLE): Low Copy Number Is a Risk Factor for and High Copy Number Is a Protective Factor against SLE Susceptibility in European Americans. *Am. J. Hum. Genet.* 80 (6), 1037–1054. doi:10.1086/518257

Yang, P., Fang, W., Meng, Q., Ren, Y., Xing, L., and Kijlstra, A. (2008). Clinical Features of Chinese Patients with Behçet's Disease. *Ophthalmology* 115, 312–318. doi:10.1016/j.ophtha.2007.04.056

Yang, T.-L., Chen, X.-D., Guo, Y., Lei, S.-F., Wang, J.-T., Zhou, Q., et al. (2008). Genome-wide Copy-Number-Variation Study Identified a Susceptibility Gene, UGT2B17, for Osteoporosis. *Am. J. Hum. Genet.* 83 (6), 663–674. doi:10.1016/j.ajhg.2008.10.006

Yang, L., Du, L., Yue, Y., Huang, Y., Zhou, Q., Cao, S., et al. (2017). miRNA Copy Number Variants Confer Susceptibility to Acute Anterior Uveitis with or without Ankylosing Spondylitis. *Invest. Ophthalmol. Vis. Sci.* 58 (4), 1991–2001. doi:10.1167/iovs.16-21047

Yim, S.-H., Kim, T.-M., Hu, H.-J., Kim, J.-H., Kim, B.-J., Lee, J.-Y., et al. (2010). Copy Number Variations in East-Asian Population and Their Evolutionary and Functional Implications. *Hum. Mol. Genet.* 19, 1001–1008. doi:10.1093/hmg/ddp564

Yim, S.-H., Jung, S.-H., Chung, B., and Chung, Y.-J. (2015). Clinical Implications of Copy Number Variations in Autoimmune Disorders. *Korean J. Intern. Med.* 30 (3), 294–304. doi:10.3904/kjim.2015.30.3.294

Yu, H., Luo, L., Wu, L., Zheng, M., Zhang, L., Liu, Y., et al. (2015). FASGene Copy Numbers Are Associated with Susceptibility to Behçet Disease and VKH Syndrome in Han Chinese. *Hum. Mutat.* 36 (11), 1064–1069. doi:10.1002/humu.22829

Zhang, M., Gu, Y., Huang, S., Lou, Q., Xie, Q., Xu, Z., et al. (2019). Copy Number Variations and Polymorphisms in HSP90AB1 and Risk of Systemic Lupus Erythematosus and Efficacy of Glucocorticoids. *J. Cel Mol. Med.* 23 (8), 5340–5348. doi:10.1111/jcmm.14410

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Song, Gao, Zhao and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

SYSTEMATIC REVIEW
published: 21 January 2022
doi: 10.3389/fgene.2021.772032

Association Between TNF-α-308, +489, –238 Polymorphism, and COPD Susceptibility: An Updated Meta-Analysis and Trial Sequential Analysis

Zhiyu Xia[1†], *Yufei Wang*[1†], *Fu Liu*[1†], *Hongxin Shu*[1] *and Peng Huang*[2,3*]

[1]The Second Clinical Medical School, Nanchang University, Nanchang, China, [2]Center for Evidence-Based Medicine, School of Public Health, Nanchang University, Nanchang, China, [3]Jiangxi Province Key Laboratory of Preventive Medicine, School of Public Health, Nanchang University, Nanchang, China

The tumor necrosis factor alpha (TNF-α) polymorphism may play an important role in chronic obstructive pulmonary disease (COPD) susceptibility. However, the results are still inconclusive. Eligible studies were searched in Cochrane Library database, EMBASE, Pudmed, Web of science, China National Knowledge Infrastructure, and Wanfang database. Finally, a total of 27 case-control studies with 3473 COPD cases and 4935 controls were included in the present analysis. We also performed trial sequential analysis (TSA) to confirm our results. Overall, association between TNF-α-308G/A polymorphism and COPD susceptibility was identified in allelic model (A vs. G, OR = 1.21, 95%CI: 1.01–1.45, p = 0.04) when smoking status was not adjusted. In ethnicity subgroup analysis, we found that the TNF-α -308G/A polymorphism was associated to COPD among Asians (GA vs. GG, OR = 1.35, 95%CI: 1.04–1.77, p = 0.02) when smoking status was not adjusted. However, no significant association was found in Asian smokers or Caucasian smokers. In conclusion, our study suggest that TNF-α-308 GA genotype is related to COPD in the Asian population. In addition, the TNF-α+489G/A, - 238G/A variants do not increase the risk of COPD.

Systematic Review Registration: https://www.crd.york.ac.uk/PROSPERO/, identifier CRD42021273980.

Keywords: COPD, single nucleotide polymorphism, meta-analysis, trial sequential analysis, TNF-α

OPEN ACCESS

Edited by:
William C Cho,
QEH, Hong Kong SAR, China

Reviewed by:
Pentti Nieminen,
University of Oulu, Finland
Gloria Pérez-Rubio,
Instituto Nacional de Enfermedades Respiratorias (INER), Mexico

***Correspondence:**
Peng Huang
huangpengncu@163.com

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *07 September 2021*
Accepted: *01 December 2021*
Published: *21 January 2022*

Citation:
Xia Z, Wang Y, Liu F, Shu H and Huang P (2022) Association Between TNF-α-308, +489, –238 Polymorphism, and COPD Susceptibility: An Updated Meta-Analysis and Trial Sequential Analysis. Front. Genet. 12:772032. doi: 10.3389/fgene.2021.772032

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is characterized by progressive airflow obstruction and bronchial hyperresponsiveness, with increasing morbidity, mortality, and resource utilization worldwide (Celli et al., 2015). According to the latest figures, COPD has been estimated to become the fourth leading cause of death in the world by 2040 (Foreman et al., 2018).

Tumor necrosis factor alpha (TNF-α) gene is found on chromosome 6p21.33 (Nedwin et al., 1985). The TNF-α-308 G/A polymorphism appears to affect the multifunctional proinflammatory cytokine (Zhou et al., 2016). The signals are mediated through two transmembrane receptors, TNFR1 and TNFR2, to regulate the inflammatory cell functions such as cell proliferation, survival,

differentiation, and apoptosis (Parameswaran and Patial, 2010). Some epidemiologic research on the association between TNF-α+489 G/A and -308 G/A polymorphism and COPD susceptibility have been performed (Cui et al., 2015; Zhang et al., 2016; Reséndiz-Hernández et al., 2018; Yu et al., 2021), but the results are still inconsistent. Additionally, the scarcity of meta-analysis about the relationship between TNF-α-238G/A and COPD risk impels us to study the issue in this article. In order to do this, we pooled 27 original studies and followed a stricter criterion to obtain an explicit understanding. Subgroup analysis based on ethnicity was also performed to gain a comprehensive view.

MATERIALS AND METHODS

We followed the PRISMA guidelines and registered the review protocol on PROSPERO (CRD42021273980, September 21, 2021).

Literature Retrieval

A comprehensive search of PubMed, EMBASE, Web of science, Cochrane Library database, Wanfang databases, and China National Knowledge Infrastructure were performed to find eligible studies. We retrieved articles using search strategies: "TNF-α or Tumor Necrosis Factor-α" and "COPD or COAD or chronic obstructive lung disease or chronic obstructive pulmonary disease or chronic obstructive airway disease" and "polymorphism or SNP or genotype or variant" published before July 25, 2021. In addition, we reviewed all references cited in the identified articles.

Inclusion and Exclusion Criteria

The included studies had to meet all four criteria: (Celli et al., 2015): Case-control or cohort studies, (Foreman et al., 2018), Availability of sufficient information to the odds rations (ORs) with 95% confidence intervals (CIs), (Nedwin et al., 1985), Alleles or genotypes frequencies in control groups and case groups could be received from the studies, and (Zhou et al., 2016) Literature in English. Accordingly, the exclusion criteria were as follows: (Celli et al., 2015): Results not focused on TNF-α-308, +489, or −238; (Foreman et al., 2018); Meta-analyses, letters, reviews, editorial articles, and studies that duplicated previous publications; (Nedwin et al., 1985); Detailed genotype data were not provided; or (Zhou et al., 2016) the genotype distribution of control was not in accordance with the Hardy-Weinberg equilibrium (HWE).

Data Extraction and Literature Quality Evaluation

Two investigators retrieved data from the included studies separately, and all discordances were discussed to reach an agreement. The extraction information mainly includes the first author's name, year of publication, country of origin, ethnicity, source of control, genotype method, the overall number of cases and controls, and genotype distribution of three TNF-α gene variants. The Newcastle-Ottawa Scale (NOS) was applied to estimate the quality of included articles, and the studies were classified as high quality (scores ≥7 stars) or low quality (scores<7 stars).

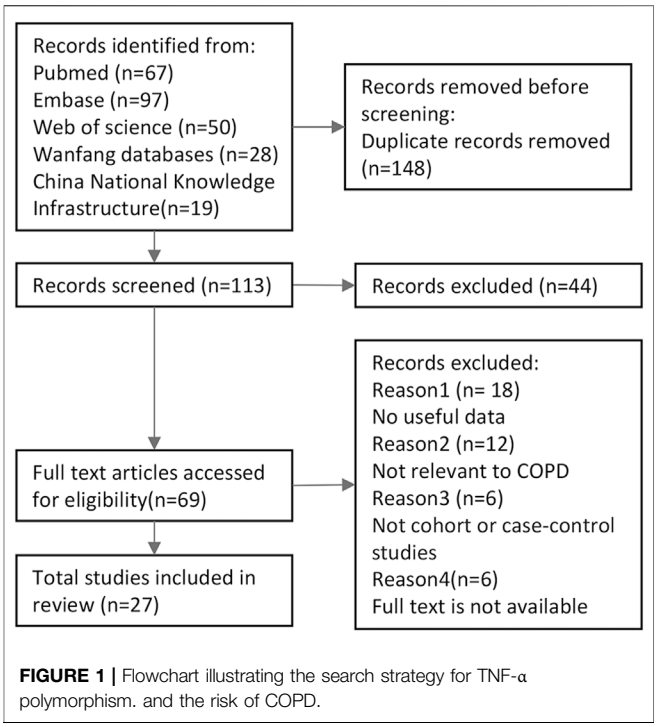


FIGURE 1 | Flowchart illustrating the search strategy for TNF-α polymorphism. and the risk of COPD.

Statistical Analyses

The odds ratio (OR) and its 95% confidence interval (95%CI) were used as an effect size to assess the risk of COPD caused by TNF-α polymorphism, as well as the strength of association between them. Z-test was used to estimate the statistical significance of pooled ORs. There were five genetic comparison models: allelic model (A vs. G), heterozygote model (GA vs. GG), homozygote model (AA vs. GG), dominant model (GA + AA vs. GG), and recessive model (AA vs. GG + GA). We checked the heterogeneity assumption by the Q-test. The outcome ($p > 0.1$ and $I^2<50\%$) indicated no heterogeneity among studies, and fixed-effects model (the Mantel-Haenszel method) was applied. Otherwise, when $I^2 \geq 50\%$ or $p \leq 0.10$, we performed random-effects model (the DerSimonian and Laird method). Publication bias of the literature was evaluated by funnel plots and Egger's test, and one-way sensitivity analyses were performed to assess the stability of the results.

The pooled OR and its 95%CI were calculated by Review Manager 5.4.1 software. All the *p*-values are two-sided.

Trial Sequential Analysis

We used the TSA v0.9.5.10 Beta software to perform the trial sequential analysis. Our study set the relative risk reduction to 20%, the first type of error $\alpha = 0.05$, and power = 80% to evaluate required information size (RIS) and the trial sequential

TABLE 1 | Main characters of studies included in this meta-analysis.

0.5	Country	Ethnicity	Source of control	Genotype methods	Cases/Controls	Genotype distribution								NOS	HWE
						Case				Control					
						TNF-α 308									
						Smoking status	GG	GA	AA	Smoking status	GG	GA	AA		
(Keatings et al., 2000)	Britain	Caucasian	HB	TaqMan	106/99	Yes	62	38	6	Yes	59	37	3	6	0.871
(Higham et al., 2000)	Britain	Caucasian	HB	RCR-RFLP	86/262	Yes	62	22	2	Mixed	181	74	7	8	0.998
					86/63					Yes	45	17	1	8	<0.05
(Ishii et al., 2000)	Japan	Asian	HB	PCR	53/65	Yes	52	1	0	Yes	64	1	0	7	1.000
(Sakao et al., 2001)	Japan	Asian	HB	PCR- RFLP	106/239	Yes	77	23	6	Mixed	209	22	8	7	0.09
					106/110					Yes	96	10	4	7	<0.05
(Küçükaycan et al., 2002)	Netherlands	Caucasian	PB	PCR	163/335	Mixed	113	49	1	Unknown	237	91	7	6	0.589
(Ferrarotti et al., 2003)	Italy	Caucasian	HB	PCR	63/86	Mixed	54	9	0	Yes	72	14	0	6	0.578
(Chierakul et al., 2005)	Thailand	Asian	HB	PCR	57/183	Yes	48	9	0	Mixed	162	21	0	6	0.585
					57/67					Yes	57	10	0	7	0.598
(Hegab et al., 2005)	Japan	Asian	HB	PCR-RFLP	88/61	Yes	86	2	0	Yes	61	0	0	6	1.000
		Caucasian	HB	PCR-RFLP	106/72	Yes	91	14	1	Yes	57	14	1	6	0.839
(Seifart et al., 2005)	Germany	Caucasian	PB	PCR	113/243	Mixed	73	36	4	Mixed	171	67	4	6	0.857
(Jiang et al., 2005)	China	Asian	PB	PCR-RFLP	101/96	Mixed	90	10	1	Mixed	90	6	0	7	0.597
(Papatheodorou et al., 2007)	Greece	Caucasian	PB	PCR	116/202	Mixed	101	14	1	Yes	169	29	4	6	0.429
(Danilko et al., 2007)	Russia	Asian	HB	PCR	319/381	Unknown	216	101	2	Unknown	273	103	5	7	0.071
(Trajkov et al., 2009)	France	Caucasian	HB	PCR-SSP	60/301	Mixed	45	14	1	Unknown	231	66	4	8	0.952
(Gingo et al., 2008)	America	Caucasian	PB	PCR	298/125	Yes	220	67	11	Yes	105	18	2	7	0.283
(Hsieh et al., 2008)	America	Asian	HB	PCR-RFLP	30/115	Mixed	23	6	1	Mixed	96	18	1	7	0.838
(Stankovic et al., 2009)	Serbia	Caucasian	PB	PCR-RFLP	97/102	Mixed	79	17	1	Mixed	71	28	3	6	0.996
(Chen et al., 2010)	China	Asian	HB	PCR-RFLP	145/139	Yes	117	28	0	Yes	109	27	3	7	0.996
(Shukla et al., 2012)	India	Asian	HB	ARMS-PCR	208/204	Mixed	178	30	0	Mixed	159	41	4	6	1.000
(Tarigan et al., 2015)	Indonesia	Asian	HB	PCR-RFLP	93/93	Yes	75	8	10	Yes	60	25	8	6	<0.05
(Yang et al., 2014)	China	Asian	HB	PCR-RFLP	101/80	Mixed	73	25	3	Mixed	71	9	0	7	0.878
					74/46	Yes	53	19	2	Yes	42	4	0	7	0.991
(Özdoğan et al., 2014)	Turkey	Caucasian	HB	PCR	60/30	Yes	44	16	0	Yes	24	6	0	6	0.543
(Akparova et al., 2017)	India	Asian	PB	PCR	55/52	Mixed	37	13	5	Mixed	36	14	2	7	0.477
(Reséndiz-Hernández et al., 2018)	Mexico	Caucasian	PB	PCR	384/674	Yes	342	42	0	Yes	641	31	2	8	0.844
(Umi Partan and Hidayat, 2019)	Indonesia	Asian	HB	PCR-RFLP	70/35	Unknown	52	17	1	Unknown	27	8	0	6	0.837
(Mir et al., 2020)	India	Asian	HB	PCR	100/163	Mixed	91	9	0	Mixed	150	13	0	7	0.989
(Yu et al., 2021)	China	Asian	HB	PCR	198/195	Mixed	173	21	4	Unknown	186	9	0	7	0.308
		—							TNF-α 489					—	—
						Smoking status	GG	GA	AA	Smoking status	GG	GA	AA		
(Küçükaycan et al., 2002)	Netherlands	Caucasian	PB	PCR	157/315	Mixed	118	38	1	Unknown	264	45	6	6	0.805
(Hegab et al., 2005)	Japan	Asian	HB	TaqMan	88/61	Yes	57	26	5	Yes	40	19	2	6	0.939
		Caucasian	HB	TaqMan	106/72	Yes	85	19	2	Yes	58	13	1	7	0.876
(Hsieh et al., 2008)	America	Caucasian	HB	PCR-RFLP	30/115	Mixed	23	7	0	Mixed	79	35	1	7	0.514
(Matokanovic et al., 2012)	Croatia	Caucasian	HB	PCR-RFLP	130/95	Mixed	87	38	5	Mixed	53	39	3	6	0.863
Yu et al., 2021)	China	Asian	HB	PCR	198/195	Mixed	140	46	12	Unknown	148	39	8	7	0.060

(Continued on following page)

TABLE 1 | (*Continued*) Main characters of studies included in this meta-analysis.

0.5	Country	Ethnicity	Source of control	Genotype methods	Cases/Controls	Genotype distribution									NOS	HWE
						Case					Control					
						TNF-α 308										
						Smoking status	GG	GA	AA	Smoking status	GG	GA	AA			
	—					TNF-α 238									—	—
						Smoking status	GG	GA	AA	Smoking status	GG	GA	AA			
(Küçükaycan et al., 2002)	Netherlands	Caucasian	PB	PCR	164/331	Mixed	151	13	0	Unknown	310	20	1		6	0.993
(Papatheodorou et al., 2007)	Greece	Caucasian	PB	PCR	116/208	Unknown	111	5	0	Yes	197	11	0		6	1.000
(Trajkov et al., 2009)	France	Caucasian	HB	PCR-SSP	60/301	Mixed	53	7	0	Unknown	276	23	2		7	0.819
(Tarigan et al., 2015)	Indonesia	Asian	HB	PCR-RFLP	93/93	Yes	85	4	4	Yes	89	1	3		6	<0.05
(Reséndiz-Hernández et al., 2018)	Mexico	Caucasian	HB	RT-PCR	384/674	Yes	349	31	4	Yes	621	52	1		8	0.472

Abbreviation: HB, hospital-based; PB, population-based.

monitoring boundary (TSMB). The results are considered reliable and stable when the Z-value crosses TSMB. At the same time, the sample size can be deemed adequate. However, if the cumulative Z value does not cross the TSMB or RIS threshold, it means the sample size is not sufficient. And it still needs more studies to confirm the result.

RESULTS

Features of Recruited Studies

The search strategy yielded 261 potentially relevant articles, and 27 publications with 3473 COPD cases and 4935 controls were ultimately included in the present analysis (**Figure 1**). All the cases were confirmed by the diagnostic criteria of COPD. There were 25 articles that studied the TNF-α-308 variant, five articles on the TNF-α+489 variant, and four articles on the TNF-α-238 variant. These studies fell into two groups: 16 Asian articles and 20 Caucasian articles. The genotype distributions of TNF-α were found to be in the Hardy-Weinberg equilibrium. Smoking status was adjusted for in 10 studies, and these were included for the second meta-analysis (**Table 1**).

Association of COPD Susceptibility With TNF-α-308

The effect of the TNF-α-308G/A polymorphism on COPD was first investigated by pooling 25 studies where smoking status was not adjusted, comprising 3283 COPD cases and 4539 non-COPD controls. It indicated that the A allele was associated with an increased COPD risk in the overall population (A vs. G, OR = 1.21,95% CI: 1.01–1.45, p = 0.04), using an allelic model. According to the comparison between heterozygote model (GA vs. GG, OR = 1.22, 95% CI :1.02–1.45, p = 0.03) and dominant model (GA + AA vs. GG, OR = 1.22, 95%CI: 1.01–1.48, p = 0.04), GA genotype carriers of TNF-α-308 have a higher risk of developing COPD compared to GG carriers.

In the stratified analysis by ethnicity, we found that the TNF-α-308 G/A polymorphism was associated with COPD risk under the allelic model (A vs. G, OR = 1.40, 95% CI: 1.03–1.89, p = 0.03). The pooled OR of heterozygote model was 1.35 (GA vs. GG, OR = 1.35, 95% CI: 1.04–1.77, p = 0.02), indicating that the risk of COPD with TNF-α-308 GA genotype was 1.35 times higher than that with the GG genotype in an Asian population. A similar result was also observed in the dominant model (GA + AA vs. GG, OR = 1.40, 95% CI: 1.04–1.88, p = 0.03), which further supported our finding. In contrast, no significant association between the TNF-α-308 G/A SNP and COPD was found in Caucasians (**Figure 2**).

Specific environmental factors, such as smoking, may contribute to the difference in the distribution of genetic polymorphism. Therefore, we conducted a second meta-analysis with 10 studies including 2749 smokers. In this meta-analysis, TNF-α-308 polymorphism was not associated

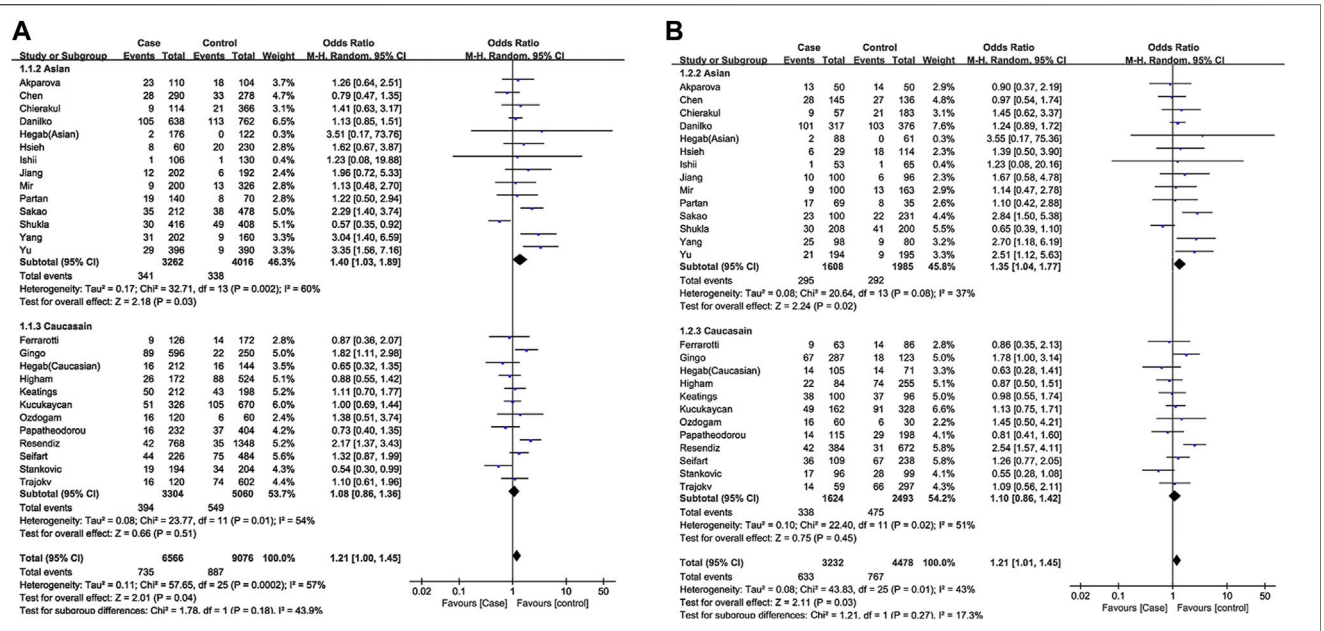


FIGURE 2 | Forest plot of TNF-α-308 polymorphism and COPD risk [**(A)**: overall for A vs. G, **(B)**: overall for GA vs. GG].

with COPD either in Asian smokers or Caucasian smokers (**Table 2**).

Association of COPD Susceptibility With TNF-α+489

For TNF-α+489G/A variant, five research papers, including 709 cases and 853 normal controls, were selected. The pooled results suggested that TNF-α+489 polymorphism was not associated with COPD risk ($p > 0.05$). Subgroups analysis by ethnicity also showed no association between TNF-α+489 and COPD among Asians ($p > 0.05$) and Caucasians ($p > 0.05$) (**Table 3**).

Association of COPD Susceptibility With TNF-α-238

624 cases and 1514 controls originated from four studies were included to probe the relevance between susceptibility to COPD and TNF-α-238G/A variant. We found that TNF-α-238 polymorphism was not associated with COPD risk in the overall population ($p > 0.05$). (**Table 3**).

Sensitivity Analyses

Sensitivity analysis was performed by sequentially excluding each study to assess the stability of the results in this meta-analysis. As shown in **Table 4**, the results based on the overall meta-analysis were not stable (**Table 4**). However, in the Asian subgroup, the pooled OR did not vary significantly, implying that the findings were reliable (**Figure 3**). We speculated that the overall stability was affected by those studies restricted to Caucasians.

Publication Bias

Begg's funnel plots and Egger's test were performed to assess the publication bias (all contrast models: $p > 0.05$), and the results suggested that there was no publication bias for the association between TNF-α-308 polymorphism and COPD in included studies (**Figure 4**).

Trial Sequential Analysis Results

We implemented TSA to reduce the risk of type I error and to evaluate the RIS by keeping the overall 5% risk of the type I error and the relative risk reduction of 20% (power of 80%). The results on TNF-α-308 showed that the Z-curve crossed the TSMB (**Figure 5A**). Therefore, it can be assumed that mutation from G to A on TNF-α-308 can increase the risk of COPD. The sample size did not reach the RIS among Asians in the allele model, heterozygote model, or dominant model (**Figure 5B**). Hence, more qualified studies are expected to confirm the relationship between TNF-α-308 variants and COPD susceptibility among Asians.

DISCUSSION

For the TNF-α-308 G/A, the GA genotype was significantly associated with increased COPD risk, especially among Asians. Presumably, the GA variant was a stronger activator of TNF-α transcription than the GG genotype (Wilson et al., 1997). Some studies had shown that TNF-α-308 GA expression may cause damage to muscle (Baumert et al., 2016) and increase bronchial responsiveness (Moffatt et al., 1999). Moreover, Zhang (Zhang et al., 2016) and other previous studies (Hu et al., 2007; Zhan et al., 2011) found an increased risk of COPD associated with TNF-α-308 AA genotype. Nevertheless, we did not find any

TABLE 2 | Meta-analysis results for relationship between the TNF-α-308 polymorphism and COPD risk.; Abbreviation: CI = confidence interval, F = fixed effect model, OR = odds ratio, R = random-effect model.

TNF-α-308		Contrast model	Number of studies	OR (95%CI)	*p*	Test for heterogeneity		Analysis model
						I^2 (%)	*p*	
Participants included	Overall	A vs. G	25	1.21 (1.01, 1.45)	0.04	57	0.00	R
		GA vs. GG	25	1.22 (1.02, 1.45)	0.03	43	0.01	R
		AA vs. GG	25	1.18 (0.80, 1.73)	0.41	0	0.45	F
		AA vs. GG + GA	25	1.14 (0.77, 1.67)	0.52	0	0.55	F
		GA + AA vs. GG	25	1.22 (1.01, 1.48)	0.04	52	0.00	R
	Asian	A vs. G	14	1.40 (1.03, 1.89)	0.03	60	0.00	R
		GA vs. GG	14	1.35 (1.04, 1.77)	0.02	37	0.08	R
		AA vs. GG	14	1.33 (0.76, 2.33)	0.32	22	0.24	F
		AA vs. GG + GA	14	1.26 (0.72, 2.21)	0.42	17	0.29	F
		GA + AA vs. GG	14	1.40 (1.04, 1.88)	0.03	51	0.01	R
	Caucasian	A vs. G	12	1.08 (0.86, 1.36)	0.51	54	0.01	R
		GA vs. GG	12	1.10 (0.86, 1.42)	0.45	51	0.02	R
		AA vs. GG	12	1.05 (0.61, 1.79)	0.87	0	0.62	F
		AA vs. GG + GA	12	1.03 (0.60, 1.77)	0.91	0	0.69	F
		GA + AA vs. GG	12	1.09 (0.85, 1.41)	0.49	53	0.01	R
Smokers	Overall	A vs. G	9	1.35 (0.97, 1.89)	0.08	52	0.03	R
		GA vs. GG	9	1.41 (0.99, 2.00)	0.06	47	0.06	R
		AA vs. GG	9	1.31 (0.62, 2.75)	0.48	0	0.45	F
		AA vs. GG + GA	9	1.24 (0.59, 2.60)	0.57	0	0.49	F
		GA + AA vs. GG	9	1.40 (0.98, 2.00)	0.06	50	0.04	R
	Asian	A vs. G	5	1.19 (0.80, 1.78)	0.39	47	0.11	F
		GA vs. GG	5	1.32 (0.85, 2.05)	0.21	17	0.30	F
		AA vs. GG	5	0.71 (0.03, 19.93)	0.84	59	0.12	R
		AA vs. GG + GA	5	0.64 (0.03, 14.48)	0.78	53	0.14	R
		GA + AA vs. GG	5	1.27 (0.83, 1.95)	0.27	37	0.17	F
	Caucasian	A vs. G	5	1.38 (0.92, 2.07)	0.12	59	0.05	R
		GA vs. GG	5	1.38 (0.84, 2.27)	0.21	65	0.02	R
		AA vs. GG	5	1.63 (0.68, 3.88)	0.27	0	0.62	F
		AA vs. GG + GA	5	1.56 (0.65, 3.71)	0.32	0	0.65	F
		GA + AA vs. GG	5	1.40 (0.88, 2.23)	0.16	62	0.03	R

TABLE 3 | Meta-analysis results for relationship between the TNF-α+489, -238 polymorphism and COPD risk.

SNPs	Contrast model	Subgroup	Number of studies	OR (95% CI)	*p*	Test for heterogeneity		Analysis model
						I^2 (%)	*p*	
TNF-α+489	A vs. G	Overall	5	1.10 (0.89, 1.35)	0.38	41	0.15	F
	GA vs. GG	Overall	5	1.04 (0.68, 1.57)	0.87	63	0.03	R
	AA vs. GG	Overall	5	1.23 (0.66, 2.26)	0.51	0	0.79	F
	AA vs. GG + GA	Overall	5	1.22 (0.67, 2.25)	0.51	0	0.77	F
	GA + AA vs. GG	Overall	5	1.05 (0.73, 1.53)	0.79	57	0.05	R
	A vs. G	Asian	2	1.24 (0.90, 1.71)	0.18	0	0.64	F
	GA vs. GG	Asian	2	1.15 (0.77, 1.72)	0.50	0	0.55	F
	AA vs. GG	Asian	2	1.62 (0.72, 3.65)	0.24	0	0.92	F
	AA vs. GG + GA	Asian	2	1.57 (0.70, 3.50)	0.27	0	0.87	F
	GA + AA vs. GG	Asian	2	1.22 (0.84, 1.77)	0.30	0	0.58	F
	A vs. G	Caucasian	4	1.01 (0.77, 1.32)	0.97	47	0.13	F
	GA vs. GG	Caucasian	4	0.97 (0.52, 1.80)	0.92	71	0.02	R
	AA vs. GG	Caucasian	4	0.81 (0.30, 2.15)	0.67	0	0.84	F
	AA vs. GG + GA	Caucasian	4	0.86 (0.33, 2.23)	0.75	0	0.75	F
	GA + AA vs. GG	Caucasian	4	0.96 (0.56, 1.66)	0.89	65	0.04	R
TNF-α-238	A vs. G	Caucasian	4	1.22 (0.89, 1.67)	0.23	0	0.89	F
	GA vs. GG	Caucasian	4	1.14 (0.81, 1.61)	0.44	0	0.76	F
	AA vs. GG	Caucasian	4	2.64 (0.70, 9.93)	0.15	0	0.40	F
	AA vs. GG + GA	Caucasian	4	2.59 (0.69, 9.69)	0.16	0	0.39	F
	GA + AA vs. GG	Caucasian	4	1.18 (0.85, 1.65)	0.32	0	0.87	F

Abbreviation: CI, confidence interval, F = fixed effect model, OR, odds ratio, R = random-effect model.

TABLE 4 | Sensitivity analysis of the association between TNF-α-308 and COPD risk, under the allelic model in overall population.; Abbreviation: CI = confidence interval, OR = odds ratio.

Author	OR (95%CI)	Z value	*p*	I^2 (%)
Akparova	1.21 (1.00, 1.46)	1.94	0.05	59
Chen	1.23 (1.02, 1.49)	2.19	0.03	57
Chierakul	1.20 (0.99, 1.45)	1.92	0.06	59
Danilko	1.22 (1.00, 1.48)	1.92	0.05	59
Ferrarotti	1.22 (1.01, 1.47)	2.08	0.04	59
Gingo	1.18 (0.98, 1.43)	1.75	0.08	57
Hegab	1.23 (1.02, 1.48)	2.17	0.03	58
Higham	1.23 (1.02, 1.49)	2.12	0.03	58
Hsieh	1.20 (0.99, 1.45)	1.89	0.06	59
Ishii	1.20 (1.00, 1.45)	2.00	0.05	59
Keatings	1.21 (1.00, 1.47)	1.97	0.05	59
Jiang	1.20 (0.99, 1.44)	1.87	0.06	58
Kucukaycan	1.22 (1.00, 1.49)	2.04	0.04	58
Mir	1.21 (1.00, 1.46)	1.99	0.05	59
Partan	1.21 (1.00, 1.46)	1.97	0.05	59
Resendiz	1.17 (0.97, 1.40)	1.67	0.09	53
Sakao	1.16 (0.97, 1.39)	1.67	0.10	53
Seifart	1.20 (0.99, 1.46)	1.87	0.06	59
Shukla	1.25 (1.05, 1.50)	2.53	0.01	51
Stankovic	1.25 (1.04, 1.49)	2.43	0.02	54
Trajokv	1.21 (1.00, 1.47)	1.99	0.05	59
Ozdogan	1.20 (1.00, 1.45)	1.94	0.05	59
Papatheodorou	1.23 (1.02, 1.49)	2.21	0.03	57
Yang	1.17 (0.97, 1.40)	1.72	0.09	54
Yu	1.16 (0.98, 1.39)	1.69	0.09	53

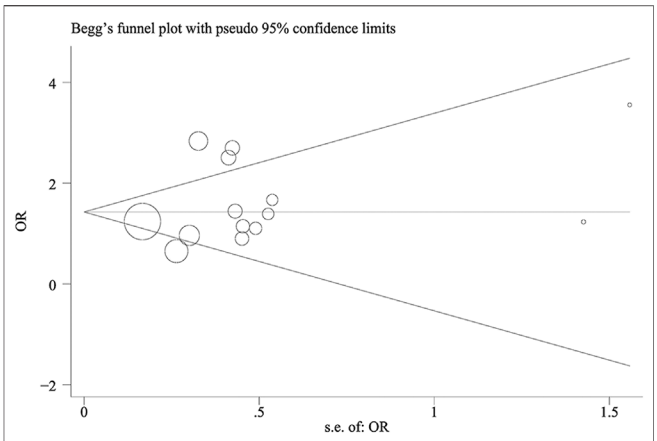


FIGURE 4 | Begg's funnel plot of TNF-α-308 polymorphism and COPD risk (Asian population for GA vs. GG).

significant relationship between COPD susceptibility and TNF-α-308 AA genotype. Our meta-analysis updated five case-control studies (Akparova et al., 2017; Reséndiz-Hernández et al., 2018; Umi Partan and Hidayat, 2019; Mir et al., 2020; Yu et al., 2021) compared with Zhang's research, and the cumulative Z-value crossed the TSA boundary. The TSA result could be due to the fact that our study had enough COPD cases and controls. This may be the reason for the difference between our analysis and Zhang's study. Furthermore, we conducted the ethnicity subgroup analysis to gain a more comprehensive view (Salimi Asl et al., 2019; Shi and Zhao, 2019). In our study, a second meta-analysis restricted to smokers was also performed, and it indicated that TNF-α-308G/A polymorphism was not associated with COPD susceptibility among smokers. The result implied that cigarette smoking may conceal the influence of TNF-α polymorphism on COPD. The effect of smoking on the gene factor may be attributed to the following: 1) smoking might exert an interaction with TNF-α-308G/A polymorphism, and have a potential influence on TNF-α gene expression; and 2) smoking is the major risk factor for COPD. We speculated that compared with the effect of TNF-α-308 GA

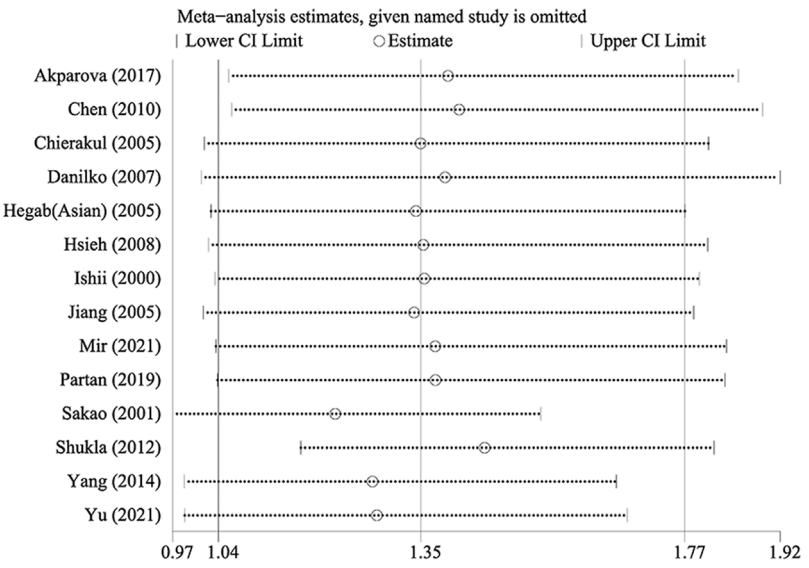


FIGURE 3 | Sensitivity analysis of TNF-α-308 polymorphism and COPD risk (Asian population for GA vs. GG).

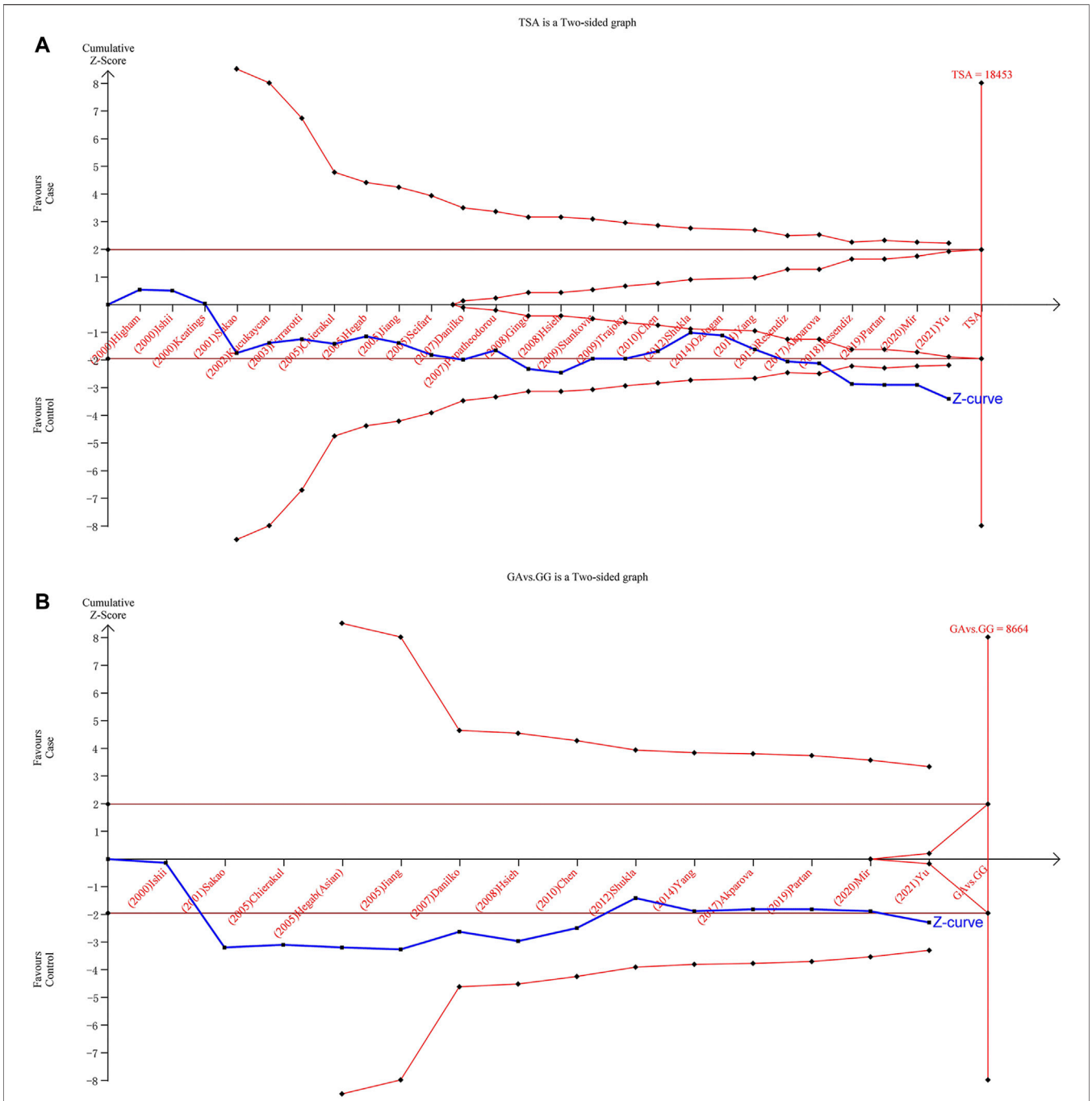


FIGURE 5 | Results of TSA with TNF-α-308 variant. The required information size was calculated based on a two-sided a = 5%, β = 15% (power 80%), and a relative risk reduction of 20%. [**(A)**: overall for A vs. G, **(B)**: Asian population for GA vs. GG].

variant on COPD, there was a stronger association between smoking and COPD risk.

As for the TNF-α+489G/A, no significant association was found between TNF-α+489G/A polymorphism and COPD susceptibility. Due to the new case-control study (Yu et al., 2021) included, we proposed a converse conclusion compared with previous meta-analysis (Cui et al., 2015). The result remained controversial because of the limited sample size. Therefore, more case-control studies and cohort studies are in urgent need to determine a stable conclusion.

To the best of our knowledge, this is the first meta-analysis to analyze TNF-α -238G/A variant and its contribution to risk of COPD. Similarly, the pooled results suggested that TNF-α -238G/A polymorphism had no significant association with COPD risk. All the case-control studies about TNF-α-238G/A we retrieved were restricted to Caucasians. Hence, we are looking forward to

more original studies on other ethnicities to confirm our conclusions.

Our study features some advantages. Firstly, we performed TSA to reduce the risk of type I error and evaluate required information size. The TSA result showed that the conclusion about TNF-α-308 G/A was based on a sufficient number of cases and controls. Secondly, we carried out a second meta-analysis restricted to smokers. As smoking is the major risk factor for the development of COPD, this sub-group analysis illustrates the additional information genetic polymorphism provides when studying risk factors of COPD. Thirdly, NOS quality test was performed to estimate the quality of the case-control studies included, making the results more reliable. In addition, sensitivity analysis and Egger's test were performed to assess the publication bias. And the test showed our results were stable in the Asian subgroup.

Nevertheless, there are still some limitations that cannot be avoided in our meta-analysis. First, due to the lack of case-control studies related to nonsmokers, we failed to ascertain the association between the TNF-α-308 G/A variant and COPD risk in nonsmoking populations. Second, some unpublished studies might not be retrieved. Third, there were limited data to conduct subgroup analyses in Asians and Caucasians. The conclusions for TNF-α+489G/A and TNF-α -238G/A required more original studies to make them convincing.

Together, the present meta-analysis indicated that TNF-α-308G/A polymorphism was associated with an increased risk of COPD among Asians but not in Caucasians, and GA genotype carriers of TNF-α-308 had a higher risk of developing COPD compared to GG carriers when smoking status was not adjusted. However, in the meta-analysis with restrictions to smokers, no association was found between TNF-α-308 polymorphism and COPD susceptibility either in Asian smokers or Caucasian smokers. Moreover, the TNF-α+489G/A, -238G/A variants did not have an increased risk of COPD susceptibility.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Study design: ZX and YW; Data collection: FL and HS; Data analysis: ZX, YW, and PH; Writing: ZX and YW.

REFERENCES

Akparova, A., Abdrakhmanova, B., Banerjee, N., and Bersimbaev, R. (2017). EPHX1 Y113H Polymorphism Is Associated with Increased Risk of Chronic Obstructive Pulmonary Disease in Kazakhstan Population. *Mutat. Research/Genetic Toxicol. Environ. Mutagenesis* 816-817, 1–6. doi:10.1016/j.mrgentox.2017.02.004

Baumert, P., Lake, M. J., Stewart, C. E., Drust, B., and Erskine, R. M. (2016). Genetic Variation and Exercise-Induced Muscle Damage: Implications for Athletic Performance, Injury and Ageing. *Eur. J. Appl. Physiol.* 116 (9), 1595–1625. doi:10.1007/s00421-016-3411-1

Celli, B. R., Decramer, M., Wedzicha, J. A., Wilson, K. C., Agustí, A., Criner, G. J., et al. (2015). An Official American Thoracic Society/European Respiratory Society Statement: Research Questions in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* 191 (7), e4–e27. doi:10.1164/rccm.201501-0044ST

Chen, Y.-C., Liu, S.-F., Chin, C.-H., Wu, C.-C., Chen, C.-J., Chang, H.-W., et al. (2010). Association of Tumor Necrosis Factor-α-863c/A Gene Polymorphism with Chronic Obstructive Pulmonary Disease. *Lung* 188 (4), 339–347. doi:10.1007/s00408-010-9236-5

Chierakul, N., Wongwisutikul, P., Vejbaesya, S., and Chotvilaiwan, K. (2005). Tumor Necrosis Factor-Alpha Gene Promoter Polymorphism Is Not Associated with Smoking-Related COPD in Thailand. *Respirology* 10 (1), 36–39. doi:10.1111/j.1440-1843.2005.00626.x

Cui, K., Ge, X. Y., and Ma, H. L. (2015). Association of the TNF-A+489 G/A Polymorphism with Chronic Obstructive Pulmonary Disease Risk in Asians: Meta-Analysis. *Genet. Mol. Res.* 14 (2), 5210–5220. doi:10.4238/2015.may.18.12

Danilko, K. V., Korytyna, G. F., Akhmadishina, L. Z., Yanbaeva, D. G., Zagidullin, S. Z., and Victorova, T. V. (2007). Association of Polymorphisms of Cytokine Genes (IL1B, IL1RN, TNFA, LTA, IL6, IL8, and IL10) with Chronic Obstructive Pulmonary Disease. *Mol. Biol.* 41 (1), 22–31. doi:10.1134/s0026893307010049

Ferrarotti, I., Zorzetto, M., Beccaria, M., Gilè, L. S., Porta, R., Ambrosino, N., et al. (2003). Tumour Necrosis Factor Family Genes in a Phenotype of COPD Associated with Emphysema. *Eur. Respir. J.* 21 (3), 444–449. doi:10.1183/09031936.03.00051303

Foreman, K. J., Marquez, N., Dolgert, A., Fukutaki, K., Fullman, N., McGaughey, M., et al. (2018). Forecasting Life Expectancy, Years of Life Lost, and All-Cause and Cause-specific Mortality for 250 Causes of Death: Reference and Alternative Scenarios for 2016-40 for 195 Countries and Territories. *The Lancet* 392 (10159), 2052–2090. doi:10.1016/s0140-6736(18)31694-5

Gingo, M. R., Silveira, L. J., Miller, Y. E., Friedlander, A. L., Cosgrove, G. P., Chan, E. D., et al. (2008). Tumour Necrosis Factor Gene Polymorphisms Are Associated with COPD. *Eur. Respir. J.* 31 (5), 1005–1012. doi:10.1183/09031936.00100307

Hegab, A. E., Sakamoto, T., Saitoh, W., Nomura, A., Ishii, Y., Morishima, Y., et al. (2005). Polymorphisms of TNFα, IL1β, and IL1RN Genes in Chronic Obstructive Pulmonary Disease. *Biochem. Biophysical Res. Commun.* 329 (4), 1246–1252. doi:10.1016/j.bbrc.2005.02.099

Higham, M. A., Pride, N. B., Alikhan, A., and Morrell, N. W. (2000). Tumour Necrosis Factor-α Gene Promoter Polymorphism in Chronic Obstructive Pulmonary Disease. *Eur. Respir. J.* 15 (2), 281–284. doi:10.1034/j.1399-3003.2000.15b10.x

Hsieh, M.-H., Chong, I.-W., Hwang, J.-J., Lee, C.-H., Ho, C.-K., Yu, M.-L., et al. (2008). Lack of Associations between Several Polymorphisms in Cytokine Genes and the Risk of Chronic Obstructive Pulmonary Diseases in Taiwan. *Kaohsiung J. Med. Sci.* 24 (3), 126–137. doi:10.1016/s1607-551x(08)70140-2

Hu, G. P., Peng, G. Y., Hu, J. X., and Ran, P. X. (2007). Association of Tumor Necrosis Factor Alpha 308 G/A Gene Promoter Polymorphism with the Presence of Chronic Obstructive Pulmonary Disease: a Meta-Analysis. *Zhonghua Jie He He Hu Xi Za Zhi* 30 (8), 588–594.

Ishii, T., Matsuse, T., Teramoto, S., Matsui, H., Miyao, M., Hosoi, T., et al. (2000). Neither IL-1 β, IL-1 Receptor Antagonist, Nor TNF- α Polymorphisms Are Associated with Susceptibility to COPD. *Respir. Med.* 94 (9), 847–851. doi:10.1053/rmed.2000.0808

Jiang, L., He, B., Zhao, M. W., Ning, L. D., Li, X. Y., and Yao, W. Z. (2005). Association of Gene Polymorphisms of Tumour Necrosis Factor-Alpha and Interleukin-13 with Chronic Obstructive Pulmonary Disease in Han Nationality in Beijing. *Chin. Med. J. (Engl)* 118 (7), 541–547.

Keatings, V. M., Cave, S. J., Henry, M. J., Morgan, K., O'Connor, C. M., FitzGerald, M. X., et al. (2000). A Polymorphism in the Tumor Necrosis Factor-α Gene Promoter Region May Predispose to a Poor Prognosis in COPD. *Chest* 118 (4), 971–975. doi:10.1378/chest.118.4.971

Küçükaycan, M., Van Krugten, M., Pennings, H. J., Huizinga, T. W., Buurman, W. A., Dentener, M. A., et al. (2002). Tumor Necrosis Factor-Alpha +489G/A Gene Polymorphism Is Associated with Chronic Obstructive Pulmonary Disease. *Respir. Res.* 3, 29. doi:10.1186/rr194

Matokanovic, M., Rumora, L., Popovic-Grle, S., Cepelak, I., Culic, O., and Barisic, K. (2012). Association of Hsp70-2 (+1267A/G), Hsp70-Hom (+2437T/C), HMOX-1 (Number of GT Repeats) and TNF-Alpha (+489G/A) Polymorphisms with COPD in Croatian Population. *Clin. Biochem.* 45 (10-11), 770–774.

Mir, H., Koul, P. A., Bhat, D., and Shah, Z. A. (2020). A Case-Control Study of Tumor Necrosis Factor-Alpha Promoter Polymorphism and its Serum Levels in Patients with Chronic Obstructive Pulmonary Disease in Kashmir, North India. *Lung India* 37 (3), 204–209. doi:10.4103/lungindia.lungindia_477_19

Moffatt, M. F., James, A., Ryan, G., Musk, A. W., and Cookson, W. O. C. M. (1999). Extended Tumour Necrosis Factor/HLA-DR Haplotypes and Asthma in an Australian Population Sample. *Thorax* 54 (9), 757–761. doi:10.1136/thx.54.9.757

Nedwin, G. E., Naylor, S. L., Sakaguchi, A. Y., Smith, D., Jarrett-Nedwin, J., Pennica, D., et al. (1985). Human Lymphotoxin and Tumor Necrosis Factor Genes: Structure, Homology and Chromosomal Localization. *Nucl. Acids Res.* 13 (17), 6361–6373. doi:10.1093/nar/13.17.6361

Özdoğan, N., Tutar, N., Demir, R., Saatçi, Ç., Kanbay, A., and Büyükoğlan, H. (2014). Is TNF-α Gene Polymorphism Related to Pulmonary Functions and Prognosis as Determined by FEV1, BMI, COPD Exacerbation and Hospitalization in Patients with Smoking-Related COPD in a Turkish Population? *Revista Portuguesa de Pneumologia* 20 (6), 305–310. doi:10.1016/j.rppneu.2014.03.003

Papatheodorou, A., Latsi, P., Vrettou, C., Dimakou, A., Chroneou, A., Makrythanasis, P., et al. (2007). Development of a Novel Microarray Methodology for the Study of SNPs in the Promoter Region of the TNF-α Gene-Their Association with Obstructive Pulmonary Disease in Greek Patients. *Clin. Biochem.* 40 (12), 843–850. doi:10.1016/j.clinbiochem.2007.03.024

Parameswaran, N., and Patial, S. (2010). Tumor Necrosis Factor-α Signaling in Macrophages. *Crit. Rev. Eukar Gene Expr.* 20 (2), 87–103. doi:10.1615/critreveukargeneexpr.v20.i2.10

Reséndiz-Hernández, J. M., Ambrocio-Ortiz, E., Pérez-Rubio, G., Lopez-Flores, L. A., Abarca-Rojano, E., Pavón Romero, G. F., et al. (2018). TNF Promoter Polymorphisms Are Associated with Genetic Susceptibility in COPD Secondary to Tobacco Smoking and Biomass Burning. *Copd* Vol. 13, 627–637. doi:10.2147/copd.s147688

Sakao, S., Tatsumi, K., Igari, H., Shino, Y., Shirasawa, H., and Kuriyama, T. (2001). Association of Tumor Necrosis Factor α Gene Promoter Polymorphism with the Presence of Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* 163 (2), 420–422. doi:10.1164/ajrccm.163.2.2006031

Salimi Asl, M., Ahmadi, A., Salimian, J., Shohani, S., Azimzadeh Jamalkandi, S., and Ghanei, M. (2019). TNF-α −308 G/A Variant and Susceptibility to Chronic Obstructive Pulmonary Disease: A Systematic Review and Meta-Analysis. *Cytokine* 123, 154763. doi:10.1016/j.cyto.2019.154763

Seifart, C., Dempfle, A., Plagens, A., Seifart, U., Clostermann, U., Müller, B., et al. (2005). TNF-α-, TNF-β-, IL-6-, and IL-10-promoter Polymorphisms in Patients with Chronic Obstructive Pulmonary Disease. *Tissue Antigens* 65 (1), 93–100. doi:10.1111/j.1399-0039.2005.00343.x

Shi, C., and Zhao, H. (2019). Association between Tumor Necrosis Factor-308 G/A Polymorphism and Chronic Obstructive Pulmonary Disease Risk in Chinese Population: Evidence from a Meta-Analysis. *Clin. Lab.* 65 (10). doi:10.7754/Clin.Lab.2019.190313

Shukla, R. K., Kant, S., Bhattacharya, S., and Mittal, B. (2012). Association of Cytokine Gene Polymorphisms in Patients with Chronic Obstructive Pulmonary Disease. *Oman Med. J.* 27 (4), 285–290. doi:10.5001/omj.2012.71

Stankovic, M., Nestorovic, A., Tomovic, A., Petrovic-Stanojevic, N., Andjelic-Jelic, M., Dopudja-Pantic, V., et al. (2009). TNF-α-308 Promotor Polymorphism in Patients with Chronic Obstructive Pulmonary Disease and Lung Cancer. *neo* 56 (4), 348–352. doi:10.4149/neo_2009_04_348

Tarigan, A. P., Syafiuddin, T., Yunus, F., and Suradi (2015). Association of Tumor Necrosis Factor Alpha and Lymphotoxin Alpha Gene Polymorphisms with the Presence of Chronic Obstructive Pulmonary Disease. *Acta Med. Indones* 47 (4), 283–290.

Trajkov, D., Mirkovska-Stojkovikj, J., Petlichkovski, A., Strezova, A., Efinska-Mladenovska, O., Sandevska, E., et al. (2009). Association of Cytokine Gene Polymorphisms with Chronic Obstructive Pulmonary Disease in Macedonians. *Iran J. Allergy Asthma Immunol.* 8 (1), 31–42. doi:10.08.01/ijaai.3142

Umi Partan, R., and Hidayat, R. (2019). The Relationship between TNF-α Gene Polymorphism, Pro-inflammatory Cytokines and Bone Turnover Markers in COPD Patients with Osteoporosis. *J. Phys. Conf. Ser.*, 1246. doi:10.1088/1742-6596/1246/1/012035

Wilson, A. G., Symons, J. A., McDowell, T. L., McDevitt, H. O., and Duff, G. W. (1997). Effects of a Polymorphism in the Human Tumor Necrosis Factor Promoter on Transcriptional Activation. *Proc. Natl. Acad. Sci.* 94 (7), 3195–3199. doi:10.1073/pnas.94.7.3195

Yang, L., Li, F., Yan, M., and Su, X. (2014). Association of the CYP1A1 MspI and TNFα-308 Polymorphisms with Chronic Obstructive Pulmonary Disease in Inner Mongolia. *Genet. Mol. Res.* 13 (2), 3209–3217. doi:10.4238/2014.april.25.6

Yu, S., Xue, M., Yan, Z., Song, B., Hong, H., and Gao, X. (2021). Correlation between TNF-α -308 and +489 Gene Polymorphism and Acute Exacerbation of Chronic Obstructive Pulmonary Diseases. *Biomed. Res. Int.* 2021, 6661281. doi:10.1155/2021/6661281

Zhan, P., Wang, J., Wei, S.-Z., Qian, Q., Qiu, L.-X., Yu, L.-K., et al. (2011). TNF-308 Gene Polymorphism Is Associated with COPD Risk Among Asians: Meta-Analysis of Data for 6,118 Subjects. *Mol. Biol. Rep.* 38 (1), 219–227. doi:10.1007/s11033-010-0098-y

Zhang, L., Gu, H., Gu, Y., and Zeng, X. (2016). Association between TNF-α -308 G/A Polymorphism and COPD Susceptibility: a Meta-Analysis Update. *Int. J. Chron. Obstruct Pulmon Dis.* 11, 1367–1379. doi:10.2147/COPD.S105394

Zhou, J., Liu, B., Liang, C., Li, Y., and Song, Y.-H. (2016). Cytokine Signaling in Skeletal Muscle Wasting. *Trends Endocrinol. Metab.* 27 (5), 335–347. doi:10.1016/j.tem.2016.03.002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Xia, Wang, Liu, Shu and Huang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

CASE REPORT
published: 24 January 2022
doi: 10.3389/fgene.2022.816987

Case Report: Mutation in AIMP2/P38, the Scaffold for the Multi-Trna Synthetase Complex, and Association With Progressive Neurodevelopmental Disorders

Mahta Mazaheri[1,2,3†], Mahdie Yavari[3,4†], Hadi Zare Marzouni[5], Angela Stufano[6,7], Piero Lovreglio[7], Simona D'Amore[8] and Hamid Reza Jahantigh[6,7]*

[1]Department of Medical Genetics, School of Medicine, Shahid Sadoughi University of Medical Sciences, Yazd, Iran, [2]Mother and Newborn Health Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran, [3]Dr. Mazaheri's Medical Genetics Lab, Yazd, Iran, [4]Division of Genetics, Department of Cell and Molecular Biology and Microbiology, Faculty of Science and Biotechnology, University of Isfahan, Isfahan, Iran, [5]Qaen School of Nursing and Midwifery, Birjand University of Medical Sciences, Birjand, Iran, [6]Department of Veterinary Medicine, University of Bari Aldo Moro, Bari, Italy, [7]Interdisciplinary Department of Medicine - Section of Occupational Medicine, University of Bari, Bari, Italy, [8]Department of Medicine, University of Cambridge, Cambridge, United Kingdom

OPEN ACCESS

Edited by:
William C. Cho,
QEH, Hong Kong SAR, China

Reviewed by:
Henna Tyynismaa,
University of Helsinki, Finland
Oscar Vargas-Rodriguez,
Yale University, United States
Meng Wang,
Zhejiang University, China

***Correspondence:**
Angela Stufano
angela.stufano@uniba.it

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: 17 November 2021
Accepted: 04 January 2022
Published: 24 January 2022

Citation:
Mazaheri M, Yavari M, Zare Marzouni H, Stufano A, Lovreglio P, D'Amore S and Jahantigh HR (2022) Case Report: Mutation in AIMP2/P38, the Scaffold for the Multi-Trna Synthetase Complex, and Association With Progressive Neurodevelopmental Disorders. Front. Genet. 13:816987. doi: 10.3389/fgene.2022.816987

Background: Leukodystrophies constitute a heterogeneous group of inherited disorders primarily affecting the white matter of the central nervous system. Aminoacyl-tRNA synthetases (ARSs) catalyze the attachment of an amino acids to their cognate transfer RNAs (tRNAs). Pathogenic variants in both cytosolic and mitochondrial ARSs have been linked to a broad range of neurological disorders, including hypomyelinating leukodystrophies and pontocerebellar hypoplasias (PCH). Aminoacyl tRNA synthetase-interacting multifunctional protein 2 (AIMP2), one of the three non-catalytic components of multi ARS complex, harbors anti-proliferative activity and functions as a proapoptotic factor thus promoting cell death. We report a case of a 7-month-old infant with a complex clinical presentation, including weight loss, severe anemia, skeletal abnormalities, microcephaly and MR imaging features of leukodystrophy with a novel mutation in AIMP2.

Methods: Whole-exome sequencing (WES) was performed on the proband. Parental samples were analyzed by PCR amplification and Sanger sequencing.

Results: Whole-exome sequencing revealed a novel variant c.A463T in the homozygous state in exon 3 (NM_001,326,607) of AIMP2 [p.(K155X)] in the proband. Parental carrier status was confirmed by target sequencing.

Conclusion: Here, we present an Iranian case with leukodystrophy with a novel AIMP2 mutation. This finding broadens the mutational and phenotypic spectra of AIMP2-related

Abbreviations: ARSs, Aminoacyl-tRNA synthetases; PCH, pontocerebellar hypoplasias; AIMP2, Aminoacyl tRNA synthetase-interacting multifunctional protein 2; WES, Whole-exome sequencing; GLIA, Global Leukodystrophy Initiative; MRI, magnetic resonance imaging; WGS, whole genome sequencingl; BWA, Burrows-Wheeler Aligner; VCF, variant call format; CADD, Combined Annotation Dependent Depletion; TNF-a, tumor necrosis factor-a.

leukodystrophy and offers guidance for proper genetic counselling for pre- and post-natal screenings as well as for disease management.

Keywords: leukodystrophies, WES, AIMP2/P38, neurodevelopmental disorders, multi-tRNA synthetase complex

INTRODUCTION

Leukodystrophies are heritable heterogeneous multisystem conditions that primarily affect the white matter of the central nervous system with or without peripheral nervous system involvement (Van Der Knaap et al., 1999). The main neuropathological sites in leukodystrophies are represented by the myelin sheath and myelin-generating cells; however, in some disorders, damage is suspected to originate at the axonal level (Ashrafi and Tavasoli, 2017). The Global Leukodystrophy Initiative (GLIA) Consortium has distinguished on the basis of magnetic resonance imaging (MRI) characteristics two types of classic leukodystrophy: hypomyelinating leukodystrophies, with mild hyperintensity of cerebral white matter found in the T2W sequence of MRI, and demyelinating leukodystrophies, characterized by hyperintensity in T2W and relevant hypointensity in T1W sequences (Parikh et al., 2015; Ashrafi et al., 2020). Moreover, categorization of leukodystrophies based on cellular pathology and metabolic and molecular approaches were proposed in recent years (Van Der Knaap et al., 1999). Currently, the incidence of heritable white matter disorders in pediatric subjects is estimated to be between 1.2/100,000 and 1/6–7,700 live births (Bonkowsky et al., 2010; Numata et al., 2014). Previous studies on leukodystrophies have observed that this condition is often progressive, with nonspecific manifestations and a similar clinical scenario found in most individuals (Vanderver, 2016). Each type of leukodystrophy affects a different part of the myelin sheath and is associated with several different neurological problems. The most common clinical manifestation is the progressive deterioration or regression of the neurological function, with motor deficit due to myelin destruction as most common neurological sign (Vanderver et al., 2015). Typically, hypotonia is more common in the early stages of the disease, especially in hypomyelinating leukodystrophies, whereas a combination of truncal hypotonia and appendicular spasticity is more frequent in the later stages. Signs of involvement of the corticospinal tract (central hypotonia, spasticity), basal ganglia (various types of movement disorders), peripheral nerves (sensory ataxia, abnormal gait), and cerebellar signs (ataxia, nystagmus) are additional important neurological features found in affected individuals (Van Der Knaap et al., 1999). In addition to the neurologic findings, a variety of extra-neural features can be helpful in orientating toward a specific diagnosis. Endocrine disturbances, ophthalmologic, cutaneous, skeletal radiographic abnormalities, dysmorphic facial features, and gastrointestinal symptoms may be detected in patients with leukodystrophy (Parikh et al., 2015).

Although assessment of cerebral white matter involvement by standard brain magnetic resonance imaging (MRI) is the diagnostic tool of choice for leukodystrophy (Van Der Knaap et al., 1999), genetic testing has also taken on a key role in the diagnostic processes of heritable childhood white matter disorders in recent years. Whole-exome sequencing (WES) and whole-genome sequencing (WGS) have been widely used for a better understanding of cases of unknown etiology in several fields of pediatric neurology, with important advantages in terms of predicting possible complications and/or symptoms that may arise during the clinical course of a disease, to identify unexpected clinical presentations associated with genes whose alterations are already known to be pathogenic, to broaden the clinical presentation of already known disorders, to determine prognosis, and finally to identify new genes whose mutations cause disease. (Srivastava et al., 2014).

Aminoacyl-tRNA synthetases (ARSs) catalyze the attachment of an amino acid to its corresponding tRNA, ensuring the translation of genetic information into functional proteins. Pathogenic variants in both cytosolic and mitochondrial ARSs are associated with a wide range of neurological disorders, including hypomyelinating leukodystrophies and pontocerebellar hypoplasia (PCH) (Ashrafi and Tavasoli, 2017). The multifunctional aminoacyl tRNA synthetase-interacting protein 2 (AIMP2, also known as p38), is one of three noncatalytic components (AIMP1, 2, and 3) that form the mammalian multi-tRNA synthetase complex in combination with nine aminoacyl tRNA synthetases. AIMP2 is also involved in other activities besides the multi-tRNA synthetase complex and may determine cell fate through anti-proliferative and pro-apoptotic activities. Specifically, AIMP2 can promote cell death through several modalities. For example, in response to DNA damage it may exert pro-apoptotic activity by modulating p53 activity (Park et al., 2010). In addition, AIMP2 may induce cell death by mediating apoptotic TNF signaling through ubiquitin-mediated destruction of TRAF2 (Choi et al., 2009). An homozygous nonsense variant (c.105C > A; p. Tyr35Ter) of AIMP2 has recently been associated with severe neurodevelopmental alterations (Shukla et al., 2018) like those resulting from other ARS mutations. We herein update the literature and describe a novel AIMP2 variant from one Iranian infant with progressive neurological disorder characterized by lack of development, microcephaly, and skeletal abnormalities. Whole-exome sequencing (WES) revealed a variant c.A463T in the homozygous state in exon 3 (NM_001,326,607) of AIMP2 [p.(K155X)]. MRI of the brain showed global cerebral atrophy and extensive white matter involvement.

MATERIALS AND METHODS

Patients

The proband was referred to Dr. Mazaheri's lab, Yazd Medical University, Yazd, Iran to confirm the clinical diagnosis. Before being referred for the WES analysis, the proband was reviewed by a metabolic specialist and referred to the lab for the genetic

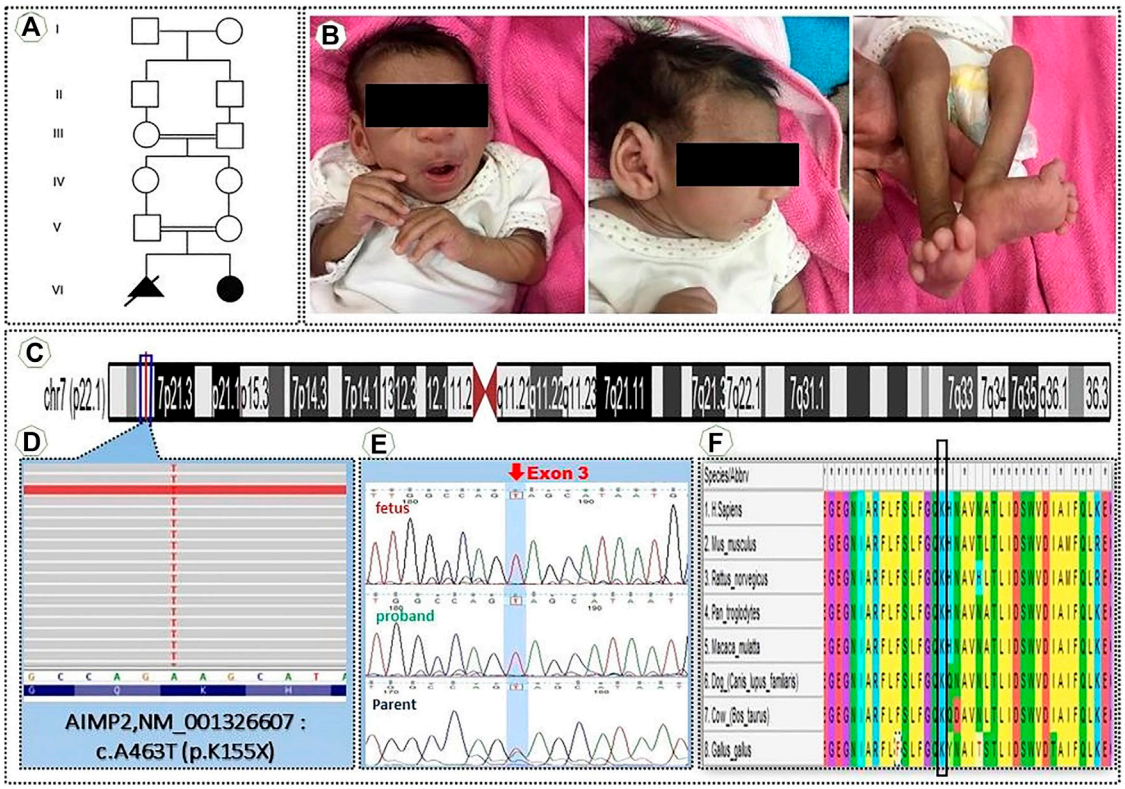


FIGURE 1 | (A) Pedigrees of the family **(B)** Clinical manifestations of the proband **(C)** An Integrative Genomic Viewer (IGV) of homozygous nonsense variant AIMP2″ NM001326607 c.A463T, K155X **(D)** Position in the homozygous state in exon 3 of the variant c.A463T (NM_001,326,607) of AIMP2 [p.(K155X)]. **(E)** Sanger sequencing of genomic DNA confirmed the presence of a homozygous mutation in the proband and heterozygous mutation in the parents. **(F)** This alignment shows this position (variant) highly conserved between vertebrates.

etiology. After written informed consent was obtained from the proband's parents, a blood sample was collected to perform a WES test. At the time of referral, clinical details and MRI results were provided by the patient's family. All genomic DNA was isolated from the peripheral leukocytes using a QIAamp DNA Blood Midi kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The DNA samples were stored at −20°C until use. DNA integrity was evaluated by performing 1% agarose gel electrophoresis. Written informed consent was obtained from the patient's proband for publication of this case report and any accompanying images.

Clinical Presentation

A female infant was born by normal vaginal delivery at 34 weeks gestation to consanguineous healthy parents (**Figure 1A**). Proband's family did not refers any health problems. The second trimester prenatal ultrasonic screening for fetal malformation didn't reveal any abnormality. The birth weight was 1,060 g, and her head circumference was 24.5 cm. She was in the neonatal intensive care unit for approximately 25 days after birth due to respiratory failure, hypotension, and clinical features of oligohydramnios. She had delayed cry at birth, microcephaly, and shortened extremities. At 3 months of age, the proband was referred to the medical Centre for a weight reduction of 110 gr. During hospitalization, she received Folic acid tablets, vitamin B6 tablets, pantoprazole tablets and Lanoxin syrup and after 1 week with good health conditions, she was discharged from the hospital. Additionally, karyotype revealed normal female constitution (46, XX). At 7 months of age, because of weight loss, severe anemia, microcephaly, and tilted ankle the proband was referred to our Specialist Centre for further evaluation, including a WES analysis (**Figure 1B**).

Radiological Findings

Brain MRI in axial T1WI, T2WI and FLAIR images and sagittal and coronal T2WI were performed. MRI of the brain on a postnatal day 5 showed minor microcephaly, brain atrophy with increased sub arachnoid space, but no changes in the white/gray matter intensity in both hemispheres. A follow-up MRI was performed at 2 months of age and revealed a mildly delayed myelination of brain parenchyma and mild cortical atrophy. At 7 months of age the MRI showed microcephaly,

extreme cerebral atrophy, and white matter loss associated with dilatation of lateral ventricle and third vertical. Periventricular white matter hyperintensity in T2/FLAIR in both central hemispheres is seen that could be related to gliosis.

Whole-Exome Sequencing

Whole-exome sequencing was performed on peripheral leukocyte DNA of the proband. After DNA extraction, exome capture was done using Agilent SureSelect Human All Exon platform following the manufacturer's protocol (Chen et al., 2015). The quality of FASTQ files was inspected making use of FastQC. Then reads were mapped to GRCh38 utilizing Burrows-Wheeler Aligner (BWA) and duplicates were marked using Picard, complied with by base recalibration, variant calling, and genotyping making use of Genome Analysis Toolkit (GATK). Variations (SNP and INDEL) were filtered based on GATK advised criteria. Ultimately, variants were filtered to maintain those of medical significance. Only variations of exonic or splice site, with less than 1% frequency in the 1,000 Genomes and ExAC databases that were not identified as benign in the clinical data sources passed the filters. These shortlisted annotated variations were further studied for analysis of pathogenic variants. The interpretation of the pathogenicity of the sequence variants is based on the most recent criteria released by the American College of Medical Genetics and Genomics (ACMG) (Richards et al., 2015). Sanger sequencing and PCR using specific primers F-5′-CACCCTTTCCCATGTCATCAG-3′ and R-5′- CCTTCAGTT TAGCGTCATTCCA-3 ′ primers for AIMP2 were used to validate the variant in both parents.

RESULTS

Genetic Analysis Identified AIMP2 as Candidate Gene

For confirmation of the genetic diagnosis of leukodystrophy, the proband's DNA was submitted to WES. The number of annotated, potentially pathogenic variants was reduced to 510 complying with bioinformatics filtering, which allows to identify exonic/splicing variants, excluding synonymous and benign sequence variants, and reported homozygous variants with minor allele frequency (MAF) < 0.001. BAM files were converted into variant call format (VCF), and VCF files were used as input for PLINK, a whole genome association analysis toolset, to elucidate the degree of homozygosity from the WES data. This previously unreported mutation most likely influences the AIMP2 [p(K155X)] at exon 3 (**Figures 1C,D**). This variant was confirmed as homozygous in the proband by bidirectional Sanger sequencing. Sequencing of the parental samples confirmed their carrier status of the novel mutation thus ruling out other causes of apparent homozygosity, such as uniparental isodisomy, allele dropout or copy number variations. (**Figure 1E**). The Combined Annotation Dependent Depletion (CADD) score, which allows to score the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome, was 42 for this variant (https://cadd.gs.washington.edu/snv), and the allele frequency of this variant in 1,000 genomes of different populaces in heterozygous and homozygous states was 0. (https://gnomad.broadinstitute.org/). This mutation is located in a region that is highly conserved among vertebrates (**Figure 1F**).

DISCUSSION

Here we report one case with a phenotype of a severe neurodevelopmental disorder with microcephaly and skeletal radiographic abnormalities with a nonsense variant in AIMP2.

ARSs are essential enzymes that bind specific amino acids to tRNAs prior to protein synthesis. Three non-enzymatic proteins-the ARS-interacting multifunctional proteins (AIMPs)-associate nine different ARSs into a multisynthetic macromolecular complex in higher eukaryotes. Many of these complex-forming ARSs are involved in a wide variety of regulatory processes such as transcription, translation, splicing, inflammation, angiogenesis, and apoptosis. Similar to ARS, AIMPs have functions unrelated to their supporting role in protein synthesis, acting as a cytokine in the control of angiogenesis, immune response, and wound repair, and have crucial regulatory actions in cell proliferation and DNA repair processes (Park et al., 2005). Previous studies have observed associations between several mutations in ARSs and encephalopathies, peripheral neuropathies, and other neurological disorders. In particular, mutations affecting 10 cytosolic ARSs appear to be related to Charcot-Marie-Tooth disease and related neuropathies, whereas mutations affecting 14 mitochondrial ARSs appear to be associated with severe leukoencephalopathies (Ognjenović and Simonović, 2018).

Among auxiliary proteins, p43/AIMP1 has been associated with hypomyelinating leukodystrophy-3 characterized by progressive neurodegeneration, microcephaly, generalized brain atrophy, progressive contractures, and spasticity (Elia et al., 2012; Accogli et al., 2019). AIMP2/p38 is a non-synthetase part of the multi-ARS structure. The p38 protein contains a lysyl tRNA synthetase binding domain, a presumptive leucine-zipper theme, and a C-terminal glutathione S-transferase-like domain, as well as having sequence patterns, which are the binding sites for protein-protein communications (http://www.ebi.ac.uk/InterPro/protein/Q131 55). In addition to its key action in assembling the multi-ARS complex, AIMP2/p38 also is able to suppress cell proliferation by down-regulating c-Myc (Kim et al., 2013). In addition, AIMP2 would also appear to be involved in the pathogenesis of Parkinson's disease by inducing neural cell death (Ochiai et al., 2021). AIMP2 enhances the ubiquitin-mediated degradation of TNF receptor-associated factor 2, an essential regulator of the tumor necrosis factor-a (TNF-a) signaling pathway, by enhancing the apoptotic response of cells to TNF-a (Choi et al., 2009). In addition, through downregulation of c-MYC it regulates the anti-proliferative activity of transforming growth factor (TGF)-b (Kim et al., 2019). Here we identified a novel pathogenic variant (c.A463T) in AIMP2 [p(K155X)]. To date, only in a single study a nonsense variant in AIMP2 has been described in two unrelated consanguineous families with three affected children each

with microcephaly, intellectual disability, seizures refractory to therapy, and spastic quadriparesis (Shukla et al., 2018). MRI showed cerebral, cerebellar, and spinal cord atrophy, with symmetrical T2 hypo-intensities in the bilateral basal ganglia and thinning of the corpus callosum. Whole-exome sequencing of three affected individuals showed c.105C > A (p. Tyr35Ter) variant in AIMP2. In agreement with the previous study, our results suggest that deleterious variants in AIMP2,, might be associated with neurodevelopmental disturbances in humans.

The study has some limitation. First, no functional studies are performed to corroborate the effect of the A463T mutation, such as histochemical staining, RNA extraction and RT-qPCR or Western blot experiments, so it was not possible to evaluate mRNA AIMP2 levels, and possible correlation between mRNA levels and phenotypical presentation. Moreover, the analysis is limited to only one patient.

Despite the limitation, this case confirms the importance of a genetic diagnosis, which provides additional information in the diagnosis of the proband and parents as well as appropriate genetic counselling for the family, including prenatal diagnosis.

CONCLUSION

In conclusion, we present a novel AIMP2 mutation in an Iranian infant with clinical and radiological signs of leukodystrophy. The crucial factor in the diagnosis of leukodystrophy is the high importance of medical signs, genetic testing, and MRI findings. Due to the relatively high cost of straight sequencing of genes, these findings could serve for an earlier and definitive diagnosis, which represents a major milestone in the patient's journey to inform for disease-specific therapies, research eligibility and for symptomatic care.

DATA AVAILABILITY STATEMENT

The datasets for this article are not publicly available due to concerns regarding participant/patient anonymity. Requests to access the datasets should be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Shahid Sadoughi University of Medical Sciences (IR.SSU.REC.1399.159). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

HJ, AS, and MM designed the study and drafted the manuscript. MY performed the laboratory analysis and data analysis. HZ and AS supported the interpretation of the data. SD'a and HJ revised and edited the manuscript. MM and PL critically reviewed the final version of the article. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

We are grateful to patients and families for their cooperation.

REFERENCES

Accogli, A., Russell, L., Sébire, G., Rivière, J.-B., St-Onge, J., Addour-Boudrahem, N., et al. (2019). Pathogenic Variants in AIMP1 Cause Pontocerebellar Hypoplasia. *Neurogenetics* 20 (2), 103–108. doi:10.1007/s10048-019-00572-7

Ashrafi, M. R., and Tavasoli, A. R. (2017). Childhood Leukodystrophies: a Literature Review of Updates on New Definitions, Classification, Diagnostic Approach and Management. *Brain Dev.* 39 (5), 369–385. doi:10.1016/j.braindev.2017.01.001

Ashrafi, M. R., Amanat, M., Garshasbi, M., Kameli, R., Nilipour, Y., Heidari, M., et al. (2020). An Update on Clinical, Pathological, Diagnostic, and Therapeutic Perspectives of Childhood Leukodystrophies. *Expert Rev. Neurother.* 20 (1), 65–84. doi:10.1080/14737175.2020.1699060

Bonkowsky, J. L., Nelson, C., Kingston, J. L., Filloux, F. M., Mundorff, M. B., and Srivastava, R. (2010). The burden of Inherited Leukodystrophies in Children. *Neurology* 75 (8), 718–725. doi:10.1212/wnl.0b013e3181eee46b

Chen, R., Im, H., and Snyder, M. (2015). Whole-exome Enrichment with the Agilent Sureselect Human All Exon Platform. *Cold Spring Harb. Protoc.* 2015 (7), 626–633. doi:10.1101/pdb.prot083659

Choi, J. W., Um, J. Y., Kundu, J. K., Surh, Y.-J., and Kim, S. (2009). Multidirectional Tumor-Suppressive Activity of AIMP2/p38 and the Enhanced Susceptibility of AIMP2 Heterozygous Mice to Carcinogenesis. *Carcinogenesis* 30 (9), 1638–1644. doi:10.1093/carcin/bgp170

Elia, M., Amato, C., Bottitta, M., Grillo, L., Calabrese, G., Esposito, M., et al. (2012). An Atypical Patient with Cowden Syndrome and PTEN Gene Mutation Presenting with Cortical Malformation and Focal Epilepsy. *Brain Dev.* 34 (10), 873–876. doi:10.1016/j.braindev.2012.03.005

Kim, J. H., Han, J. M., and Kim, S. (2013). Protein-Protein Interactions and Multi-Component Complexes of Aminoacyl-tRNA Synthetases. *Top Curr. Chem.* 344, 119–144. doi:10.1007/128_2013_479

Kim, M., Kim, H., Kim, D., Park, C., Huh, Y., Jung, J., et al. (2019). Fluorescence-based Analysis of Noncanonical Functions of Aminoacyl-tRNA Synthetase-Interacting Multifunctional Proteins (AIMPs) in Peripheral Nerves. *Materials* 12 (7), 1064. doi:10.3390/ma12071064

Numata, Y., Gotoh, L., Iwaki, A., Kurosawa, K., Takanashi, J.-i., Deguchi, K., et al. (2014). Epidemiological, Clinical, and Genetic Landscapes of Hypomyelinating Leukodystrophies. *J. Neurol.* 261 (4), 752–758. doi:10.1007/s00415-014-7263-5

Ochiai, A., Sawaguchi, S., Memezawa, S., Seki, Y., Morimoto, T., Oizumi, H., et al. (2021). Knockdown of Golgi Stress-Responsive Caspase-2 Ameliorates HLD17-Associated AIMP2 Mutant-Mediated Inhibition of Oligodendroglial Cell Morphological Differentiation. *Neurochem. Res.*, 1–15. doi:10.1007/s11064-021-03451-6

Ognjenović, J., and Simonović, M. (2018). Human Aminoacyl-tRNA Synthetases in Diseases of the Nervous System. *RNA Biol.* 15 (4-5), 623–634. doi:10.1080/15476286.2017.1330245

Parikh, S., Bernard, G., Leventer, R. J., van der Knaap, M. S., van Hove, J., Pizzino, A., et al. (2015). A Clinical Approach to the Diagnosis of Patients with

Leukodystrophies and Genetic Leukoencephelopathies. *Mol. Genet. Metab.* 114 (4), 501–515. doi:10.1016/j.ymgme.2014.12.434

Park, S. G., Ewalt, K. L., and Kim, S. (2005). Functional Expansion of Aminoacyl-tRNA Synthetases and Their Interacting Factors: New Perspectives on Housekeepers. *Trends Biochem. Sci.* 30 (10), 569–574. doi:10.1016/j.tibs.2005.08.004

Park, S. G., Choi, E. C., and Kim, S. (2010). Aminoacyl-tRNA Synthetase-Interacting Multifunctional Proteins (AIMPs): a Triad for Cellular Homeostasis. *IUBMB life* 62 (4), 296–302. doi:10.1002/iub.324

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17 (5), 405–424. doi:10.1038/gim.2015.30

Shukla, A., Das Bhowmik, A., Hebbar, M., Rajagopal, K. V., Girisha, K. M., Gupta, N., et al. (2018). Homozygosity for a Nonsense Variant in AIMP2 Is Associated with a Progressive Neurodevelopmental Disorder with Microcephaly, Seizures, and Spastic Quadriparesis. *J. Hum. Genet.* 63 (1), 19–25. doi:10.1038/s10038-017-0363-1

Srivastava, S., Cohen, J., Vernon, H., Bara nano, K., McClellan, R., Jamal, L., et al. (2014). Clinical Whole Exome Sequencing in Child Neurology Practice. *Ann. Neurol.* 76 (4). 473–483. doi:10.1002/ana.24251

Van Der Knaap, M. S., Breiter, S. N., Naidu, S., Hart, A. A. M., and Valk, J. (1999). Defining and Categorizing Leukoencephalopathies of Unknown Origin: MR Imaging Approach. *Radiology* 213 (1), 121–133. doi:10.1148/radiology.213.1.r99se01121

Vanderver, A., Prust, M., Tonduti, D., Mochel, F., Hussey, H. M., Helman, G., et al. (2015). Case Definition and Classification of Leukodystrophies and Leukoencephalopathies. *Mol. Genet. Metab.* 114 (4), 494–500. doi:10.1016/j.ymgme.2015.01.006

Vanderver, A. (2016). Genetic Leukoencephalopathies in Adults. *Continuum: Lifelong Learn. Neurol.* 22 (3), 916–942. doi:10.1212/con.0000000000000338

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Mazaheri, Yavari, Zare Marzouni, Stufano, Lovreglio, D'Amore and Jahantigh. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 24 January 2022
doi: 10.3389/fgene.2021.728827

Emerging Roles for LGR4 in Organ Development, Energy Metabolism and Carcinogenesis

Linlin Yang[1,2,3], *Jing Wang*[1,2,3], *Xiaodi Gong*[1,2,3], *Qiong Fan*[1,2,3], *Xiaoming Yang*[1,2,3], *Yunxia Cui*[1,2,3], *Xiaoyan Gao*[1,2,3], *Lijuan Li*[1,2,3], *Xiao Sun*[1,2,3], *Yuhong Li*[1,2,3]* and *Yudong Wang*[1,2,3]*

[1]Department of Gynecological Oncology, The International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China, [2]Shanghai Municipal Key Clinical Specialty, Shanghai, China, [3]Shanghai Key Laboratory of Embryo Original Disease, Shanghai, China

The leucine-rich repeats containing G protein-coupled receptor 4 (LGR4) belonging to G protein-coupled receptors (GPCRs) family, had various regulatory roles at multiple cellular types and numerous targeting sites, and aberrant LGR4 signaling played crucial roles in diseases and carcinogenesis. On the basis of these facts, LGR4 may become an appealing therapeutic target for the treatment of diseases and tumors. However, a comprehensive investigation of its functions and applications was still lacking. Hence, this paper provided an overview of the molecular characteristics and signaling mechanisms of LGR4, its involvement in multiple organ development and participation in the modulation of immunology related diseases, metabolic diseases, and oxidative stress damage along with cancer progression. Given that GPCRs accounted for almost a third of current clinical drug targets, the in-depth understanding of the sophisticated connections of LGR4 and its ligands would not only enrich their regulatory networks, but also shed new light on designing novel molecular targeted drugs and small molecule blockers for revolutionizing the treatment of various diseases and tumors.

Keywords: LGR4, development, immunity, metabolism, malignancy, targeted therapy

OPEN ACCESS

Edited by:
Joao Batista Teixeira da Rocha, Federal University of Santa Maria, Brazil

Reviewed by:
Qiang Hou, Wenzhou Medical University, China
Nikolai V. Litviakov, Tomsk National Research Medical Center of the Russian Academy of Sciences, Russia

***Correspondence:**
Yuhong Li
13022107898@163.com
Yudong Wang
wangyudong@shsmu.edu.cn

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *22 June 2021*
Accepted: *30 December 2021*
Published: *24 January 2022*

Citation:
Yang L, Wang J, Gong X, Fan Q, Yang X, Cui Y, Gao X, Li L, Sun X, Li Y and Wang Y (2022) Emerging Roles for LGR4 in Organ Development, Energy Metabolism and Carcinogenesis. Front. Genet. 12:728827. doi: 10.3389/fgene.2021.728827

INTRODUCTION

The leucine-rich repeat-containing G protein-coupled receptors (LGRs) are highly conserved proteins of the G protein-coupled receptors (GPCRs) family, identified as multiple repeats of leucine-rich repeats (LRRs) in the extracellular domain (Luo and Hsueh, 2006). The leucines in LRRs act as the dominant hydrophobic residue and play a critical role in the interactions between proteins (Kobe and Deisenhofer, 1993; 1995), which also allow forming a unique tertiary structure (Kajava, 1998). LRRs are connected to a seven-transmembrane (TM) domain that are related to G protein activation via a cysteine-rich region (Xu et al., 2013) and exerts its biological effects by binding to the ligand (Wang D. et al., 2013). LGRs could be subdivided into three groups (groups A–C) (Barker et al., 2013). Group A is consisted of lutenizing hormone receptor (LHR), follicle-stimulating hormone receptor (FSHR) and thyroid stimulating hormone receptor (TSHR) (Van Loy et al., 2008). Group B contains LGR4, LGR5, and LGR6, which exhibit a high degree of homology and function as receptors for the Wnt-activating R-spondins (Dubey et al., 2020). Group C is consisted of LGR7 and LGR8, which recognize relaxin (Yoshino et al., 2020) and insulin-like peptide 3 (INSL3) (Yeom et al., 2021), respectively.

The leucine-rich repeat-containing GPCR 4 (also called as LGR4) molecule is 107 kb and located on chromosome 11 at position 11p14-p13. It is composed of 17 LRRs and each contains 24 amino acids (McDonald et al., 1998). LGR4 signaling plays a functional role in self-renewal of stem cells by binding to R-spondin, thus potentiating Wnt signaling. R-spondin interacts with LGR4 inhibiting the expression of ZNRF3 and RNF43, the negative mediators of Wnt signaling that induce degradation of the Wnt receptor Frz and coreceptors LRP5/6 (Hao et al., 2012), thereby elevating the concentration of Wnt receptors and increasing the signaling response. R-spondin-bound LGR4 could also bind directly to LRP6 to boost the phosphorylation of LRP6 in response to Wnt-Fzd combination (Carmon et al., 2011). Clathrin (Glinka et al., 2011) and Norrin (Deng et al., 2013) were also reported to be needed for LGR4-mediated Wnt signaling. The ligand activated LGR4 triggers G-protein through GTP binding as well. Then coupled Gαs is dissociated from LGR4 and activates adenylyl cyclase (AC) elevating the level of second messenger cyclic AMP (cAMP), which activates protein kinase A (PKA) and in turn, phosphorylates cre-binding protein (CREB), thus enhancing the expression of its target genes, such as mineralocorticoid receptor (Wang et al., 2012), estrogen receptor α (Li et al., 2010). However, the ligands initiating cAMP/PKA pathway by LGR4 still remains unidentified.

Accumulating evidence supported by recent studies has shown that LGR4 is indispensable in embryonic growth, multiple organ development (Knight and Hankenson, 2014), energy metabolism (Li et al., 2014), ischemia/reperfusion injury (Li Z. et al., 2019) and the maintenance of stem cell self-renewal in intestine (Mustata et al., 2011), prostate (Luo et al., 2013), and mammary gland (Wang Y. et al., 2013). LGR4, as a new RANKL receptor, could counteract RANKL-driven osteoclastogenesis and enhance osteoblast maturation, mineralization (Luo et al., 2016) and vascular calcifcation (Carrillo-López et al., 2020). It also plays an oncogenic role in various human cancers, such as multiple myeloma (van Andel et al., 2017), thyroid carcinoma (Kang et al., 2017), and ovarian cancer (Wang Z. et al., 2020), etc. This paper will systematically summarize LGR4's role in organ development, energy metabolism and tumor formation, which may provide the fundamental basis for the targeted gene therapy in the future.

METHODS

We screened MEDLINE, PubMed, and Google Scholar for relevant literatures from 2000 to 2021 and subjected the corresponding references to this review. This compiling was limited to studies written in English by using the terms "Leucine-rich repeats containing G protein-coupled receptor 4", "LGR4", "GPR48", focusing on the biological function of LGR4, and various roles of LGR4 in development, immunity, energy metabolism, oxidative stress, and carcinogenesis.

SIGNAL TRANSDUCTION OF LGR4 GENE IN CELLS

LGR4 is a transmembrane receptor of the GPCRs superfamily that is characterized by a large extracellular Leucine-rich domain that recognizes and interacts with its ligands (Van Loy et al., 2008), thus regulating numerous cellular processes (**Figure 1**). Many studies have explored the mechanisms of LGR4 gene. Researchers demonstrated that LGR, R-spondin, and ZNRF3 or RNF43 formed a ternary complex (Hao et al., 2012; Koo et al., 2012), alleviating ZNRF3/RNF43 clearance of Frizzled-LRP Wnt coreceptor, thus activating Wnt signaling (Hao et al., 2012). Rspos-LGRs signaling is essential for embryogenesis and cell protection (Knight and Hankenson, 2014). Another study found that RSPO-LGR4-IQGAP1 promoted MEK1/2-modulated phosphorylation of LRP5/6 in β-catenin-dependent manner or regulated actin dynamics in β-catenin-independent way, thus potentiating Wnt signaling (Carmon et al., 2014). Wang, D. et al. showed that the furin-like cysteine-rich domains of RSPO1 could interact with LGR4, thus inducing its biological activities (Wang D. et al., 2013). In addition, Park, S. et al. explored that full-length LGR4 interacted with E3 ligases RNF43 and ZNRF3 forming a complex to reduce ubiquitylation degradation of Wnt receptors and activate Wnt/β-catenin signaling (Park et al., 2020). They also explored that RSPO2 activated Wnt/β-catenin signaling with no binding to LGR4 or LGR5 (Park et al., 2018), which differentiated from other RSPO molecules (**Figure 2**). Moreover, Geng, A. et al. demonstrated that Rspo1 interacted with LGR4 and then activated cAMP-PKA signaling to elevate Esr1 expression and increase mammary side branches in a Wnt-independent manner (Geng et al., 2020), which provided a novel mechanism for estrogen-related diseases.

LGR4 IN ORGAN DEVELOPMENT

The Role of LGR4 Gene in the Development of Maxillofacial Organs

The gene of LGR4 was widely expressed in adrenal gland, kidney, heart, stomach, intestine, bone/cartilage and other tissues, and was first found to be associated with developmental processes (Yi et al., 2013), which was validated by immunohistochemical staining in wild-type mice. LGR4 deletion led to the severe pre- and postnatal lethality of mice explaining the significant importance of the LGR4 signaling for cell survival and growth (Mazerbourg et al., 2004). Meanwhile, LGR4 deficiency attenuated the expression of ATF4 via cAMP-PKA-CREB pathway to modulate definitive erythropoiesis (Song et al., 2008). Furthermore, LGR4 knockdown blocked GnRH neuron development by impairing Wnt/β-catenin signaling, leading to delayed puberty (DP) (Mancini et al., 2020). In addition, LGR4 regulated long term depression (LTD) at parallel fiber-PC (PF-PC) by modulating Creb signaling, suggesting its role in cerebellar ataxia (Guan et al., 2014). As a marker for hair follicle stem cell (Kim et al., 2019), LGR4 promoted the hair cycle progression by activating Akt/mTOR signaling, Wnt/β-catenin signaling and decreasing BMP signaling, thus regulating the development of

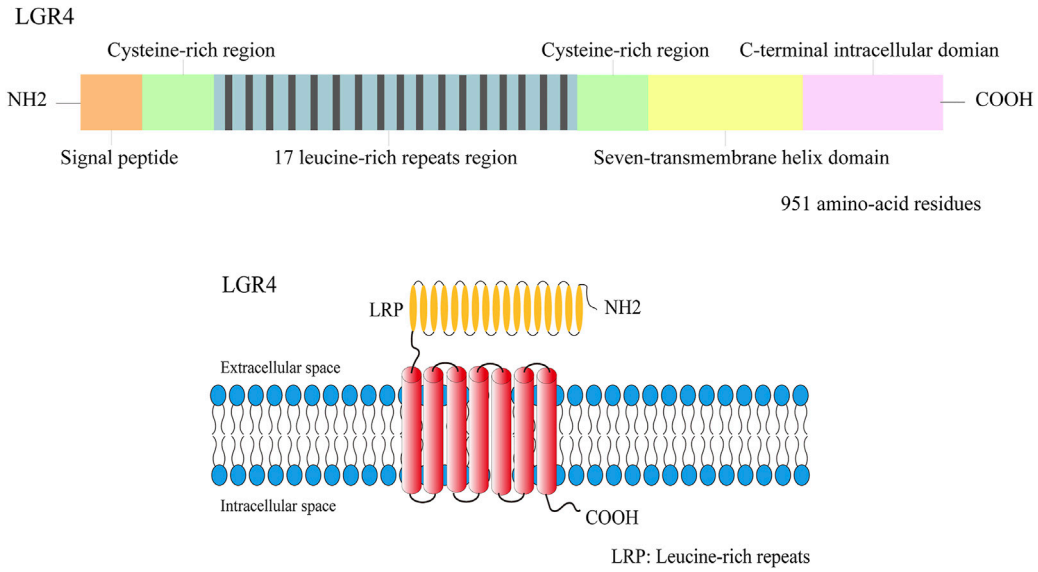


FIGURE 1 | The structure illustration of the LGR4 protein. LGR4 was a member of transmembrane receptor, its N-terminal domain was comprised of 17 leucine-rich repeats region which was flanked by the N-/C- cysteine-rich regions. A seven-transmembrane domain and a C-terminal intracellular region were detected in LGR4.

hair follicle (HF). The skin epithelia-specific deletion of LGR4 lead to reduced numbers of LGR5$^+$, and actively proliferating HF stem cells without affecting the number of quiescent HF stem cells, resulting in compromised HF regeneration after transplantation (Ren et al., 2020). It also played a critical role in controlling hair cell differentiation in cochlea (Żak et al., 2016). In eye development, Jin, C. et al. showed that LGR4 was highly expressed in cells of eyelids. LGR4 deficiency inhibited the phosphorylation of EGFR, thus blocking epithelial cell proliferation and migration in eyelid development (Jin et al., 2008). Additionally, LGR4 regulated keratinocyte proliferation through EGFR signaling pathway and the inhibitor of EGFR tyrosine kinase or its ligand HB-EGF could suppress cellular processes (Wang et al., 2010). The keratinocyte motility was reduced in LGR4 deleting mice leading to eye-open at birth (EOB) phenotype (Kato et al., 2007). Weng, J. et al. showed that deletion of LGR4 downregulated Pitx2 via cAMP-CREB signaling, thus inducing ocular anterior segment dysgenesis (ASD) (Weng et al., 2008). Further, the antioxidant enzymes CAT and SOD1 were downregulated in the lens epithelial cells of LGR4 deficiency mice, resulting in cataract formation (Zhu et al., 2015). MicroRNA let-7b promoted the apoptosis of lens epithelial cell by targeting LGR4, thus inducing age-related cataract (Dong et al., 2016). LGR4 could also be directly targeted by miR-34a and the downregulation of LGR4 inhibited the proliferation, migration of retinal pigment epithelial cell (Hou et al., 2016). LGR4 inactivation decreased histone demethylases Jmjd2a and Fbxl10 via cAMP-CREB signaling, thereby reducing the expression of development-related genes and increasing cell apoptosis, resulting in aniridia-mental retardation syndrome (Yi et al., 2014). Intriguingly, RSPO2 cooperated with WNT9b potentiating WNT/β-catenin signaling to regulate mouse facial development, while Jin, Y. R. et al. reported that LGR4/5/6 receptors played less critical roles in their cooperation to control facial development (Jin et al., 2020). Consistently, Szenker-Ravi, E. et al. explored that RSPO2, without the interaction with LGR4/5/6 receptors, served as a direct antagonistic to the ligases of RNF43 and ZNRF3, which together governed limb development (Szenker-Ravi et al., 2018). Then more relevant studies were needed to validate this specific mechanism.

LGR4 in Bone Differentiation and Mineralization

LGR4 was believed to be a novel receptor for RANKL, it could induce the cAMP-PKA-CREB signaling to control the expression of Atf4 and its target genes Ocn, Bsp and collagen in osteoblasts. LGR4 deficiency in murine led to a delay in osteoblast differentiation, while increasing the activity of osteoclasts, thus regulating bone remodeling (Luo et al., 2009). LGR4 could also compete with the canonical receptor RANK to bind RANKL, suppress RANKL-RANK-TRAF6 signaling cascade and activate the Gαq and GSK3-β signaling, thus inhibiting the activity of NFATC1 and blocking RANKL-induced osteoclast differentiation (Luo et al., 2016). Jang, Y. et al. identified that the mutated RANKL protein acted as a competitive inhibitor of RANKL, bound only to the receptor LGR4, induced GSK-3β phosphorylation and inhibited NFATc1 nuclear translocation, and thereby preventing osteoclast differentiation (Jang et al., 2021). Additionally, miR-34c promoted osteoclast differentiation through targeting LGR4, activating NF-κB and GSK3-β signaling (Cong et al., 2017). LGR4 was found to be preferentially expressed in osteoblasts and played a vital role in canonical Wnt signaling, thus regulating osteoblastogenesis and bone homeostasis (Zhang et al., 2021). MiR-193a-3p inhibited

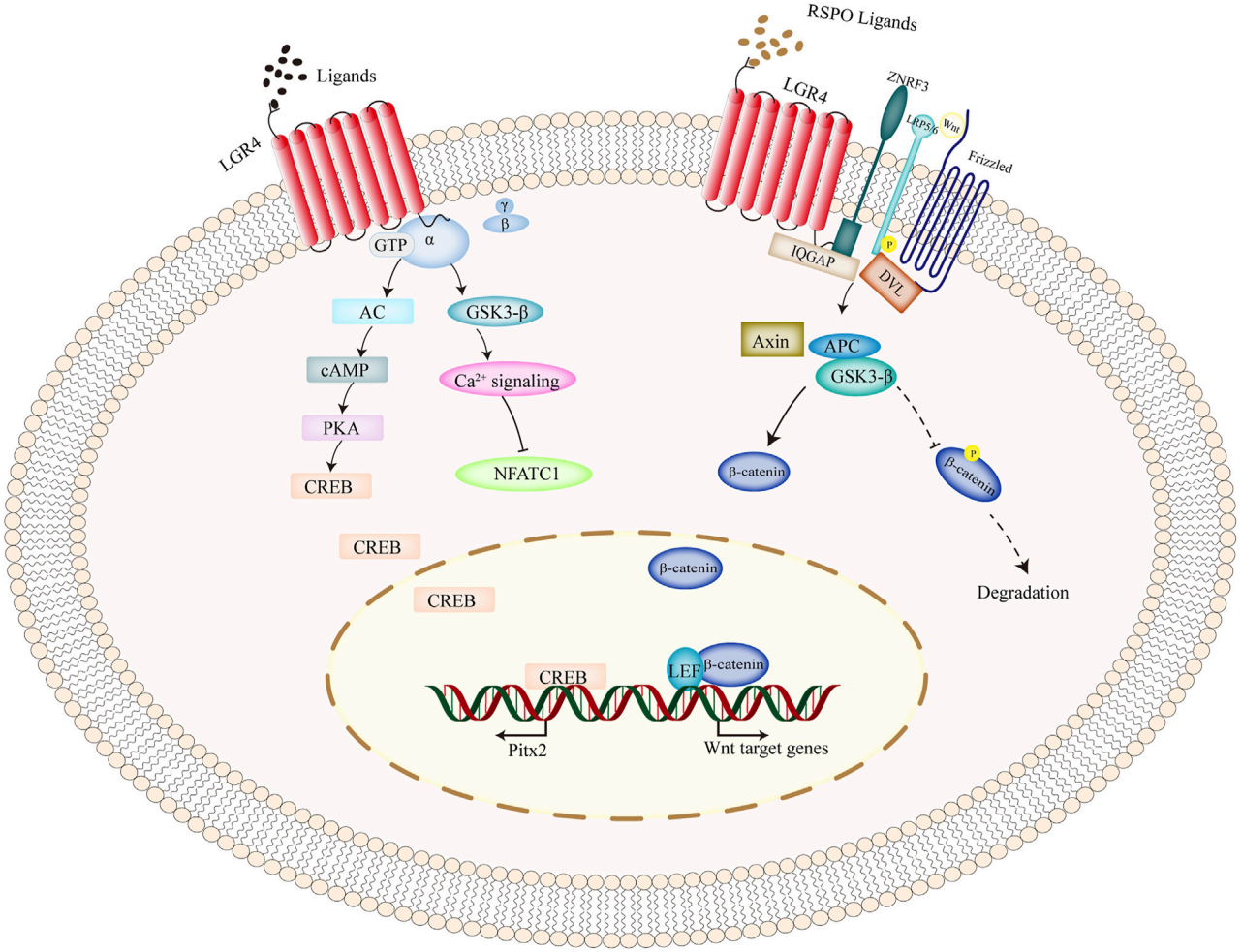


FIGURE 2 | Intracellular signaling pathways of LGR4 gene. As a classical GPCRs molecule, upon ligands binding to LGR4, and it activated heterotrimeric G-proteins to transduce the cytoplasmic signal. Moreover, upon the stimulation of RSPO ligands, simultaneous binding of ZNRF3 and LGR4 suppressed the ubiquitination of frizzled receptor, promoted LRP phosphorylation, recruited IQGAP1 and increased its interaction with DVL, and thus activating the canonical Wnt signaling. β-catenin was prevented from degradation, translocated into the nucleus, and interacted with the transcription factors of TCF/LEF to induce the transcription of its target genes. Abbreviations: GPCRs, G-protein-coupled receptors; LGR4, leucine-rich repeat-containing G protein-coupled receptor 4; RSPO, R-spondin.

osteoblast differentiation through regulating LGR4/ATF4 signaling (Wang et al., 2018). Zhang, M. et al. reported that RSPO3-LGR4 system inhibited osteogenesis of human adipose-derived stem cells by negatively regulating ERK/FGF signalling (Zhang M. et al., 2017). The compressive force (CF) in alveolar bone led to the elevation of RANK and decrease of LGR4, thus inducing bone differentiation (Matsuike et al., 2018). LGR4 played an essentia role in the sequential development of molars by Wnt/β-catenin/LEF1 signaling (Yamakami et al., 2016). The silencing of LGR4 suppressed proliferation and osteogenic differentiation of stem cells from apical papillae (SCAPs) through inhibiting the Wnt/β-catenin pathway (Zhou et al., 2017). Arima, M. et al. reported that RSPO2-LGR4 accelerated osteoblastic differentiation by Wnt/β-catenin signaling in immature human periodontal ligament cells (Arima et al., 2019).

Further study indicated that there was a close correlation between LGR4 genotypes and bone mineral density (BMD), including the association between rs11029986 of LGR4 and total fat mass (TFM) (Yu et al., 2020). Additionally, researchers identified that a rare nonsense mutation within LGR4 gene (c.376C > T) was strongly associated with lower BMD and osteoporotic fractures by whole-genome sequencing of Icelandic individuals (Styrkarsdottir et al., 2013). Meanwhile, by the technology of next generation sequencing (NGS), Li, C. et al. showed that LGR4 was significantly differentially expressed between postmenopausal cases with impaired BMD and control group with normal values (Li C. et al., 2020). Moreover, a study also reported that LGR4-deficiency inhibited the differentiation of bone marrow mesenchymal stem cells (BMSCs), reduced bone mass, thus suppressing fracture healing (Sun et al., 2019). MiR-137 was correlated

with an increased risk of fracture in patients with osteoporosis by targeting LGR4/ALP expression (Liu and Xu, 2018). In addition, a latest finding showed that the novel RANKL variant induced the expression of LGR4 by the GSK3-β signaling, thus suppressing the activity of NFATc1 and inhibiting osteoporosis (Ko et al., 2021). Shi, G. X. et al. identified that Rspo1/LGR4 could enhance osteogenesis by Wnt/β-catenin signaling. LGR4 might be a novel molecular protein in the transmission of mechanical stimuli to bone reorganization (Shi et al., 2017). Further research found that LGR4 induced the expression of pyruvate dehydrogenase kinase 1 (pdk1) via the canonical Wnt/β-catenin signaling. Loss-of-function experiments indicated that LGR4 deficiency resulted in decreased osteogenic effects together with aerobic glycolysis (Yang et al., 2021). The above studies revealed the important mechanisms of LGR4 in bone differentiation and development, indicating its great potential in the treatment of osteolysis diseases.

LGR4 Gene in the Development of Heart, Liver, Kidney, Gonads, and Other Important Organs

Recent study indicated that Rspo3-LGR4 axis played a crucial role in heart development (Da Silva et al., 2018). LGR4 was found to be a molecular biomarker for cardiac progenitors (den Hartogh et al., 2016). In addition, study found that RSPO-LGR4/5-ZNRF3/RNF43 system regulated metabolic liver zonation by Wnt/β-catenin signalling (Planas-Paz et al., 2016). In contrast, Planas-Paz, L. et al. explored that LGR4/5-modulated WNT/β-Catenin signaling was dispensable for ductular reaction (DR) in biliary epithelial cells (BECs), while YAP and mTORC1 signaling were necessary for this process. LGR5 and AXIN2 were detected in hepatocytes to facilitate liver regeneration (Planas-Paz et al., 2019). More researches were needed on the important role of LGR4 gene in liver metabolism. Additionally, researchers proposed that the N-termini and 7TM domains of LGR5/LGR4 modulated WNT signaling in a ligand-dependent manner, while their C-termini and rhodopsin-like 7TM domains activated NF-κB signaling in a ligand-independent manner to control the survival of LGR5+ stem cells and intestinal crypts (Lai et al., 2020). Moreover, Dang, Y. et al. identified that the deficiency of LGR4 led to polycystic lesions and renal fibrosis by regulating Wnt/PCP signaling but not the TGF-β/Smad pathway (Dang et al., 2014). The serious renal hypoplasia was observed in LGR4 null mice (Kato et al., 2006). Conversely, Vidal, V. P. et al. explored that knockout of LGR4/5/6, the receptors of R-spondins, did not intervene with MET of nephron progenitor, revealing LGR-independent role in kidney development (Vidal et al., 2020). It was possible that the differences in mouse species and experimental conditions led to the differences in conclusion. A study revealed that high parathyroid hormone (PTH) elevated the expression of LGR4 and RANKL to facilitate vascular calcification (VC) by PTH1R/PKA activation (Carrillo-López et al., 2020). Luo, W. et al. identified that LGR4 promoted prostate development and stem cell differentiation by Wnt, Hedgehog and Notch1 signaling (Luo et al., 2013). The secretome from activation of stromal-androgen receptor (AR) maintained the basal state of epithelial cells by LGR4/β-Catenin/ΔNP63α signaling and did not induce the clonogenic growth of benign prostate hyperplasia (BPH) (Chauhan et al., 2020). In addition, researchers found that LGR4, not LGR5 was indispensable for the hematopoietic differentiation of human pluripotent stem cells (hPSCs) by regulating transforming growth factor beta (TGF-beta)-SMAD2/SMAD3 signaling, thus controlling mesoderm induction and hematopoietic development (Wang Y. et al., 2020).

In mammary gland, LGR4 induced the expression of Sox2 to facilitate mammary development via Wnt/β-catenin/Lef1 signaling pathway (Wang Y. et al., 2013). LGR4 could also promote corpus luteum maturation by WNT-mediated EGFR-ERK signaling, thus maintaining female fertility (Pan et al., 2014). Hsu, P. J. et al. explored that the LGR4 splice variant which encoded only the ectodomain of LGR4 (LGR4-ED) acted as an antagonist to suppress the LGR4/RSPO2/Norrin-mediated Wnt signaling thus controlling gonadal development (Hsu et al., 2014). Meanwhile, LGR4 could activate ERalpha by cAMP/PKA signaling to control the development of male reproductive tract (Li et al., 2010), and LGR4 inactivation led to the abnormal organization of it (Mendive et al., 2006). Hoshii, T. et al. explored that LGR4 knockout reduced the expression of estrogen receptor (ESR1), controlling elongation and differentiation of epididymal ducts (Hoshii et al., 2007). Further study explored that abnormal development of female gonads was observed in LGR4 (−/−) female mice. Rspo1/LGR4 was essential for ovarian somatic cell development via the Wnt/beta-catenin/Lefl/Axin2 signaling (Koizumi et al., 2015). In uterine receptivity, Kida, T. et al. explored that the phosphorylated PR was significantly reduced and persistent epithelial E2 receptor α was activated in LGR4 knockout mice, leading to impaired uterine receptivity (Kida et al., 2014). The reduced uterine glands and decidualization was observed in LGR4 knockout female mice by decreasing the secretion of LIF, implying the function of LGR4 in uterine gland development (Sone et al., 2013). Moreover, Gαq/11-coupled LGR4 promoted uterine receptivity by triggering PR signaling (de Oliveira et al., 2019). MiR-449a could promote caprine endometrial receptivity by targeting 3′-untranslated region of LGR4 (An et al., 2017). By using a bovine endometrial epithelial cell inflammation model and a mouse lipopolysaccharide-mediated endometritis model, the author confirmed that miR-34a/miR-193a-3p was upregulated by IL-1β and suppressed the level of the LGR4 3′UTR, which in turn amplified the inflammatory response through activating the phosphorylation of NF-κB p65 pathway, suggesting miR-34a/miR-193a-3p-LGR4 playing a pivotal role in endometritis (Ma et al., 2021; Yin et al., 2021). Furthermore, the gene of LGR4 modulated a WNT-NR5A2 signaling cascade facilitating the secretion, maturation and steroidogenesis of oviduct epithelial cells to safeguard the development and function of oviduct in mice (Tan et al., 2021). Other studies reported that akermanite

TABLE 1 | Diverse roles of LGR4 in organ development.

Organ development	Signaling	Effect	References
Cerebellum	LGR4-Creb signaling	LGR4 (–/–) mice led to impairing long term depression	Guan et al. (2014)
Hair follicle	LGR4-Akt/mTOR signaling, Wnt/β-catenin signaling and decreasing BMP signaling	LGR4 promotes the normal hair cycle	Ren et al. (2020)
Facial organs	WNT9b:RSPO2-WNT/β-catenin	Wnt9b; Rspo2 double mutant mice displayed facial defects	Jin et al. (2020)
Pubertal development	LGR4-Wnt/β-catenin signaling	Mice deficient in LGR4 had delayed onset of puberty	Mancini et al. (2020)
Ocular cells	MicroRNA let-7b- LGR4; LGR4- cAMP-CREB-Pitx2; LGR4-EGFR	The antioxidant enzymes were decreased in LGR4 (–/–) mice	Jin et al. (2008); Weng et al. (2008); Wang et al. (2010); Zhu et al. (2015); Dong et al. (2016)
Liver	RSPO-LGR4/5-ZNRF3/RNF43- Wnt/β-catenin	Recombinant RSPO1 protein increased liver size	Planas-Paz et al. (2016)
Kidney	LGR4-WNT signaling; LGR4-cAMP-CREB-Jmjd2a/Fbxl10; MiR-34a- LGR4	LGR4 deficiency led to polycystic lesions and renal fibrosis	Dang et al. (2014); Yi et al. (2014); Hou et al. (2016)
Intestine	LGR5/4-NF-κB and WNT signaling	LGR5/LGR4 promoted the growth of intestinal crypts	Lai et al. (2020)
Vascular cells	PTH-PTH1R/PKA-LGR4	High PTH increases LGR4 thereby favouring vascular calcification	Carrillo-López et al. (2020)
Hematopoietic cell	LGR4-cAMP-PKA-CREB-ATF4; R-spondin1/R-spondin3/LGR4/ZNRF3-TGF-beta-SMAD2/SMAD3 signaling	$LGR4^{-/-}$ fetuses displayed anemia, deletion of LGR4 limited hematopoietic differentiation	Song et al., 2008; Wang et al., 2020b
Osteoclast	LGR4-Gαq-GSK3-β-NFATC1; RSPO-LGR4-IQGAP1-Wnt/β-catenin; MiR-34c-LGR4-NF-□b/GSK3-β; MiR-137-LGR4-ALP; RANKL-GSK3-β signaling-LGR4	LGR4 deficiency exhibit osteoclast hyperactivation	Carmon et al. (2014); Luo et al. (2016); Cong et al. (2017); Liu and Xu, (2018); Park et al. (2020); Jang et al. (2021); Ko et al. (2021)
Osteoblast	LGR4-cAMP-PKA-CREB-Atf4; LGR4-WNT/β-catenin; MiR-193a-3p-LGR4/ATF4; RSPO3-LGR4-ERK/FGF; RSPO1/2-LGR4-Wnt/β-catenin; LGR4-Wnt/β-catenin-pdk1/LEF1	Deletion of LGR4 results in a delay in osteoblast differentiation	Luo et al. (2009); Yamakami et al. (2016); Shi et al. (2017); Zhang et al. (2017a); Zhou et al. (2017); Wang et al. (2018); Arima et al. (2019); Yang et al. (2021); Zhang et al. (2021)
Gonads	LGR4-ED-LGR4/RSPO2/Norrin-Wnt; RSPO1/LGR4-Wnt/β-catenin- Lefl/Axin2; LGR4-WNT-EGFR-ERK signaling	LGR4-ED acted as an antagonist controlling gonadal development	Hsu et al. (2014); Pan et al. (2014); Koizumi et al. (2015)
Mammary gland	LGR4-Wnt/β-catenin/Lef1-Sox2; Rspo1-LGR4-cAMP-PKA-Esr1	LGR4 (–/–) mice had delayed ductal development	Wang et al. (2013c); Geng et al. (2020)
Prostate	LGR4-Wnt, Notch, Sonic Hedgehog signaling; LGR4/β-Catenin/ΔNP63α; LGR4-cAMP/PKA-ERalpha	LGR4 loss blocked differentiation of prostate cells	Li et al. (2010); Luo et al. (2013); Chauhan et al. (2020)
Uterine	LGR4-PR/LIF; Gαq/11-LGR4-PR; MiR-449a-LGR4; MiR-34a/miR-193a-3p-LGR4-NF-κB	LGR4 KO down-regulated progesterone signaling, affecting uterine receptivity and led to endometritis	Sone et al. (2013); Kida et al. (2014); An et al. (2017); de Oliveira et al. (2019); Ma et al. (2021); Yin et al. (2021)
Oviduct	LGR4-WNT-NR5A2 signaling	The loss of LGR4 ultimately impaired the epithelial secretion	Tan et al. (2021)

elevated the expressions of integrinβ1, LGR4, LGR5, and LGR6, accompanied by triggering the Wnt/β-catenin pathway (Wang F. et al., 2020), thereby accelerating re-epithelialization in wound healing (**Table 1** and **Figure 3**). Taken together, LGR4 was widely expressed in various tissues and played a fundamental role in modulating their development in a tissue-specific manner.

REGULATION OF LGR4 GENE IN IMMUNE-RELATED DISEASES

Cancer immunotherapy has demonstrated marvelous efficacy in clinical trials targeting negative immune checkpoint mediators including CTLA-4 and PD-1 (Pardoll, 2012). Tumor-associated macrophages (TAMs) constituted the major leukocytic infiltrate and could be polarized to a proinflammatory "M1" or a immunosuppressive "M2" phenotype by tumor-derived chemokines or cytokines (Mantovani et al., 2004). Tan, B. et al. demonstrated that Rspo-LGR4 axis functioned as a novel pathway aggravating M2-like macrophage polarization through noncanonical Erk/Stat3 signaling by recruiting IQGAP1 and MEK1/2, thus maintaining protumoral TAMs, promoting tumor progression and enhancing the resistance of Lewis lung carcinoma (LLC) cells to anti-PD1 treatment. Accordingly, LGR4 deficiency only in macrophages was able to activate both macrophage-mediated innate and T-cell–mediated adaptive antitumor immune responses, thus overcoming resistance to checkpoint blockade therapy, which gave this axis great potential as a promising therapeutic target in macrophage-targeting strategies (Tan et al., 2018). Moreover, LGR4 negatively modulated CD14 transcriptional activation and inhibited TLR2/4-associated immune response via cAMP-PKA-

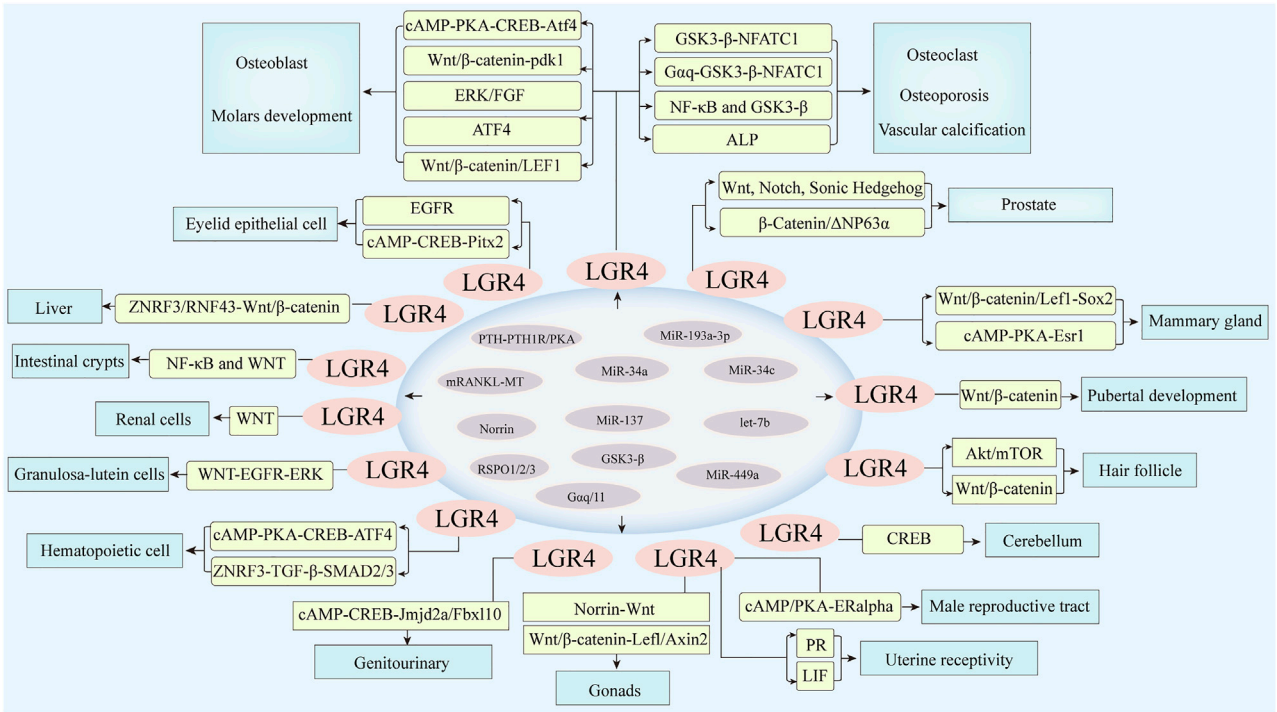


FIGURE 3 | LGR4 was expressed in multiple tissues involved in the regulation of cell differentiation and tissue development. MicroRNAs including miR-34a, miR-34c, miR-193a-3p, miR-137, and miR-449a, let-7b could bind to the 3′ untranslated region of its target gene LGR4, resulting in its translational repression or degradation. Meanwhile, RSPO1/2/3, Norrin, and RANKL could act as ligands of LGR4. PTH-PTH1R/PKA, Gαq/11, GSK3-β could also activate LGR4, which then stimulated corresponding molecules and induced downstream signaling, such as Wnt/β-catenin, cAMP/PKA, Akt/mTOR, and so forth, thus regulating the differentiation of various organs.

TABLE 2 | The functions of LGR4 in immune-related diseases.

Diseases	Signaling	Effect	References
Lung carcinoma associated macrophage	Rspo/LGR4/Erk/Stat3-enhanced macrophage M2 polarization	Blocking Rspo-LGR4 signaling overcame lung carcinoma resistance and suppressed tumor growth	Tan et al. (2018)
Macrophage in innate immunity	LGR4-cAMP-PKA-CREB signaling-CD14	LGR4-deficiency led to increased immune response	Du et al. (2013)
Intestinal inflammation	LGR4-Wnt/β-catenin signaling	LGR4 (−/−) mice exhibited stronger intestinal inflammation	Liu et al. (2013b)
Traumatic osteoarthritis	LGR4-NF-κB signaling	Upregulation of LGR4 expression can inhibit the secretion of the inflammatory factors	Ge et al. (2019)
Keratinocyte in venous ulcers	MiR-34-LGR4-GSK-3β-induced p65 phosphorylation-NF-κB signaling	Knockout of LGR4 impaired wound closure with enhanced inflammation	Wu et al. (2020)
Vesicular stomatitis virus	Vesicular stomatitis virus-LGR4	LGR4 knockdown suppressed VSV infection	Zhang et al. (2017b)

CREB signaling (Du et al., 2013). Liu, S. et al. identified that LGR4 prevented intestinal inflammation by modulation of the Wnt/β-catenin signaling (Liu S. et al., 2013). Another study explored that the expression of LGR4 was reduced in traumatic osteoarthritis. Overexpression of LGR4 could suppress the joint inflammation by inhibiting NF-κB pathway (Ge et al., 2019). MiR-34a and miR-34c enhanced inflammatory response and delayed chronic wound healing of venous ulcers by direct targeting LGR4. MiR-34-LGR4 axis reduced GSK-3β-induced phosphorylation of p65 at Ser468, while enhancing phosphorylation at Ser536, activating NF-kB signaling, thus regulating inflammatory response of keratinocyte (Wu et al., 2020). Zhang, N. et al. showed that LGR4 functioned as a vesicular stomatitis virus (VSV)-specific host factor enhancing VSV infection, LGR4 knockdown reduced the levels of VSV(Zhang N. et al., 2017). Furthermore, LGR4 was related to the infection of severe acute respiratory syndrome coronavirus (SARS-CoV) (Liu H.-L. et al., 2020) (**Table 2**), and may be a potential gene target for therapy. Together these data pointed towards a key role for LGR4 in immune-related diseases, though further studies were needed to better establish this, especially in the field of tumor immunotherapy.

TABLE 3 | Regulatory effects of LGR4 gene on metabolic diseases.

Pathway	Effect	References
LGR4-cAMP/PKA	LGR4 KO mice had aldosterone resistance	Wang et al. (2012)
Rspo1/Rspo3/LGR4	Injection of Rspo1 or Rspo3 inhibited food intake	Li et al. (2014)
R-spondin1-LGR4-Pomc	LGR4 KO mice didn't exhibit a suppressed appetite	Otsuka et al. (2019)
LGR4-Ampk/Sirt1/Pgc1α pathway	LGR4 ablation enhanced fuel shift	Sun et al. (2015)
Rspo1/Rspo3-LGR4-AMPKα-SREBP2 pathway	Rspo1/Rspo3-LGR4 signaling suppresses cholesterol synthesis	Liu et al. (2020b)

THE ROLE OF LGR4 IN METABOLIC DISEASES

The expression of LGR4 was detected by *in situ* hybridization assay and was elevated in hypothalamic energy homeostatic areas and co-localizated with some energy homeostatic neurons suggesting that it may regulate energy homeostasis (Van Schoore et al., 2005). In line with this, Otsuka, A. et al. showed that R-spondin1-LGR4 suppressed appetite of mice by upregulating Pomc gene expression. The suppressed food intake was not observed in LGR4 knockdown mice (Otsuka et al., 2019). Rspo1/Rspo3/LGR4 forming novel system to regulate feeding behavior (Li et al., 2014). Moreover, LGR4 ablation promoted the energy switch from glucose to fatty acid by activating Ampk/Sirt1/Pgc1α pathway (Sun et al., 2015). Importantly, the LGR4 A750T variant was investigated by Sanger sequencing, and the result found that it was correlated with central obesity (Zou et al., 2017). In addition, Shi et al. demonstrated that genetic polymorphisms of the LGR4 gene were related to bone and obesity phenotypes in Chinese nuclear families with female children (Shi et al., 2021). Furthermore, LGR4 was found to modulate energy balance and body weight through regulating the translation of white fat into brown fat (Wang J. et al., 2013). A recent study reported that RSPO1/LGR4 axis was involved in obesity-related renal fibrosis through promoting Wnt/β-catenin signaling pathway, providing a potential therapeutic target for the obesity-related chronic kidney disease (CKD) (Su et al., 2021). Likewise, LGR4, as a adipocytokine, was closely related to the progression of diabetes and hypertension (Li B. et al., 2019). Notably, Wang, J. et al. identified that LGR4 elevated mineralocorticoid receptor (MR) expression by cAMP/protein kinase A pathway to improve aldosterone responsiveness and maintain electrolyte homeostasis (Wang et al., 2012). Meanwhile, Rspo1/Rspo3-LGR4 signaling attenuated cholesterol synthesis in hepatocytes by activating the phosphorylation of AMPKα and suppressing SREBP2 nuclear translocation (Liu S. et al., 2020) (**Table 3**). Hence, the function of LGR4 in energy metabolism was being broadly studied, a better understanding of the molecular mechanism underlying various metabolic pathways involved by LGR4 will help in future development of new treatments for metabolic diseases.

THE ROLE OF LGR4 IN OXIDATIVE STRESS DAMAGE

Oxidative stress exerted an increased impact on pathophysiology of osteoporosis. However, the correlation between LGR4 and oxidative stress remained unknown. Pawaputanon Na et al. explored that the treatment of hydrogen peroxide decreased the expression of LGR4 in osteoblastic cells (Pawaputanon Na Mahasarakham et al., 2017). Rspo1-LGR4 axis protected hepatocytes against acute injury by suppressing NF-κB-p65 via Wnt3a/β-catenin pathway (Li Z. et al., 2019). Of note, Liu, S. et al. reported that Rspo3-LGR4 system protected hepatocytes from dimethyloxalylglycine (DMOG)-caused hypoxia/reoxygenation (H/R) damage by Wnt3a/β-catenin (Liu et al., 2018). LGR4 also controlled mitochondrial function and oxidative stress by activating ERK signaling, thus protecting myocardium against ischemia-reperfusion (I/R) damage (Chen T. et al., 2021). Additionally, Singla, B. et al. showed that RSPO2-LGR4 interaction reduced lymphangiogenesis by inhibiting PI3K-AKT-eNOS and Wnt-β-catenin pathway. LGR4 silencing could block this process and facilitate cholesterol drainage from atherosclerotic arteries (Singla et al., 2020). Conversely, Huang, C. K. et al. showed that LGR4 could induce proinflammatory responses in myocardial infarction (MI) by elevating the expression of AP-1 via CREB-regulated c-Fos, Fosl1, and Fosb activation. Knockout of LGR4 could mitigate ischemic injury (Huang et al., 2020). Study uncovered that radiation therapy elevated the expression of Rspo1 and LGR4 in bone mesenchymal stem cells (BMSCs). Exogenous Rspo1 reduced radiation-induced bone damage by Rspo1-LGR4-mTOR-autophagy signaling (Chen X. et al., 2021) (**Table 4**). Therefore, LGR4 was involved in oxidative stress and cellular damage, a deeper exploration of its mechanisms was still required. From a broader and longer-term perspective, these investigations may slow or even reverse the onset of stress injuries by targeting LGR4 gene.

SUGGESTED ROLES OF LGR4 IN CARCINOMAS

The gene of LGR4 emerged as a critical player in regulation of tumor growth and progression (Gong et al., 2015; Liang et al., 2015). Zhu, Y. B. et al. identified that LGR4 induced Wnt/β-catenin signaling to promote cancer cell growth and migration (Zhu et al., 2013). Accordingly, Kang, Y. E. et al. demonstrated that R-spondin 2 and LGR4 were overexpressed in thyroid cancer. The upregulated LGR4 enhanced cell proliferation and migration by inducing the phosphorylation of ERK and GSK3β and activating β-catenin pathway (Kang et al., 2017). The expression of LGR4 was elevated after prostate cancer radiotherapy. LGR4 ablation inhibited AR/CREB1 expression, promoted γH2A.X staining and reduced tumor growth (Liang

TABLE 4 | Increased impacts of LGR4 gene on oxidative stress response.

Pathway	Effect	References
Rspo1-LGR4-Wnt3a/β-catenin-NF-κB-p65	LGR4 protected hepatocytes from injury	Liu et al. (2018); Li et al. (2019b)
LGR4-ERK signaling-oxidative stress	LGR4 protected cardiomyocyte against I/R	Chen et al. (2021a)
RSPO2-LGR4-PI3K-AKT-Enos/Wnt-β-catenin	LGR4 silencing promoted lymphangiogenesis	Singla et al. (2020)
LGR4-CREB-mediated c-Fos/Fosl1/Fosb/AP-1	LGR4 knockout infarcts had reduced inflammatory	Huang et al. (2020)
Rspo1-LGR4-mTOR-autophagy	Exogenous Rspo1-LGR4 alleviated radiation-induced bone loss	Chen et al. (2021b)

et al., 2021). Moreover, LGR4 was overexpressed in human prostate cancer and correlated with shorter disease-free survival. The knockdown of LGR4 inhibited cell migration and reversed EMT by elevating the expression of E-cadherin (Luo et al., 2017). Also, Zhang, J. et al. revealed that LGR4 promoted Jmjd2a/AR signaling to enhance AR binding to PSA promoter, thus contributing to inducing prostate tumorigenesis (Zhang et al., 2016). LGR4 may also facilitate the growth of prostate cancer via the PI3K/Akt/mTOR signaling (Liang et al., 2015). Furthermore, hypoxia exposure downregulated the expression of miR-137, which targeted LGR4 and prevent the migration and EMT of prostate cancer by inhibiting EGFR/ERK signaling (Zhang et al., 2020). When a hypomorphic mutation of LGR4 was observed in chronic myelogenous leukaemia (CML) stem cells, it induced inadequate disease-initiating capacity of CML cells in mice (Naka et al., 2020). With regard to acute myeloid leukemia (AML), Salik, B. et al. also identified that LGR4 was upregulated and cooperated with HOXA9 in AML. RSPO3-LGR4 interaction enhanced proliferation and self-renewal of AML blasts (Salik et al., 2020). Study identified that LGR4 expression was associated with poor prognosis in breast cancer. The down-regulation of LGR4 inhibited Wnt signaling and epithelial-mesenchymal transition (EMT), suppressing mammary tumorigenesis (Yue et al., 2018). Additionally, Yue et al. uncovered a Wnt-independent LGR4-EGFR signaling axis enhancing breast cancer cell metastasis with broad implications for the targeted therapy of breast cancer (Yue et al., 2021). Importantly, Gao, Y. et al. identified that the short hairpin RNA of LGR4 could suppress the invasion and metastasis of HeLa cells (Gao et al., 2009). In addition, Berti FCB. et al. found that for HPV16-cervical squamous carcinoma (CESC), eleven miRNAs were shared by XIST/LGR4 and XIST/ZNF81 lncRNA-mRNA co-expressed pairs, implying an increased effect on their ultimate biological effect. Moreover, XIST/miR-23a-3p/LGR4 had a remarkable impact on the overall survival of HPV18- and HPV16-CESC patients (Berti et al., 2021). LGR4/ELF3 axis could also enhance the epithelial phenotype of serous ovarian cancer and was mediated by WNT7B/FZD5 interaction (Wang Z. et al., 2020). By analyzing 122 serous ovarian cancer tissues and 41 paired para-carcinoma tissues, Zeng, Z. et al. showed that the expression of LGR4 was elevated in serous ovarian cancer and it could be used as an independent prognostic predictor (Zeng et al., 2020). Specifically, a study revealed that LGR4 was upregulated in colorectal cancer (CRC) and enhanced tumor metastasis by PI3K/Akt and MAPK/ERK1/2-mediated β-catenin/TCF signaling (Wu et al., 2013). Gao, Y. et al. demonstrated that enhanced expression of LGR4 caused by p27Kip1 deficiency promoted metastasis of colon cancer cells (Gao et al., 2006). Currently, CircLGR4 was highly expressed in colorectal tumors. CircLGR4-derived peptide activated LGR4 and then promoted Wnt/β-catenin signaling, thus driving colorectal tumorigenesis (Zhi et al., 2019). Additionally, LncGata6 recruited the NURF complex onto the promoter of Ehf to enhance its transcription, which elevated the expression of LGR4/5 to activate Wnt signaling, thus promoting the progression of colorectal cancer (Zhu et al., 2018). Wang, Y. et al. showed that the expression of LGR4 was elevated in uveal melanoma cells. MiR-34a negatively controlled the expression level of LGR4, thus downregulating the markers of the EMT and MMP2, thereby impacting the aggressiveness of uveal melanoma (Hou et al., 2019). Study reported that LGR4 was considered as an independent prognostic marker for patients with non-small cell lung cancer (NSCLC) (Li R. et al., 2020). MiR-449b as a tumor suppressor prevented the proliferation of NSCLC by downregulating LGR4 (Yang et al., 2018) and LGR4 was perceived as a high-risk immune gene in NSCLC (Sun et al., 2020). The aberrant activation of RSPO3-LGR4-IQGAP1 system promoted tumor aggressiveness in Keap1-deficient lung adenocarcinomas. Knockdown of LGR4 led to reduction in cell proliferation (Gong et al., 2015). Additionally, Zhang, L. et al. showed that Rspo2-LGR4 system promoted the growth and migration of tongue squamous cell carcinoma (TSCC). It potentiated β-catenin pathway by enhancing phosphorylation of LRP6, while reducing phosphorylation of GSK-3β, contributing to subsequent upregulation of TCF-1 and its downstream genes CD44, c-Myc, and Cyclin D1, thus facilitating the progression of TSCC (Zhang et al., 2019). LGR4 could also promote the proliferation of glioma by activating Wnt/β-catenin signaling (Yu et al., 2013). Stat3 could elevate LGR4 expression by binding to LGR4 promoter, thereby regulating osteosarcoma progression (Liu J. et al., 2013). Furthermore, van Andel, H. et al. identified that R-spondin/LGR4 axis promoted multiple myeloma (MM) by activating aberrant Wnt/β-catenin signaling (van Andel et al., 2017). LGR4 acted as an essential positive factor for inducing skin tumorigenesis by activating MEK1/ERK1/2/AP-1 and Wnt/β-catenin pathways (Xu et al., 2016). Conversely, Souza, S. M. et al. detected that LGR4 was expressed in a larger number of cells in normal gastric mucosa than in primary gastric carcinomas and not specific to gastric cancer cells, predominantly affecting the expression of β-catenin in membrane-complex but rarely in nucleus, suggesting a controversial function of LGR4, and which was positively correlated with cell proliferation but inversely related to cancer progression (Souza et al., 2019) (**Table 5** and **Figure 4**). Due to various carcinogenic factors, multiple cellular microenvironment and a variety of cell types, the complicated functions of LGR4 in

TABLE 5 | The gene of LGR4 involving in the process of multiple tumors.

Cancer subtypes	Pathway	Effect	References
Skin carcinoma	LGR4-MEK1/ERK1/2/AP-1 and Wnt/β-catenin pathways	LGR4 was crucial for skin carcinogenesis	Xu et al. (2016)
Glioma	LGR4-Wnt/β-catenin	LGR4 overexpression promoted cell proliferation	Yu et al. (2013)
Uveal melanoma	MiR-34a-LGR4-MMP2	Knockdown of LGR4 attenuated the aggressiveness	Hou et al. (2019)
Tongue carcinoma	Rspo2-LGR4-Wnt/β-catenin	Elevated LGR4 promoted growth	Zhang et al. (2019)
Thyroid carcinomas	R-spondin2-LGR4-p-ERK-p-LRP6–p-GSK3β-β-catenin	Elevated expression of LGR4 promoted proliferation and migration	Kang et al. (2017)
Lung cancer	RSPO3-LGR4- IQGAP1, MiR-449b-LGR4	Knockdown of LGR4 decreased tumor growth	Gong et al. (2015); Yang et al. (2018)
Colon cancer	p27Kip1- LGR4, CircLGR4-LGR4-Wnt/β-catenin, LGR4-GSK3β-PI3K/Akt-MAPK-ERK1/2-catenin/TCF-Cyclin D1/c-Myc, LncGata6-NURF-Ehf-LGR4/5-Wnt	LGR4 expression was associated with colorectal tumorigenesis	Gao et al. (2006); Wu et al. (2013); Zhu et al. (2018); Zhi et al. (2019)
Acute myeloid leukemia	RSPO3-LGR4-HOXA9	RSPO3-LGR4 interaction promoted proliferation	Salik et al. (2020)
Multiple myeloma	IL-6/STAT3-LGR4/R-spondin- Wnt/β-catenin	LGR4 expression was driven by IL-6/STAT3 signaling and allowed MM cells to hijack R-spondins	van Andel et al. (2017)
Osteosarcoma	Stat3-LGR4	Overexpression of Stat3 promoted LGR4 expression	Liu et al. (2013a)
Breast cancer	LGR4- Wnt/β-catenin signaling, LGR4-EGFR signaling	LGR4 down-regulation decreased tumor growth and lung metastasis	Zhu et al. (2013); Yue et al. (2018); Yue et al. (2021)
Prostate cancer	LGR4-EMT, LGR4-Jmjd2a/AR signaling-PSA, MiR-137-LGR4-EGFR/ERK, LGR4-AR/CREB1 expression, LGR4-PI3K/Akt/mTOR	LGR4 knockdown impaired cell migration	Liang et al. (2015); Zhang et al. (2016); Luo et al. (2017); Zhang et al. (2020); Liang et al. (2021)
Ovarian cancer	WNT7B/FZD5-LGR4/ELF3 axis	LGR4 overexpression enhanced tumorisphere formation capacity	Wang et al. (2020c)
Cervical cancer	XIST/LGR4, XIST/miR-23a-3p/LGR4	—	Berti et al. (2021)

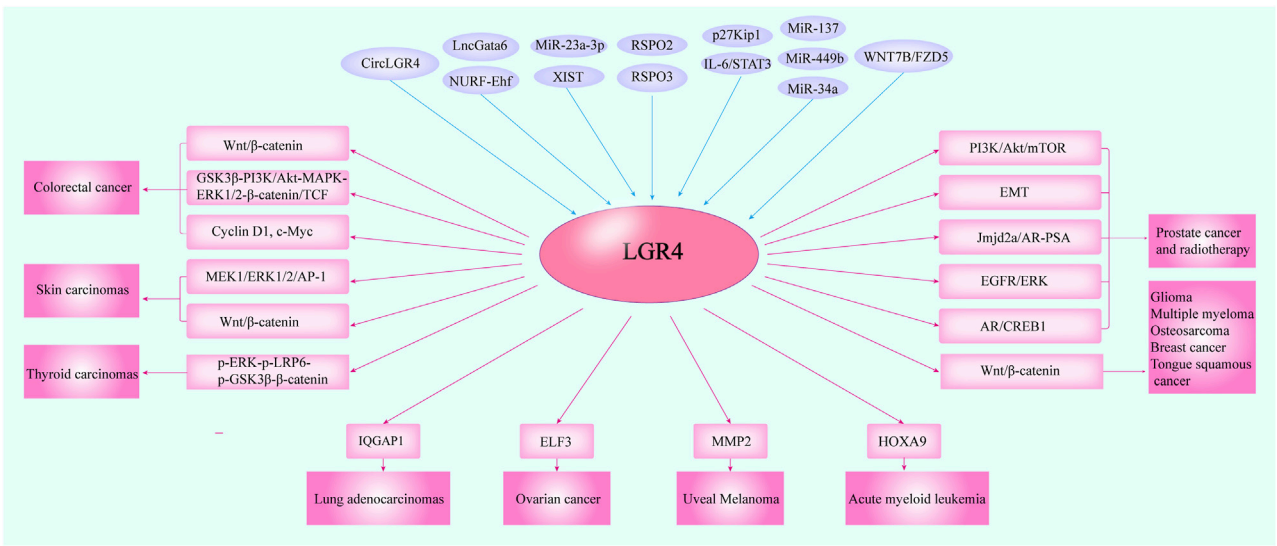


FIGURE 4 | The cross-talking of LGR4 with other molecules in multiple carcinomas. Several studies indicated that LGR4 was overexpressed in cancer tissues. The molecules including circLGR4, lncGata6-NURF-Ehf, XIST/miR-23a-3p, miR-137, miR-449b, miR-34a, IL-6/STAT3, p27Kip1, WNT7B/FZD5, and RSPO2/3 could function as upstream regulators of LGR4 and mediate its expression. Then, LGR4 could modulate the proteins such as IQGAP1, ELF3, MMP2, and HOXA9 to control tumor progression and trigger the Wnt/β-catenin, MEK1/ERK1/2/AP-1, PI3K/Akt pathways to promote the initiation and metastasis of a variety of malignancies.

carcinogenesis needed further exploration. Therefore, given the multitude of indications for the relevant oncogenic role of LGR4, it remained valuable to further investigate the clinical potential of anti-LGR4 monoclonal antibodies specifically in cancer patients that harbored LGR4 alterations, which would hopefully provide more insight beneficial to the development of novel treatment strategies against LGR4 driven cancer.

DISCUSSION AND FUTURE DIRECTIONS

In general, a growing body of evidence indicated that LGR4 was widely expressed in diverse tissues from the early embryogenesis to adulthood, participated in the differentiation and development of various organs, involved in immune-related diseases, metabolic diseases as well as oxidative stress damage and contributed to multiple cancer progression. The repertoire of the physiological and pathological roles of LGR4 in numerous cellular processes provided a systematic and comprehensive understanding of its functional characteristics, thereby offering a novel diagnostic biomarker and therapeutic target for a range of diseases.

The interaction between LGR4 and its ligands including RSPOs, Norrin, RANKL, and could activate the downstream Wnt pathway and other G protein-associated pathways. Several researches implicated that inhibitors or antagonists of the RSPOs/LGR4/Wnt/β-catenin axis could suppress tumor metastasis and recurrence. Moreover, blockage of the LGR4 signaling would result in a decreased population and impaired the migration ability of cancer stem cells, which may lay a solid theoretical basis for the development of small molecule blockers and antagonizing antibodies to suppress LGR4 pathway.

Additionally, with regard to the LGR4/cAMP/PKA signaling, the endogenous ligands that activated LGR4 were yet to be elucidated clearly. Meanwhile, the function of LGR4 through its newly discovered ligand RANKL was mainly detected in maintaining homeostasis of bone tissue, whereas their involvement in malignancies was rarely explored. Hence, further detailed mechanism investigation was crucial for representing the spatiotemporal profile of LGR4 and might open up a new avenue for molecular targeted therapy for tumors and other diseases.

AUTHOR CONTRIBUTIONS

LY designed and wrote this paper. JW, XG, and QF participated in literature searching and sorting. XY, YC, XG, LL, and XS analyzed the data, YL and YW assessed and edited this work. All authors have read and approved the final version of the manuscript.

FUNDING

This work was funded by Shanghai Municipal Key Clinical Specialty (No. shslczdzk06302), National Natural Science Foundation of China (No. 81172477, 81402135), the Project of the Science and Technology Commission of Shanghai Municipality (No. 17441907400), and Shanghai Jiao Tong University Medicine-Engineering Fund (No. YG2017MS41).

ACKNOWLEDGMENTS

The authors wish to thank the International Peace Maternity and Child Health Hospital for the excellent scientific research platform.

REFERENCES

An, X., Liu, X., Zhang, L., Liu, J., Zhao, X., Chen, K., et al. (2017). MiR-449a Regulates Caprine Endometrial Stromal Cell Apoptosis and Endometrial Receptivity. *Sci. Rep.* 7 (1), 12248. doi:10.1038/s41598-017-12451-y

Arima, M., Hasegawa, D., Yoshida, S., Mitarai, H., Tomokiyo, A., Hamano, S., et al. (2019). R-spondin 2 Promotes Osteoblastic Differentiation of Immature Human Periodontal Ligament Cells through the Wnt/β-Catenin Signaling Pathway. *J. Periodont Res.* 54 (2), 143–153. doi:10.1111/jre.12611

Barker, N., Tan, S., and Clevers, H. (2013). Lgr Proteins in Epithelial Stem Cell Biology. *Development* 140 (12), 2484–2494. doi:10.1242/dev.083113

Berti, F. C. B., Mathias, C., Garcia, L. E., Gradia, D. F., de Araújo-Souza, P. S., Cipolla, G. A., et al. (2021). Comprehensive Analysis of ceRNA Networks in HPV16- and HPV18-Mediated Cervical Cancers Reveals XIST as a Pivotal Competing Endogenous RNA. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1867, 166172. doi:10.1016/j.bbadis.2021.166172

Carmon, K. S., Gong, X., Lin, Q., Thomas, A., and Liu, Q. (2011). R-spondins Function as Ligands of the Orphan Receptors LGR4 and LGR5 to Regulate Wnt/-catenin Signaling. *Proc. Natl. Acad. Sci.* 108 (28), 11452–11457. doi:10.1073/pnas.1106083108

Carmon, K. S., Gong, X., Yi, J., Thomas, A., and Liu, Q. (2014). RSPO-LGR4 Functions via IQGAP1 to Potentiate Wnt Signaling. *Proc. Natl. Acad. Sci. USA* 111 (13), E1221–E1229. doi:10.1073/pnas.1323106111

Carrillo-López, N., Martínez-Arias, L., Alonso-Montes, C., Martín-Carro, B., Martín-Vírgala, J., Ruiz-Ortega, M., et al. (2020). The Receptor Activator of Nuclear Factor κB Ligand Receptor Leucine-Rich Repeat-Containing G-Protein-Coupled Receptor 4 Contributes to Parathyroid Hormone-Induced Vascular Calcification. *Nephrol. Dial. Transpl.* 36, 618–631. doi:10.1093/ndt/gfaa290

Chauhan, G., Mehta, A., and Gupta, S. (2020). Stromal-AR Influences the Growth of Epithelial Cells in the Development of Benign Prostate Hyperplasia. *Mol. Cel Biochem* 471 (1-2), 129–142. doi:10.1007/s11010-020-03773-z

Chen, T., Qiao, X., Cheng, L., Liu, M., Deng, Y., and Zhuo, X. (2021a). LGR4 Silence Aggravates Ischemic Injury by Modulating Mitochondrial Function and Oxidative Stress via ERK Signaling Pathway in H9c2 Cells. *J. Mol. Histol.* 52, 363–371. doi:10.1007/s10735-021-09957-1

Chen, X., Chen, L., Tan, J., Zhang, L., Xia, J., Cheng, B., et al. (2021b). Rspo1-LGR4 axis in BMSCs Protects Bone against Radiation-induced Injury through the mTOR-dependent Autophagy Pathway. *J. Cel Physiol* 236, 4273–4289. doi:10.1002/jcp.30051

Cong, F., Wu, N., Tian, X., Fan, J., Liu, J., Song, T., et al. (2017). MicroRNA-34c Promotes Osteoclast Differentiation through Targeting LGR4. *Gene* 610, 1–8. doi:10.1016/j.gene.2017.01.028

Da Silva, F., Massa, F., Motamedi, F. J., Vidal, V., Rocha, A. S., Gregoire, E. P., et al. (2018). Myocardial-specific R-Spondin3 Drives Proliferation of the Coronary Stems Primarily through the Leucine Rich Repeat G Protein Coupled Receptor LGR4. *Dev. Biol.* 441 (1), 42–51. doi:10.1016/j.ydbio.2018.05.024

Dang, Y., Liu, B., Xu, P., Zhu, P., Zhai, Y., Liu, M., et al. (2014). Gpr48 Deficiency Induces Polycystic Kidney Lesions and Renal Fibrosis in Mice by Activating Wnt Signal Pathway. *PLoS One* 9 (3), e89835. doi:10.1371/journal.pone.0089835

den Hartogh, S. C., Wolstencroft, K., Mummery, C. L., and Passier, R. (2016). A Comprehensive Gene Expression Analysis at Sequential Stages of *In Vitro* Cardiac Differentiation from Isolated MESP1-Expressing-Mesoderm Progenitors. *Sci. Rep.* 6, 19386. doi:10.1038/srep19386

Deng, C., Reddy, P., Cheng, Y., Luo, C.-W., Hsiao, C.-L., and Hsueh, A. J. W. (2013). Multi-functional Norrin Is a Ligand for the LGR4 Receptor. *J. Cel Sci* 126 (Pt 9), 2060–2068. doi:10.1242/jcs.123471

Dong, Y., Zheng, Y., Xiao, J., Zhu, C., and Zhao, M. (2016). MicroRNA Let-7b Induces Lens Epithelial Cell Apoptosis by Targeting Leucine-Rich Repeat Containing G Protein-Coupled Receptor 4 (Lgr4) in Age-Related Cataract. *Exp. Eye Res.* 147, 98–104. doi:10.1016/j.exer.2016.04.018

Du, B., Luo, W., Li, R., Tan, B., Han, H., Lu, X., et al. (2013). Lgr4/Gpr48 Negatively Regulates TLR2/4-Associated Pattern Recognition and Innate Immunity by Targeting CD14 Expression. *J. Biol. Chem.* 288 (21), 15131–15141. doi:10.1074/jbc.M113.455535

Dubey, R., van Kerkhof, P., Jordens, I., Malinauskas, T., Pusapati, G. V., McKenna, J. K., et al. (2020). R-spondins Engage Heparan Sulfate Proteoglycans to Potentiate WNT Signaling. *Elife* 9. doi:10.7554/eLife.54469

Gao, Y., Shan, Z. Y., Wang, H., Zhang, H. M., and Teng, W. P. (2009). Inhibitory Effect of shRNA Targeting GPR48 on Invasion and Metastasis of Human Cervical Carcinoma Cell Line HeLa. *Ai Zheng* 28 (2), 104–107.

Gao, Y., Kitagawa, K., Hiramatsu, Y., Kikuchi, H., Isobe, T., Shimada, M., et al. (2006). Up-regulation of GPR48 Induced by Down-Regulation of p27Kip1 Enhances Carcinoma Cell Invasiveness and Metastasis. *Cancer Res.* 66 (24), 11623–11631. doi:10.1158/0008-5472.Can-06-2629

Ge, J. J., Lu, W. K., Zhu, N., Gao, X. J., Yang, F. Q., and Wu, X. Z. (2019). Effect of LGR4 on Synovial Cells and Inflammatory Factors in Rats with Traumatic Osteoarthritis. *Eur. Rev. Med. Pharmacol. Sci.* 23 (16), 6783–6790. doi:10.26355/eurrev_201908_18716

Geng, A., Wu, T., Cai, C., Song, W., Wang, J., Yu, Q. C., et al. (2020). A Novel Function of R-Spondin1 in Regulating Estrogen Receptor Expression Independent of Wnt/β-Catenin Signaling. *Elife* 9. doi:10.7554/eLife.56434

Glinka, A., Dolde, C., Kirsch, N., Huang, Y. L., Kazanskaya, O., Ingelfinger, D., et al. (2011). LGR4 and LGR5 Are R-spondin Receptors Mediating Wnt/β-catenin and Wnt/PCP Signalling. *EMBO Rep.* 12 (10), 1055–1061. doi:10.1038/embor.2011.175

Gong, X., Yi, J., Carmon, K. S., Crumbley, C. A., Xiong, W., Thomas, A., et al. (2015). Aberrant RSPO3-LGR4 Signaling in Keap1-Deficient Lung Adenocarcinomas Promotes Tumor Aggressiveness. *Oncogene* 34 (36), 4692–4701. doi:10.1038/onc.2014.417

Guan, X., Duan, Y., Zeng, Q., Pan, H., Qian, Y., Li, D., et al. (2014). Lgr4 Protein Deficiency Induces Ataxia-like Phenotype in Mice and Impairs Long Term Depression at Cerebellar Parallel Fiber-Purkinje Cell Synapses. *J. Biol. Chem.* 289 (38), 26492–26504. doi:10.1074/jbc.M114.564138

Hao, H.-X., Xie, Y., Zhang, Y., Charlat, O., Oster, E., Avello, M., et al. (2012). ZNRF3 Promotes Wnt Receptor Turnover in an R-Spondin-Sensitive Manner. *Nature* 485 (7397), 195–200. doi:10.1038/nature11019

Hoshii, T., Takeo, T., Nakagata, N., Takeya, M., Araki, K., and Yamamura, K.-i. (2007). LGR4 Regulates the Postnatal Development and Integrity of Male Reproductive Tracts in Mice1. *Biol. Reprod.* 76 (2), 303–313. doi:10.1095/biolreprod.106.054619

Hou, Q., Han, S., Yang, L., Chen, S., Chen, J., Ma, N., et al. (2019). The Interplay of MicroRNA-34a, LGR4, EMT-Associated Factors, and MMP2 in Regulating Uveal Melanoma Cells. *Invest. Ophthalmol. Vis. Sci.* 60 (13), 4503–4510. doi:10.1167/iovs.18-26477

Hou, Q., Zhou, L., Tang, J., Ma, N., Xu, A., Tang, J., et al. (2016). LGR4 Is a Direct Target of MicroRNA-34a and Modulates the Proliferation and Migration of Retinal Pigment Epithelial ARPE-19 Cells. *PLoS One* 11 (12), e0168320. doi:10.1371/journal.pone.0168320

Hsu, P.-J., Wu, F.-J., Kudo, M., Hsiao, C.-L., Hsueh, A. J. W., and Luo, C.-W. (2014). A Naturally Occurring Lgr4 Splice Variant Encodes a Soluble Antagonist Useful for Demonstrating the Gonadal Roles of Lgr4 in Mammals. *PLoS One* 9 (9), e106804. doi:10.1371/journal.pone.0106804

Huang, C.-K., Dai, D., Xie, H., Zhu, Z., Hu, J., Su, M., et al. (2020). Lgr4 Governs a Pro-inflammatory Program in Macrophages to Antagonize Post-Infarction Cardiac Repair. *Circ. Res.* 127 (8), 953–973. doi:10.1161/circresaha.119.315807

Jang, Y., Sohn, H. M., Ko, Y. J., Hyun, H., and Lim, W. (2021). Inhibition of RANKL-Induced Osteoclastogenesis by Novel Mutant RANKL. *Ijms* 22 (1), 434. doi:10.3390/ijms22010434

Jin, C., Yin, F., Lin, M., Li, H., Wang, Z., Weng, J., et al. (2008). GPR48 Regulates Epithelial Cell Proliferation and Migration by Activating EGFR during Eyelid Development. *Invest. Ophthalmol. Vis. Sci.* 49 (10), 4245–4253. doi:10.1167/iovs.08-1860

Jin, Y.-R., Han, X. H., Nishimori, K., Ben-Avraham, D., Oh, Y. J., Shim, J.-w., et al. (2020). Canonical WNT/β-Catenin Signaling Activated by WNT9b and RSPO2 Cooperation Regulates Facial Morphogenesis in Mice. *Front. Cel Dev. Biol.* 8, 264. doi:10.3389/fcell.2020.00264

Kajava, A. V. (1998). Structural Diversity of Leucine-Rich Repeat Proteins 1 1Edited by F. Cohen. *J. Mol. Biol.* 277 (3), 519–527. doi:10.1006/jmbi.1998.1643

Kang, Y. E., Kim, J.-M., Kim, K. S., Chang, J. Y., Jung, M., Lee, J., et al. (2017). Upregulation of RSPO2-Gpr48/lgr4 Signaling in Papillary Thyroid Carcinoma Contributes to Tumor Progression. *Oncotarget* 8 (70), 114980–114994. doi:10.18632/oncotarget.22692

Kato, S., Matsubara, M., Matsuo, T., Mohri, Y., Kazama, I., Hatano, R., et al. (2006). Leucine-rich Repeat-Containing G Protein-Coupled Receptor-4 (LGR4, Gpr48) Is Essential for Renal Development in Mice. *Nephron Exp. Nephrol.* 104 (2), e63–e75. doi:10.1159/000093999

Kato, S., Mohri, Y., Matsuo, T., Ogawa, E., Umezawa, A., Okuyama, R., et al. (2007). Eye-open at Birth Phenotype with Reduced Keratinocyte Motility in LGR4 Null Mice. *FEBS Lett.* 581 (24), 4685–4690. doi:10.1016/j.febslet.2007.08.064

Kida, T., Oyama, K., Sone, M., Koizumi, M., Hidema, S., and Nishimori, K. (2014). Lgr4 Is Required for Endometrial Receptivity Acquired through Ovarian Hormone Signaling. *Biosci. Biotechnol. Biochem.* 78 (11), 1813–1816. doi:10.1080/09168451.2014.936353

Kim, H. S., Kim, Y. S., Lee, C., Shin, M. S., Kim, J. W., and Jang, B. G. (2019). Expression Profile of Sonic Hedgehog Signaling-Related Molecules in Basal Cell Carcinoma. *PLoS One* 14 (11), e0225511. doi:10.1371/journal.pone.0225511

Knight, M. N., and Hankenson, K. D. (2014). R-spondins: Novel Matricellular Regulators of the Skeleton. *Matrix Biol.* 37, 157–161. doi:10.1016/j.matbio.2014.06.003

Ko, Y. J., Sohn, H. M., Jang, Y., Park, M., Kim, B., Kim, B., et al. (2021). A Novel Modified RANKL Variant Can Prevent Osteoporosis by Acting as a Vaccine and an Inhibitor. *Clin. Translational Med.* 11 (3), e368. doi:10.1002/ctm2.368

Kobe, B., and Deisenhofer, J. (1995). A Structural Basis of the Interactions between Leucine-Rich Repeats and Protein Ligands. *Nature* 374 (6518), 183–186. doi:10.1038/374183a0

Kobe, B., and Deisenhofer, J. (1993). Crystal Structure of Porcine Ribonuclease Inhibitor, a Protein with Leucine-Rich Repeats. *Nature* 366 (6457), 751–756. doi:10.1038/366751a0

Koizumi, M., Oyama, K., Yamakami, Y., Kida, T., Satoh, R., Kato, S., et al. (2015). Lgr4 Controls Specialization of Female Gonads in Mice1. *Biol. Reprod.* 93 (4), 90. doi:10.1095/biolreprod.114.123638

Koo, B.-K., Spit, M., Jordens, I., Low, T. Y., Stange, D. E., van de Wetering, M., et al. (2012). Tumour Suppressor RNF43 Is a Stem-Cell E3 Ligase that Induces Endocytosis of Wnt Receptors. *Nature* 488 (7413), 665–669. doi:10.1038/nature11308

Lai, S., Cheng, R., Gao, D., Chen, Y. G., and Deng, C. (2020). LGR5 Constitutively Activates NF-κB Signaling to Regulate the Growth of Intestinal Crypts. *FASEB j.* 34 (11), 15605–15620. doi:10.1096/fj.202001329R

Li, B., Yao, Q., Guo, S., Ma, S., Dong, Y., Xin, H., et al. (2019a). Type 2 Diabetes with Hypertensive Patients Results in Changes to Features of Adipocytokines: Leptin, Irisin, LGR4, and Sfrp5. *Clin. Exp. Hypertens.* 41 (7), 645–650. doi:10.1080/10641963.2018.1529779

Li, C., Huang, Q., Yang, R., Guo, X., Dai, Y., Zeng, J., et al. (2020a). Targeted Next Generation Sequencing of Nine Osteoporosis-Related Genes in the Wnt Signaling Pathway Among Chinese Postmenopausal Women. *Endocrine* 68 (3), 669–678. doi:10.1007/s12020-020-02248-x

Li, J.-Y., Chai, B., Zhang, W., Fritze, D. M., Zhang, C., and Mulholland, M. W. (2014). LGR4 and its Ligands, R-Spondin 1 and R-Spondin 3, Regulate Food Intake in the Hypothalamus of Male Rats. *Endocrinology* 155 (2), 429–440. doi:10.1210/en.2013-1550

Li, R., Liu, X., Zhou, X. J., Chen, X., Li, J. P., Yin, Y. H., et al. (2020b). Identification and Validation of the Prognostic Value of Immune-Related Genes in Non-small Cell Lung Cancer. *Am. J. Transl Res.* 12 (9), 5844–5865.

Li, X.-Y., Lu, Y., Sun, H.-Y., Wang, J.-Q., Yang, J., Zhang, H.-J., et al. (2010). G Protein-Coupled Receptor 48 Upregulates Estrogen Receptor α Expression via

cAMP/PKA Signaling in the Male Reproductive Tract. *Development* 137 (1), 151–157. doi:10.1242/dev.040659

Li, Z., Liu, S., Lou, J., Mulholland, M., and Zhang, W. (2019b). LGR4 Protects Hepatocytes from Injury in Mouse. *Am. J. Physiology-Gastrointestinal Liver Physiol.* 316 (1), G123–g131. doi:10.1152/ajpgi.00056.2018

Liang, F., Yue, J., Wang, J., Zhang, L., Fan, R., Zhang, H., et al. (2015). GPCR48/LGR4 Promotes Tumorigenesis of Prostate Cancer via PI3K/Akt Signaling Pathway. *Med. Oncol.* 32 (3), 49. doi:10.1007/s12032-015-0486-1

Liang, F., Zhang, H., Cheng, D., Gao, H., Wang, J., Yue, J., et al. (2021). Ablation of LGR4 Signaling Enhances Radiation Sensitivity of Prostate Cancer Cells. *Life Sci.* 265, 118737. doi:10.1016/j.lfs.2020.118737

Liu, H.-L., Yeh, I.-J., Phan, N. N., Wu, Y.-H., Yen, M.-C., Hung, J.-H., et al. (2020a). Gene Signatures of SARS-CoV/SARS-CoV-2-Infected Ferret Lungs in Short- and Long-Term Models. *Infect. Genet. Evol.* 85, 104438. doi:10.1016/j.meegid.2020.104438

Liu, J., Wei, W., Guo, C.-A., Han, N., Pan, J.-f., Fei, T., et al. (2013a). Stat3 Upregulates Leucine-Rich Repeat-Containing G Protein-Coupled Receptor 4 Expression in Osteosarcoma Cells. *Biomed. Res. Int.* 2013, 1–8. doi:10.1155/2013/310691

Liu, S., Gao, Y., Zhang, L., Yin, Y., and Zhang, W. (2020b). Rspo1/Rspo3-LGR4 Signaling Inhibits Hepatic Cholesterol Synthesis through the AMPKα-SREBP2 Pathway. *FASEB j.* 34 (11), 14946–14959. doi:10.1096/fj.202001234R

Liu, S., Qian, Y., Li, L., Wei, G., Guan, Y., Pan, H., et al. (2013b). Lgr4 Gene Deficiency Increases Susceptibility and Severity of Dextran Sodium Sulfate-Induced Inflammatory Bowel Disease in Mice. *J. Biol. Chem.* 288 (13), 8794–8803. doi:10.1074/jbc.M112.436204

Liu, S., Yin, Y., Yu, R., Li, Y., and Zhang, W. (2018). R-spondin3-LGR4 Signaling Protects Hepatocytes against DMOG-Induced Hypoxia/reoxygenation Injury through Activating β-catenin. *Biochem. Biophysical Res. Commun.* 499 (1), 59–65. doi:10.1016/j.bbrc.2018.03.126

Liu, X., and Xu, X. (2018). MicroRNA-137 Dysregulation Predisposes to Osteoporotic Fracture by Impeding ALP Activity and Expression via Suppression of Leucine-Rich Repeat-Containing G-Protein-Coupled Receptor 4 Expression. *Int. J. Mol. Med.* 42 (2), 1026–1033. doi:10.3892/ijmm.2018.3690

Luo, C. W., and Hsueh, A. J. (2006). Genomic Analyses of the Evolution of LGR Genes. *Chang Gung Med. J.* 29 (1), 2–8.

Luo, J., Yang, Z., Ma, Y., Yue, Z., Lin, H., Qu, G., et al. (2016). LGR4 Is a Receptor for RANKL and Negatively Regulates Osteoclast Differentiation and Bone Resorption. *Nat. Med.* 22 (5), 539–546. doi:10.1038/nm.4076

Luo, J., Zhou, W., Zhou, X., Li, D., Weng, J., Yi, Z., et al. (2009). Regulation of Bone Formation and Remodeling by G-Protein-Coupled Receptor 48. *Development* 136 (16), 2747–2756. doi:10.1242/dev.033571

Luo, W., Rodriguez, M., Valdez, J. M., Zhu, X., Tan, K., Li, D., et al. (2013). Lgr4 Is a Key Regulator of Prostate Development and Prostate Stem Cell Differentiation. *Stem Cells* 31 (11), 2492–2505. doi:10.1002/stem.1484

Luo, W., Tan, P., Rodriguez, M., He, L., Tan, K., Zeng, L., et al. (2017). Leucine-rich Repeat-Containing G Protein-Coupled Receptor 4 (Lgr4) Is Necessary for Prostate Cancer Metastasis via Epithelial-Mesenchymal Transition. *J. Biol. Chem.* 292 (37), 15525–15537. doi:10.1074/jbc.M116.771931

Ma, X., Yin, B., Guo, S., Umar, T., Liu, J., Wu, Z., et al. (2021). Enhanced Expression of miR-34a Enhances *Escherichia coli* Lipopolysaccharide-Mediated Endometritis by Targeting LGR4 to Activate the NF-Kb Pathway. *Oxidative Med. Cell Longevity* 2021, 1–18. doi:10.1155/2021/1744754

Mancini, A., Howard, S. R., Marelli, F., Cabrera, C. P., Barnes, M. R., Sternberg, M. J. E., et al. (2020). LGR4 Deficiency Results in Delayed Puberty through Impaired Wnt/β-Catenin Signaling. *JCI Insight* 5 (11). doi:10.1172/jci.insight.133434

Mantovani, A., Sica, A., Sozzani, S., Allavena, P., Vecchi, A., and Locati, M. (2004). The Chemokine System in Diverse Forms of Macrophage Activation and Polarization. *Trends Immunol.* 25 (12), 677–686. doi:10.1016/j.it.2004.09.015

Matsuike, R., Tanaka, H., Nakai, K., Kanda, M., Nagasaki, M., Murakami, F., et al. (2018). Continuous Application of Compressive Force Induces Fusion of Osteoclast-like RAW264.7 Cells via Upregulation of RANK and Downregulation of LGR4. *Life Sci.* 201, 30–36. doi:10.1016/j.lfs.2018.03.038

Mazerbourg, S., Bouley, D. M., Sudo, S., Klein, C. A., Zhang, J. V., Kawamura, K., et al. (2004). Leucine-rich Repeat-Containing, G Protein-Coupled Receptor 4 Null Mice Exhibit Intrauterine Growth Retardation Associated with Embryonic and Perinatal Lethality. *Mol. Endocrinol.* 18 (9), 2241–2254. doi:10.1210/me.2004-0133

McDonald, T., Wang, R., Bailey, W., Xie, G., Chen, F., Caskey, C. T., et al. (1998). Identification and Cloning of an Orphan G Protein-Coupled Receptor of the Glycoprotein Hormone Receptor Subfamily. *Biochem. Biophysical Res. Commun.* 247 (2), 266–270. doi:10.1006/bbrc.1998.8774

Mendive, F., Laurent, P., Van Schoore, G., Skarnes, W., Pochet, R., and Vassart, G. (2006). Defective Postnatal Development of the Male Reproductive Tract in LGR4 Knockout Mice. *Dev. Biol.* 290 (2), 421–434. doi:10.1016/j.ydbio.2005.11.043

Mustata, R. C., Van Loy, T., Lefort, A., Libert, F., Strollo, S., Vassart, G., et al. (2011). Lgr4 Is Required for Paneth Cell Differentiation and Maintenance of Intestinal Stem Cells *Ex Vivo*. *EMBO Rep.* 12 (6), 558–564. doi:10.1038/embor.2011.52

Naka, K., Ochiai, R., Matsubara, E., Kondo, C., Yang, K.-M., Hoshii, T., et al. (2020). The Lysophospholipase D Enzyme Gdpd3 Is Required to Maintain Chronic Myelogenous Leukaemia Stem Cells. *Nat. Commun.* 11 (1), 4681. doi:10.1038/s41467-020-18491-9

Oliveira, V., Schaefer, J., Calder, M., Lydon, J. P., DeMayo, F. J., Bhattacharya, M., et al. (2019). Uterine Gα Q/11 Signaling, in a Progesterone-dependent Manner, Critically Regulates the Acquisition of Uterine Receptivity in the Female Mouse. *FASEB j.* 33 (8), 9374–9387. doi:10.1096/fj.201900026R

Otsuka, A., Jinguji, A., Maejima, Y., Kasahara, Y., Shimomura, K., Hidema, S., et al. (2019). LGR4 Is Essential for R-Spondin1-Mediated Suppression of Food Intake via Pro-opiomelanocortin. *Biosci. Biotechnol. Biochem.* 83 (7), 1336–1342. doi:10.1080/09168451.2019.1591266

Pan, H., Cui, H., Liu, S., Qian, Y., Wu, H., Li, L., et al. (2014). Lgr4 Gene Regulates Corpus Luteum Maturation through Modulation of the WNT-Mediated EGFR-ERK Signaling Pathway. *Endocrinology* 155 (9), 3624–3637. doi:10.1210/en.2013-2183

Pardoll, D. M. (2012). The Blockade of Immune Checkpoints in Cancer Immunotherapy. *Nat. Rev. Cancer* 12 (4), 252–264. doi:10.1038/nrc3239

Park, S., Cui, J., Yu, W., Wu, L., Carmon, K. S., and Liu, Q. J. (2018). Differential Activities and Mechanisms of the Four R-Spondins in Potentiating Wnt/β-Catenin Signaling. *J. Biol. Chem.* 293 (25), 9759–9769. doi:10.1074/jbc.RA118.002743

Park, S., Wu, L., Tu, J., Yu, W., Toh, Y., Carmon, K. S., et al. (2020). Unlike LGR4, LGR5 Potentiates Wnt-β-Catenin Signaling without Sequestering E3 Ligases. *Sci. Signal.* 13 (660). doi:10.1126/scisignal.aaz4051

Pawaputanon Na Mahasarakham, C., Izu, Y., Nishimori, K., Izumi, Y., Noda, M., and Ezura, Y. (2017). Lgr4 Expression in Osteoblastic Cells Is Suppressed by Hydrogen Peroxide Treatment. *J. Cel. Physiol.* 232 (7), 1761–1766. doi:10.1002/jcp.25684

Planas-Paz, L., Orsini, V., Boulter, L., Calabrese, D., Pikiolek, M., Nigsch, F., et al. (2016). The RSPO-Lgr4/5-Znrf3/rnf43 Module Controls Liver Zonation and Size. *Nat. Cel Biol* 18 (5), 467–479. doi:10.1038/ncb3337

Planas-Paz, L., Sun, T., Pikiolek, M., Cochran, N. R., Bergling, S., Orsini, V., et al. (2019). YAP, but Not RSPO-Lgr4/5, Signaling in Biliary Epithelial Cells Promotes a Ductular Reaction in Response to Liver Injury. *Cell Stem Cell* 25 (1), 39–53. doi:10.1016/j.stem.2019.04.005

Ren, X., Xia, W., Xu, P., Shen, H., Dai, X., Liu, M., et al. (2020). Lgr4 Deletion Delays the Hair Cycle and Inhibits the Activation of Hair Follicle Stem Cells. *J. Invest. Dermatol.* 140 (9), 1706e1704–1712. doi:10.1016/j.jid.2019.12.034

Salik, B., Yi, H., Hassan, N., Santiappillai, N., Vick, B., Connerty, P., et al. (2020). Targeting RSPO3-LGR4 Signaling for Leukemia Stem Cell Eradication in Acute Myeloid Leukemia. *Cancer Cell* 38 (2), 263–278. doi:10.1016/j.ccell.2020.05.014

Schoore, G. V., Mendive, F., Pochet, R., and Vassart, G. (2005). Expression Pattern of the Orphan Receptor LGR4/GPR48 Gene in the Mouse. *Histochem. Cel Biol* 124 (1), 35–50. doi:10.1007/s00418-005-0002-3

Shi, G.-X., Zheng, X.-F., Zhu, C., Li, B., Wang, Y.-R., Jiang, S.-D., et al. (2017). Evidence of the Role of R-Spondin 1 and its Receptor Lgr4 in the Transmission of Mechanical Stimuli to Biological Signals for Bone Formation. *Ijms* 18 (3), 564. doi:10.3390/ijms18030564

Shi, S.-q., Li, S.-s., Zhang, X.-y., Wei, Z., Fu, W.-z., He, J.-w., et al. (2021). LGR4 Gene Polymorphisms Are Associated with Bone and Obesity Phenotypes in Chinese Female Nuclear Families. *Front. Endocrinol.* 12, 656077. doi:10.3389/fendo.2021.656077

Singla, B., Lin, H.-P., Chen, A., Ahn, W., Ghoshal, P., Cherian-Shaw, M., et al. (2020). Role of R-Spondin 2 in Arterial Lymphangiogenesis and Atherosclerosis. *Cardiovasc. Res.* 117, 1489–1509. doi:10.1093/cvr/cvaa244

Sone, M., Oyama, K., Mohri, Y., Hayashi, R., Clevers, H., and Nishimori, K. (2013). LGR4 Expressed in Uterine Epithelium Is Necessary for Uterine Gland Development and Contributes to Decidualization in Mice. *FASEB j.* 27 (12), 4917–4928. doi:10.1096/fj.13-232215

Song, H., Luo, J., Luo, W., Weng, J., Wang, Z., Li, B., et al. (2008). Inactivation of G-Protein-Coupled Receptor 48 (Gpr48/Lgr4) Impairs Definitive Erythropoiesis at Midgestation through Down-Regulation of the ATF4 Signaling Pathway. *J. Biol. Chem.* 283 (52), 36687–36697. doi:10.1074/jbc.M800721200

Souza, S. M. d., Valiente, A. E. F., Sá, K. M., Juanes, C. d. C., Rodrigues, B. J., Farias, A. C. C., et al. (2019). Immunoexpression of LGR4 and B-Catenin in Gastric Cancer and Normal Gastric Mucosa. *Asian Pac. J. Cancer Prev.* 20 (2), 519–527. doi:10.31557/apjcp.2019.20.2.519

Styrkarsdottir, U., Thorleifsson, G., Sulem, P., Gudbjartsson, D. F., Sigurdsson, A., Jonasdottir, A., et al. (2013). Nonsense Mutation in the LGR4 Gene Is Associated with Several Human Diseases and Other Traits. *Nature* 497 (7450), 517–520. doi:10.1038/nature12124

Su, X., Zhou, G., Tian, M., Wu, S., and Wang, Y. (2021). Silencing of RSPO1 Mitigates Obesity-Related Renal Fibrosis in Mice by Deactivating Wnt/β-Catenin Pathway. *Exp. Cel Res.* 405 (2), 112713. doi:10.1016/j.yexcr.2021.112713

Sun, L., Zhang, Z., Yao, Y., Li, W.-Y., and Gu, J. (2020). Analysis of Expression Differences of Immune Genes in Non-small Cell Lung Cancer Based on TCGA and ImmPort Data Sets and the Application of a Prognostic Model. *Ann. Transl Med.* 8 (8), 550. doi:10.21037/atm.2020.04.38

Sun, P., Jia, K., Zheng, C., Zhu, X., Li, J., He, L., et al. (2019). Loss of Lgr4 Inhibits Differentiation, Migration and Apoptosis, and Promotes Proliferation in Bone Mesenchymal Stem Cells. *J. Cel Physiol* 234 (7), 10855–10867. doi:10.1002/jcp.27927

Sun, Y., Hong, J., Chen, M., Ke, Y., Zhao, S., Liu, W., et al. (2015). Ablation of Lgr4 Enhances Energy Adaptation in Skeletal Muscle via Activation of Ampk/Sirt1/Pgc1α Pathway. *Biochem. Biophysical Res. Commun.* 464 (2), 396–400. doi:10.1016/j.bbrc.2015.06.066

Szenker-Ravi, E., Altunoglu, U., Leushacke, M., Bosso-Lefèvre, C., Khatoo, M., Thi Tran, H., et al. (2018). RSPO2 Inhibition of RNF43 and ZNRF3 Governs Limb Development Independently of LGR4/5/6. *Nature* 557 (7706), 564–569. doi:10.1038/s41586-018-0118-y

Tan, B., Shi, X., Zhang, J., Qin, J., Zhang, N., Ren, H., et al. (2018). Inhibition of Rspo-Lgr4 Facilitates Checkpoint Blockade Therapy by Switching Macrophage Polarization. *Cancer Res.* 78 (17), 4929–4942. doi:10.1158/0008-5472.Can-18-0152

Tan, X., Zhang, L., Li, T., Zhan, J., Qiao, K., Wu, H., et al. (2021). Lgr4 Regulates Oviductal Epithelial Secretion through the WNT Signaling Pathway. *Front Cel Dev Biol* 9, 666303. doi:10.3389/fcell.2021.666303

van Andel, H., Ren, Z., Koopmans, I., Joosten, S. P. J., Kocemba, K. A., de Lau, W., et al. (2017). Aberrantly Expressed LGR4 Empowers Wnt Signaling in Multiple Myeloma by Hijacking Osteoblast-Derived R-Spondins. *Proc. Natl. Acad. Sci. USA* 114 (2), 376–381. doi:10.1073/pnas.1618650114

Van Loy, T., Vandersmissen, H. P., Van Hiel, M. B., Poels, J., Verlinden, H., Badisco, L., et al. (2008). Comparative Genomics of Leucine-Rich Repeats Containing G Protein-Coupled Receptors and Their Ligands. *Gen. Comp. Endocrinol.* 155 (1), 14–21. doi:10.1016/j.ygcen.2007.06.022

Vidal, V. P., Jian-Motamedi, F., Rekima, S., Gregoire, E. P., Szenker-Ravi, E., Leushacke, M., et al. (2020). R-spondin Signalling Is Essential for the Maintenance and Differentiation of Mouse Nephron Progenitors. *Elife* 9. doi:10.7554/eLife.53895

Wang, D., Huang, B., Zhang, S., Yu, X., Wu, W., and Wang, X. (2013a). Structural Basis for R-Spondin Recognition by LGR4/5/6 Receptors. *Genes Dev.* 27 (12), 1339–1344. doi:10.1101/gad.219360.113

Wang, F., Wang, X., Ma, K., Zhang, C., Chang, J., and Fu, X. (2020a). Akermanite Bioceramic Enhances Wound Healing with Accelerated Reepithelialization by Promoting Proliferation, Migration, and Stemness of Epidermal Cells. *Wound Rep. Reg.* 28 (1), 16–25. doi:10.1111/wrr.12742

Wang, J., Li, X., Ke, Y., Lu, Y., Wang, F., Fan, N., et al. (2012). GPR48 Increases Mineralocorticoid Receptor Gene Expression. *Jasn* 23 (2), 281–293. doi:10.1681/asn.2011040351

Wang, J., Liu, R., Wang, F., Hong, J., Li, X., Chen, M., et al. (2013b). Ablation of LGR4 Promotes Energy Expenditure by Driving white-to-brown Fat Switch. *Nat. Cel Biol* 15 (12), 1455–1463. doi:10.1038/ncb2867

Wang, W., Chen, J., Hui, Y., Huang, M., and Yuan, P. (2018). Down-regulation of miR-193a-3p Promotes Osteoblast Differentiation through Up-Regulation of LGR4/ATF4 Signaling. *Biochem. Biophysical Res. Commun.* 503 (3), 2186–2193. doi:10.1016/j.bbrc.2018.08.011

Wang, Y., Dong, J., Li, D., Lai, L., Siwko, S., Li, Y., et al. (2013c). Lgr4 Regulates Mammary Gland Development and Stem Cell Activity through the Pluripotency Transcription Factor Sox2. *Stem Cells* 31 (9), 1921–1931. doi:10.1002/stem.1438

Wang, Y., Wang, H., Guo, J., Gao, J., Wang, M., Xia, M., et al. (2020b). LGR4, Not LGR5, Enhances hPSC Hematopoiesis by Facilitating Mesoderm Induction via TGF-Beta Signaling Activation. *Cel Rep.* 31 (5), 107600. doi:10.1016/j.celrep.2020.107600

Wang, Z., Jin, C., Li, H., Li, C., Hou, Q., Liu, M., et al. (2010). GPR48-Induced Keratinocyte Proliferation Occurs through HB-EGF Mediated EGFR Transactivation. *FEBS Lett.* 584 (18), 4057–4062. doi:10.1016/j.febslet.2010.08.028

Wang, Z., Yin, P., Sun, Y., Na, L., Gao, J., Wang, W., et al. (2020c). LGR4 Maintains HGSOC Cell Epithelial Phenotype and Stem-like Traits. *Gynecol. Oncol.* 159 (3), 839–849. doi:10.1016/j.ygyno.2020.09.020

Weng, J., Luo, J., Cheng, X., Jin, C., Zhou, X., Qu, J., et al. (2008). Deletion of G Protein-Coupled Receptor 48 Leads to Ocular Anterior Segment Dysgenesis (ASD) through Down-Regulation of Pitx2. *Proc. Natl. Acad. Sci.* 105 (16), 6081–6086. doi:10.1073/pnas.0708257105

Wu, J., Li, X., Li, D., Ren, X., Li, Y., Herter, E. K., et al. (2020). MicroRNA-34 Family Enhances Wound Inflammation by Targeting LGR4. *J. Invest. Dermatol.* 140 (2), 465–476. doi:10.1016/j.jid.2019.07.694

Wu, J., Xie, N., Xie, K., Zeng, J., Cheng, L., Lei, Y., et al. (2013). GPR48, a Poor Prognostic Factor, Promotes Tumor Metastasis and Activates β-catenin/TCF Signaling in Colorectal Cancer. *Carcinogenesis* 34 (12), 2861–2869. doi:10.1093/carcin/bgt229

Xu, K., Xu, Y., Rajashankar, K. R., Robev, D., and Nikolov, D. B. (2013). Crystal Structures of Lgr4 and its Complex with R-Spondin1. *Structure* 21 (9), 1683–1689. doi:10.1016/j.str.2013.07.001

Xu, P., Dang, Y., Wang, L., Liu, X., Ren, X., Gu, J., et al. (2016). Lgr4 Is Crucial for Skin Carcinogenesis by Regulating MEK/ERK and Wnt/β-Catenin Signaling Pathways. *Cancer Lett.* 383 (2), 161–170. doi:10.1016/j.canlet.2016.09.005

Yamakami, Y., Kohashi, K., Oyama, K., Mohri, Y., Hidema, S., and Nishimori, K. (2016). LGR4 Is Required for Sequential Molar Development. *Biochem. Biophys. Rep.* 8, 174–183. doi:10.1016/j.bbrep.2016.08.018

Yang, D., Li, J.-s., Xu, Q.-y., Xia, T., and Xia, J.-h. (2018). Inhibitory Effect of MiR-449b on Cancer Cell Growth and Invasion through LGR4 in Non-small-cell Lung Carcinoma. *Curr. Med. Sci.* 38 (4), 582–589. doi:10.1007/s11596-018-1917-y

Yang, Y. y., Zhou, Y. m., Xu, J. z., Sun, L. h., Tao, B., Wang, W. q., et al. (2021). Lgr4 Promotes Aerobic Glycolysis and Differentiation in Osteoblasts via the Canonical Wnt/β -catenin Pathway. *J. Bone Miner Res.* 36, 1605–1620. doi:10.1002/jbmr.4321

Yeom, E., Shin, H., Yoo, W., Jun, E., Kim, S., Hong, S. H., et al. (2021). Tumour-derived Dilp8/INSL3 Induces Cancer Anorexia by Regulating Feeding Neuropeptides via Lgr3/8 in the Brain. *Nat. Cel Biol* 23 (2), 172–183. doi:10.1038/s41556-020-00628-z

Yi, J., Xiong, W., Gong, X., Bellister, S., Ellis, L. M., and Liu, Q. (2013). Analysis of LGR4 Receptor Distribution in Human and Mouse Tissues. *PLoS One* 8 (10), e78144. doi:10.1371/journal.pone.0078144

Yi, T., Weng, J., Siwko, S., Luo, J., Li, D., and Liu, M. (2014). LGR4/GPR48 Inactivation Leads to Aniridia-Genitourinary Anomalies-Mental Retardation Syndrome Defects. *J. Biol. Chem.* 289 (13), 8767–8780. doi:10.1074/jbc.M113.530816

Yin, B., Umar, T., Ma, X., Chen, Y., Chen, N., Wu, Z., et al. (2021). MiR-193a-3p Targets LGR4 to Promote the Inflammatory Response in Endometritis. *Int. Immunopharmacology* 98, 107718. doi:10.1016/j.intimp.2021.107718

Yoshino, O., Ono, Y., Honda, M., Hattori, K., Sato, E., Hiraoka, T., et al. (2020). Relaxin-2 May Suppress Endometriosis by Reducing Fibrosis, Scar Formation, and Inflammation. *Biomedicines* 8 (11), 467. doi:10.3390/biomedicines8110467

Yu, C.-Y., Liang, G.-B., Du, P., and Liu, Y.-H. (2013). Lgr4 Promotes Glioma Cell Proliferation through Activation of Wnt Signaling. *Asian Pac. J. Cancer Prev.* 14 (8), 4907–4911. doi:10.7314/apjcp.2013.14.8.4907

Yu, W.-j., Zhang, Z., Fu, W.-z., He, J.-w., Wang, C., and Zhang, Z.-L. (2020). Association between LGR4 Polymorphisms and Peak Bone mineral Density

and Body Composition. *J. Bone Miner Metab.* 38 (5), 658–669. doi:10.1007/s00774-020-01106-0

Yue, F., Jiang, W., Ku, A. T., Young, A. I. J., Zhang, W., Souto, E. P., et al. (2021). A Wnt-independent LGR4-EGFR Signaling Axis in Cancer Metastasis. *Cancer Res.* 81 (17), 4441–4454. doi:10.1158/0008-5472.Can-21-1112

Yue, Z., Yuan, Z., Zeng, L., Wang, Y., Lai, L., Li, J., et al. (2018). LGR4 Modulates Breast Cancer Initiation, Metastasis, and Cancer Stem Cells. *FASEB j.* 32 (5), 2422–2437. doi:10.1096/fj.201700897R

Żak, M., van Oort, T., Hendriksen, F. G., Garcia, M.-I., Vassart, G., and Grolman, W. (2016). LGR4 and LGR5 Regulate Hair Cell Differentiation in the Sensory Epithelium of the Developing Mouse Cochlea. *Front. Cel. Neurosci.* 10, 186. doi:10.3389/fncel.2016.00186

Zeng, Z., Ji, N., Yi, J., Lv, J., Yuan, J., Lin, Z., et al. (2020). LGR4 Overexpression Is Associated with Clinical Parameters and Poor Prognosis of Serous Ovarian Cancer. *Cbm* 28 (1), 65–72. doi:10.3233/cbm-191145

Zhang, H., Liang, F., Yue, J., Liu, P., Wang, J., Wang, Z., et al. (2020). MicroRNA-137 R-egulates H-ypoxia-mediated M-igration and E-pithelial-mesenchymal T-ransition in P-rostate C-ancer by T-argeting LGR4 via the EGFR/ERK S-ignaling P-athway. *Int. J. Oncol.* 57 (2), 540–549. doi:10.3892/ijo.2020.5064

Zhang, J., Li, Q., Zhang, S., Xu, Q., and Wang, T. (2016). Lgr4 Promotes Prostate Tumorigenesis through the Jmjd2a/AR Signaling Pathway. *Exp. Cel Res.* 349 (1), 77–84. doi:10.1016/j.yexcr.2016.09.023

Zhang, L., Song, Y., Ling, Z., Li, Y., Ren, X., Yang, J., et al. (2019). R-spondin 2-LGR4 System Regulates Growth, Migration and Invasion, Epithelial-Mesenchymal Transition and Stem-like Properties of Tongue Squamous Cell Carcinoma via Wnt/β-Catenin Signaling. *EBioMedicine* 44, 275–288. doi:10.1016/j.ebiom.2019.03.076

Zhang, M., Zhang, P., Liu, Y., Lv, L., Zhang, X., Liu, H., et al. (2017a). RSPO3-LGR4 Regulates Osteogenic Differentiation of Human Adipose-Derived Stem Cells via ERK/FGF Signalling. *Sci. Rep.* 7, 42841. doi:10.1038/srep42841

Zhang, N., Huang, H., Tan, B., Wei, Y., Xiong, Q., Yan, Y., et al. (2017b). Leucine-rich Repeat-Containing G Protein-Coupled Receptor 4 Facilitates Vesicular Stomatitis Virus Infection by Binding Vesicular Stomatitis Virus Glycoprotein. *J. Biol. Chem.* 292 (40), 16527–16538. doi:10.1074/jbc.M117.802090

Zhang, X., Deng, H. W., Shen, H., and Ehrlich, M. (2021). Prioritization of Osteoporosis-Associated Genome-wide Association Study (GWAS) Single-Nucleotide Polymorphisms (SNPs) Using Epigenomics and Transcriptomics. *JBMR Plus* 5 (5), e10481. doi:10.1002/jbm4.10481

Zhi, X., Zhang, J., Cheng, Z., Bian, L., and Qin, J. (2019). Withdrawn: circLgr4 Drives Colorectal Tumorigenesis and Invasion through Lgr4-targeting Peptide. *Int. J. Cancer*. doi:10.1002/ijc.32549

Zhou, M., Guo, S., Yuan, L., Zhang, Y., Zhang, M., Chen, H., et al. (2017). Blockade of LGR4 Inhibits Proliferation and Odonto/osteogenic Differentiation of Stem Cells from Apical Papillae. *J. Mol. Hist.* 48 (5-6), 389–401. doi:10.1007/s10735-017-9737-0

Zhu, J., Hou, Q., Dong, X. D., Wang, Z., Chen, X., Zheng, D., et al. (2015). Targeted Deletion of the Murine Lgr4 Gene Decreases Lens Epithelial Cell Resistance to Oxidative Stress and Induces Age-Related Cataract Formation. *PLoS One* 10 (3), e0119599. doi:10.1371/journal.pone.0119599

Zhu, P., Wu, J., Wang, Y., Zhu, X., Lu, T., Liu, B., et al. (2018). LncGata6 Maintains Stemness of Intestinal Stem Cells and Promotes Intestinal Tumorigenesis. *Nat. Cel Biol* 20 (10), 1134–1144. doi:10.1038/s41556-018-0194-0

Zhu, Y.-B., Xu, L., Chen, M., Ma, H.-N., and Lou, F. (2013). GPR48 Promotes Multiple Cancer Cell Proliferation via Activation of Wnt Signaling. *Asian Pac. J. Cancer Prev.* 14 (8), 4775–4778. doi:10.7314/apjcp.2013.14.8.4775

Zou, Y., Ning, T., Shi, J., Chen, M., Ding, L., Huang, Y., et al. (2017). Association of a Gain-Of-Function Variant in LGR4 with central Obesity. *Obesity* 25 (1), 252–260. doi:10.1002/oby.21704

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Yang, Wang, Gong, Fan, Yang, Cui, Gao, Li, Sun, Li and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 February 2022
doi: 10.3389/fgene.2022.805651

Estimates of Variance Components and Heritability Using Random Regression Models for Semen Traits in Boars

Yifeng Hong [1,2], Limin Yan [1], Xiaoyan He [1,2], Dan Wu [1,2], Jian Ye [1,2], Gengyuan Cai [1,2], Dewu Liu [2], Zhenfang Wu [1,2] and Cheng Tan [1,2]**

[1]College of Animal Science and National Engineering Research Center for Breeding Swine Industry, South China Agricultural University, Guangzhou, China, [2]National Engineering Research Center for Breeding Swine Industry, WENS Foodstuff Group Co., Ltd., Yunfu, China

It has been proven that the random regression model has a great advantage over the repeatability model in longitudinal data analysis. At present, the random regression model has been used as a standard analysis method in longitudinal data analysis. The aim of this study was to estimate the variance components and heritability of semen traits over the reproductive lifetime of boars. The study data, including 124,941 records from 3,366 boars, were collected from seven boar AI centers in South China between 2010 and 2019. To evaluate alternative models, we compared different polynomial orders of fixed, additive, and permanent environment effects in total 216 models using Bayesian Information Criterions. The result indicated that the best model always has higher-order polynomials of permanent environment effect and lower-order polynomials of fixed effect and additive effect regression. In Landrace boars, the heritabilities ranged from 0.18 to 0.28, 0.06 to 0.43, 0.03 to 0.14, and 0.05 to 0.24 for semen volume, sperm motility, sperm concentration, and abnormal sperm percentage, respectively. In Large White boars, the heritabilities ranged from 0.20 to 0.26, 0.07 to 0.15, 0.10 to 0.23, and 0.06 to 0.34 for semen volume, sperm motility, sperm concentration, and abnormal sperm percentage, respectively.

Keywords: semen trait, random regression model, variance components, heritability, boars

OPEN ACCESS

Edited by:
Peng Xu,
Xiamen University, China

Reviewed by:
Yunlei Li,
Institute of Animal Sciences (CAAS), China
Zhili Li,
Foshan University, China

****Correspondence:***
Zhenfang Wu
wzfemail@163.com
Cheng Tan
tancheng200508@163.com

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *30 October 2021*
Accepted: *10 January 2022*
Published: *04 February 2022*

Citation:
Hong Y, Yan L, He X, Wu D, Ye J, Cai G, Liu D, Wu Z and Tan C (2022) Estimates of Variance Components and Heritability Using Random Regression Models for Semen Traits in Boars. Front. Genet. 13:805651. doi: 10.3389/fgene.2022.805651

INTRODUCTION

The use of artificial insemination (AI) is widely applied in the intensive pig industry as AI incredibly accelerates genetic progress by using semen of superior boars. However, young boars are seldom selected for AI considering semen traits. Even boars with high breeding values regarding production traits will still be culled due to poor semen quality. In addition to this, eliminative boars will receive a lower price in the commodity market (Lopez Rodriguez et al., 2017). Therefore, semen traits, such as semen volume, sperm motility, sperm concentration, and abnormal sperm percentage, affect the profitability of AI centers tremendously (Tesfay et al., 2020). So, adding semen traits to the selection index for boars at an early age not only can improve the profitability of AI centers but also increase female reproductive ability.

Some important economic traits of pigs, such as semen trait and body growth, are recoded with the age. In a few cases, the assumption of a repeatability model is invalid; however, a multivariate

animal model would be highly overparameterized (Meyer and Hill, 1997). Random regression (RR) models were explained by Henderson (Henderson, 1982), and RR models are used to explain the repeated record and the longitudinal data that were collected multiple times for a single trait during the life time of animals (Hill, 1999).

In order to elaborate a reliable selection program for semen traits, estimating variance components and heritability for semen traits is important. Recently, more and more genetic parameters related to pig semen traits have been reported. However, both genetic and population diversity can influence the variance components. In addition, most of the studies related to semen traits are based on animal models or repeatability models (Wolf, 2009b; Marques et al., 2017; Li et al., 2019), and there is little research using random RR models for semen traits. Compared with animal models and repeatability models (Oh et al., 2006), the RR model demonstrated change of meaning and covariance along with age (Schaeffer, 2004). The RR model is not only widely used to estimate genetic parameters of milk yield in cows but also used for the analysis of growth data in pigs and beef cattle (Meyer, 1999; Andersen and Pedersen, 2010; Sasaki et al., 2017). The purpose of this study was to use an RR model to estimate genetic parameters of a series of semen traits, including semen volume, sperm motility, sperm concentration, and abnormal sperm percentage, in a large data set.

MATERIALS AND METHODS

Ethical review and approval were not required for the animal study because the data used for this study were collected as part of routine data that are recorded in a commercial breeding program. Semen collections were conducted strictly in line with the Guidelines for the Care and Use of Experimental Animals established by the Ministry of Science and Technology of the People's Republic of China. All efforts were made to minimize animal suffering.

Data from seven AI centers of southern China were collected between 2010 and 2019. The total number of ejaculates was 124941 stemming from 3366 AI boars, including Landrace (LA, n = 1147) and Large White (LW, n = 2219). Semen volume (ml; VOL), sperm motility (%; MOT), sperm concentration (10 ^ 8/ml; CON), and abnormal sperm percentage (%; ABN) were considered in this study. VOL was measured by weighting each ejaculate and considering 1 g of semen to 1 ml. MOT, CON, and ABN were measured by a microscope (before 2017) and a computer-assisted sperm analysis system (after 2017).

Combined with previous studies and the characteristics of the data set, the following criteria are applied to data quality control: (a) the range age of boars between 33 and 150 weeks; (b) the first record was excluded, and the interval between two subsequent semen collections was within the range of 1–30 days; (c) animals with minimum ejaculation number (set to 6) were chosen to calculate the within-boar variation of the studies trait; (d) each fixed effect level should have at least 10 ejaculation records; (e) records on the VOL, MOT, CON, and ABN should be within 100 ml–600 ml, 10%–100%, 0.1 * 10⁸/ml–8 * 10⁸/ml, and 0.01–100%. After data filtering, the clean data of each breed are presented in **Table1**. These data and three-generation pedigree of boars were applied to the subsequent analysis.

TABLE 1 | Number of boars and ejaculates for two breeds.

Breed[1]	Number of boars	Number of ejaculates	Ejaculates per boar	SD	Min[2]	Max[3]
LD	1147	38950	33.96	23.93	6	158
LW	2219	85991	38.75	26.38	6	150

Note: [1]LD, Landrace; LW, Large White; [2]Min = minimum number of records; [3]Max = maximum number of records.

The following random regression model is used to estimate the (co) variance and breeding value:

$$y_{ijt} = AIYS_i + Interval_j + \sum_{l=0}^{nf} \varphi_n(w_t)\beta_k + \sum_{l=0}^{nr} \varphi_n(w_t)a_k + \sum_{l=0}^{nr} \varphi_n(w_t)Pe_k + e_{ijt},$$

y_{ijt} denotes the semen traits of boar recorded on day t within $AIYS$ subclass i and $Interval$ subclass j; μ is the overall mean; $AIYS_j$ is the combined effects which include the AI center, year, and month; $Interval_k$ is the interval effect between two semen collections; β_k is the fixed regression coefficients for the effect of the boar's age; a_k and Pe_k are random regression coefficients for the additive genetic and permanent environmental effects, respectively, the terms $\varphi_n(w_t)$ correspond with Legendre polynomials evaluated at standardized time w_t ($-1 \le w_t \le 1$), and the residual is given by e_{ijt}. The matrix of the model is accordingly denoted as follows: $y = X_1b_1 + X_2b_2 + Z_1a + Z_2p + e$, where y is the vector of phenotypes; b_1 is the vector of fixed effects; b_2 is the vector of fixed regression coefficients; a and p are vectors of random regressions for additive genetic and permanent environmental effect, respectively, X_1, X_2, Z_1, and Z_2 are design matrices of b_1, b_2, a, and p, respectively; and e is the vector of residuals. It was assumed that

$$var\begin{bmatrix} a \\ p \\ e \end{bmatrix} = \begin{bmatrix} A\sigma_a^2 & 0 & 0 \\ 0 & I\sigma_p^2 & 0 \\ 0 & 0 & I\sigma_e^2 \end{bmatrix},$$

where A is the numerator relationship, I is an identity matrix whose dimension is equal to the number of effect levels, σ_a^2 and σ_p^2 are co (variance) matrices of additive genetic and permanent environmental regression coefficients, respectively, and σ_e^2 stand for residual variance.

Legendre polynomials were generated using the following recursion formula:

$$(n+1)P_{n+1}(w_t) = (2n+1)w_1p_n(w_t) - nP_{n-1}(w_t),$$

where $P_0(w_t) = 1$ and $P_1(w_t) = t$. $P_n(w_t)$ is the polynomial of order n, and w_t is the standardized time variable in the interval [-1,1] as $w_t = -1 + \frac{2(a_t - a_{min})}{a_{max} - a_{min}}$, in which a_t is the boar's age when collecting semen traits, and a_{min} and a_{min} represent the first and latest boar's age when collecting semen traits, respectively. The

TABLE 2 | Description statistics for semen traits in two breeds.

Breed[1]	Trait[2]	Mean	SD	Min3	Max4
LD	VOL	240.62	77.44	100	586
	MOT	68.68	12.01	10	100
	CON	3.23	1.96	0.1	8
	ABN	12.07	11.20	0.01	100
LW	VOL	255.7	85.98	100	600
	MOT	68.76	11.57	10	100
	CON	3.12	1.29	0.1	8
	ABN	12.03	10.17	0.01	100

Note:[1]LD, Landrace; LW, Large White; [3]Trait: VOL, semen volume (ml), MOT, sperm motility (%), CON, sperm concentration (108/ml), ABN, abnormal sperm percentage (%), [2]Min = minimum number of records; [3]Max = maximum number of records.

normalized value of the nth Legendre polynomial evaluated at age t ($\varphi_n(w_t)$) is as follows:

$$\begin{aligned}\varphi_0(w_t) &= 0.7071; \varphi_1(w_t) = 1.2247t; \varphi_2(w_t)\\ &= 2.3717(t^2) - 0.7906; \varphi_3(w_t)\\ &= 4.6771(t^3) - 2.8067t; \varphi_4(w_t)\\ &= 9.2808(t^4) - 7.9550(t^2) + 0.7955; \varphi_5(w_t)\\ &= 18.4685(t^5) - 20.5206(t^3) + 4.3973t; \varphi_6(w_t)\\ &= 36.8085(t^6) - 50.1935(t^4) + 16.7312(t^2)\\ &\quad - 0.7967; \varphi_7(w_t)\\ &= 73.4291(t^7) - 118.6162(t^5) + 53.9164(t^3)\\ &\quad - 5.9907t; \varphi_8(w_t)\\ &= 146.5710(t^8) - 273.5992(t^6) + 157.8457(t^4)\\ &\quad - 28.992(t^2) + 0.7972;\end{aligned}$$

We also used general linear models (GLMs) to perform the Waller–Duncan k-ratio *t*-test on the effects of semen collection interval. Models named L (a, b, and c) indicate the order of the polynomial fitted for fixed effects (a), additive genetic (b), and permanent environmental effects. This resulted in the evaluation of 200 models. Here, random regression models are fitted to be evaluated first through eighth-order polynomial covariance functions for the fixed effects of boar age classification and second through sixth-order polynomial covariance functions for the additive genetic and permanent environmental effects. The fitness of model was tested by Bayesian Information Criterions (BICs) (Neath and Cavanaugh, 2011):

$$BIC = -2log(L) + p \times log(n);$$

where log(L) is the log-likelihood value, P stands for the number of parameters, and n is the sample size.

The semen trait change over time for all selected boars can be represented by $V = \varphi_n(w_t)\beta_k$, where V is a vector of actual semen traits from the boar's age: 36 to 136 weeks, and others are same as before.

The estimate effects of the interval between two subsequent semen collections were also obtained from the abovementioned model.

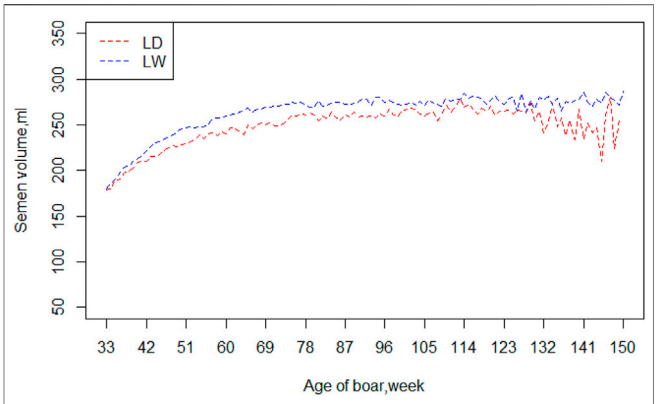


FIGURE 1 | Mean of semen volume by age of boars in two breeds. Note: LD = Landrace, LW = Large White.

TABLE 3 | Order of fit for fixed (L_F), additive genetic (L_A), and permanent environmental (K_P) effects; negative twice of log likelihood (–2logL); Bayesian Information Criterions (BIC); and ranks of BIC.

Breed and trait[1]	Model 2	P3	-2Log(L)4	BIC5	Rank6
LD/VOL	L (4,3,8)	13	351983	352121	1
	L (5,3,8)	14	351975	352123	2
	L (6,3,8)	15	351965	352124	3
LD/MOT	L (3,3,8)	12	201527	201653	1
	L (3,4,8)	13	201522	201660	2
	L (3,5,8)	14	201523	201671	3
LD/CON	L (4,8,4)	14	38829	38977	2
	L (7,5,5)	15	38592	38751	1
	L (8,7,4)	17	39525	39705	3
LD/ABN	L (3,3,8)	12	198427	198552	1
	L (3,5,8)	14	198410	198558	2
	L (4,3,8)	13	198422	198560	3
LW/VOL	L (5,3,8)	14	775951	776110	2
	L (5,4,8)	15	775939	776109	1
	L (6,3,8)	15	775940	776111	3
LW/MOT	L (3,4,8)	13	436804	436952	2
	L (3,5,8)	14	436788	436947	1
	L (4,3,8)	13	436809	436957	3
LW/CON	L (8,4,5)	15	74339	74509	3
	L (8,7,3)	16	73553	73735	2
	L (8,7,4)	17	73451	73644	1
LW/ABN	L (3,4,8)	13	421778	421925	2
	L (3,5,8)	14	421767	421927	3
	L (4,5,7)	14	413477	413636	1

Note:[1]LD, Landrace; LW, Large white; VOL, semen volume (ml), MOT, sperm motility (%), CON, sperm concentration (108/ml), ABN, abnormal sperm percentage (%). [2]L(a, b, and c) = the order of the polynomial fitted for fixed effects (a), additive genetic (b), and permanent environmental effects. [3]P = the number of parameters in random regression models. [4]Log(L) = log-likelihood value. [5]BIC, Bayesian Information Criterions. [6]Rank = The ranking of the top three models.

RESULTS

Table 2 exhibited a series of information, including means, standard deviations, and minimum and maximum of the four semen traits (VOL, DEN, MOT and ABN), for the two breeds. The means and standard deviation of semen volume were 240.62

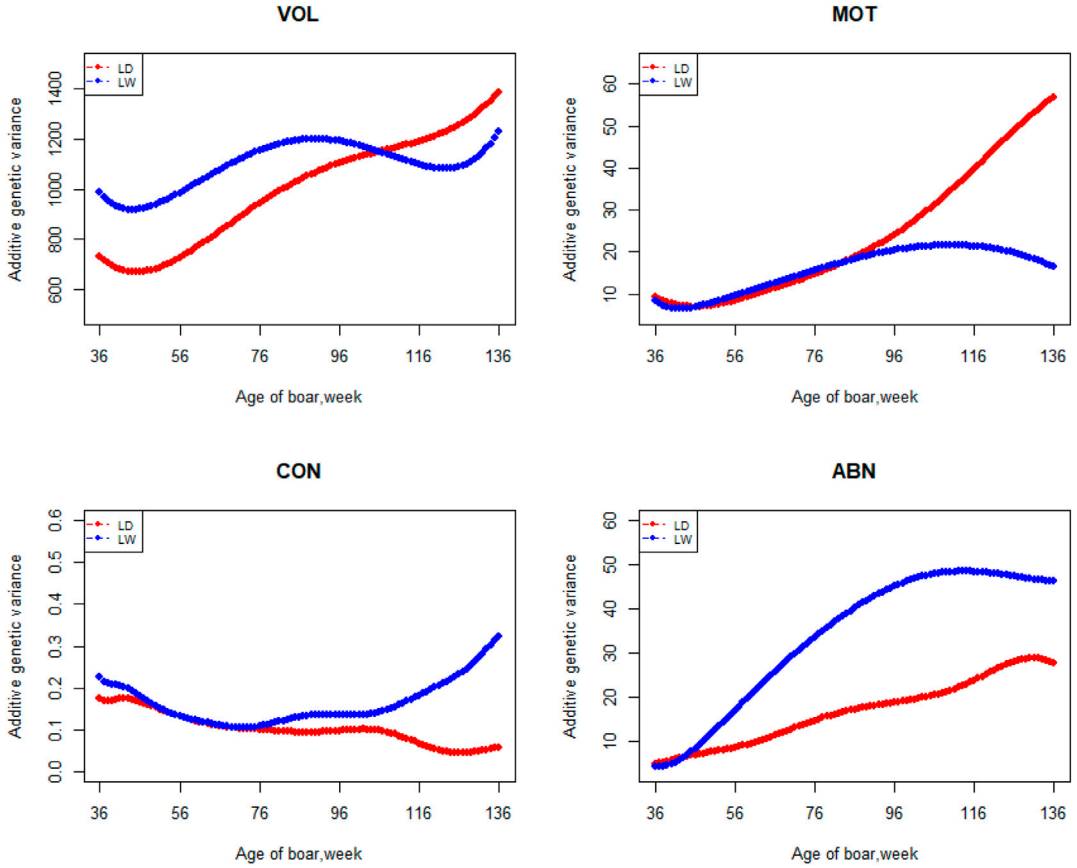


FIGURE 2 | Additive genetic variance from 36 weeks to the age of 136 weeks, estimated with a random regression animal model.

and 77.44 ml in Landrace boars and 255.7 and 85.98 ml in Large White boars, respectively. The Large White boars had a larger mean of semen volume than Landrace boars. For the other three traits (DEN, MOT, and ABN), the Landrace boars and Large White boars had similar mean value and standard deviation. **Figure 1** shows the average value of VOL in different ages of the two breeds. Large White boars always have a higher VOL value than Landrace boars over time, but their developing trends are similar. The 80th week is a turning point. Before 80 weeks, the means of semen volume increased followed with age, and it became stable after 80 weeks.

Table 3 showed the order of fit for fixed (LF), additive genetic (LA), and permanent environmental (LP) effects; number of parameters (P), -2 times log likelihood (-2log(L), and Bayesian Information Criterions (BICs). The random regression model that fits LF = 4, LA = 3, and LP = 8 and LF = 5, LA = 4, and LP = 8 coefficients for fixed, additive genetic, and permanent environmental effects showed the smallest BIC for VOL in Landrace and Large White boars, respectively. For the MOT, LF = 3, LA = 3, LP = 8 and LF = 3, LA = 5, LP = 8 coefficients for fixed, additive genetic, and permanent environmental effects showed the smallest BIC in Landrace and Large White boars, respectively. In addition to this, LF = 7, LA = 5, LP = 5 and LF = 8, LA = 7, LP = 4 coefficients for fixed, additive genetic, and permanent environmental effects showed the smallest BIC for CON in Landrace and Large White boars, respectively. For the ABN, LF = 3, LA = 3, LP = 8 and LF = 4, LA = 5, LP = 7 coefficients for fixed, additive genetic, and permanent environmental effects showed the smallest BIC in Landrace and Large White boars, respectively. In conclusion, the low order of LA and LF and the high order of LP (especially LP = 8) best fits the RR model for VOL, MOT, and ABN, but not CON.

The variation of the additive genetic variance, permanent environmental variance, heritability, and repeatability for four semen traits in the two breeds changes over time as shown in **Figures 2**–**5**. In Landrace, the genetic variance, heritability, and repeatability estimates for VOL, MOT and ABN increased as the boar matured, while the CON decreased or remained stable along with time. In Large White, the additive variance and heritability estimates of ABN increase clearly as the boar matured, while the VOL constantly increase with age, and the additive variance and heritability for CON and MOT fluctuate somewhat over time. In summary, in Landrace, the additive variance ranged from 699.99 to 1384.12, 6.98 to 56.97, 0.05 to 0.17, and 4.81 to 28.73 for VOL, MOT, CON, and ABN, respectively. The permanent environmental variance ranged from 288.94 to 1440.16, 17.02 to 412.41, 0.20 to 0.64, and 41.16 to 152.06 for VOL, MOT, CON, and ABN, respectively. The heritabilities ranged from 0.18 to

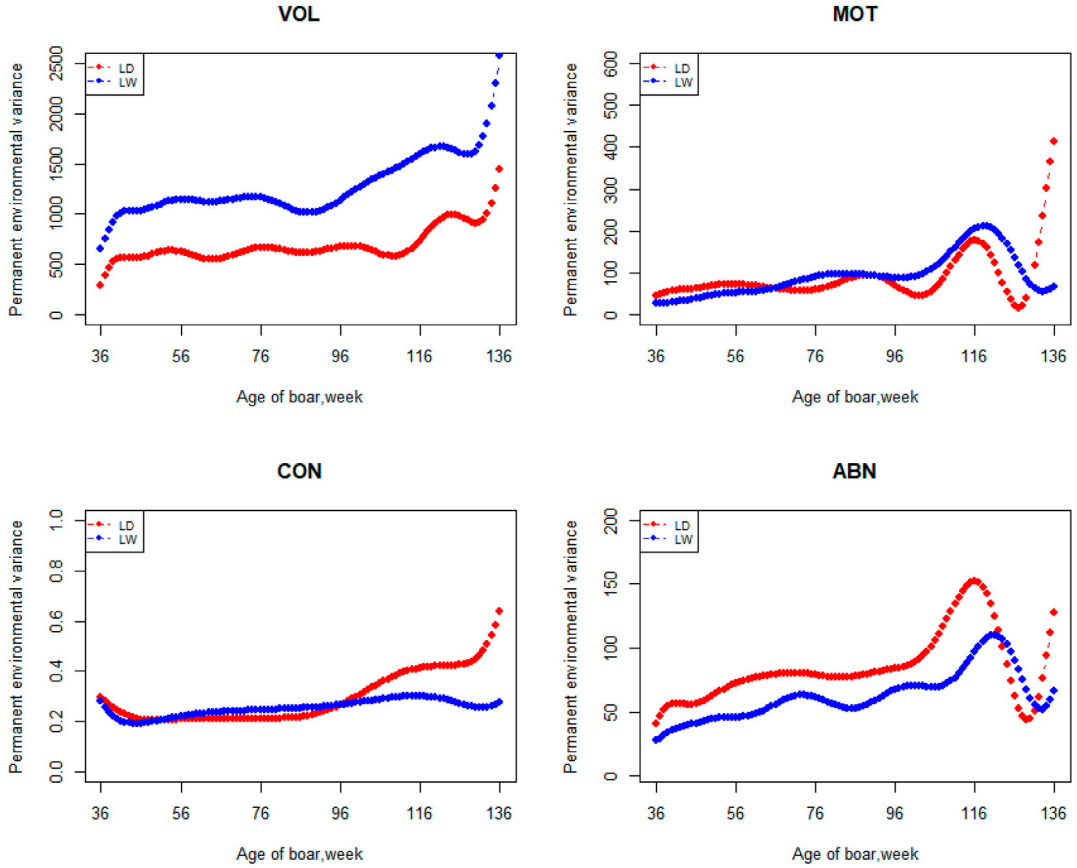


FIGURE 3 | Permanent environmental variance from 36 weeks to the age of 136 weeks, estimated with a random regression animal model.

0.28, 0.06 to 0.43, 0.03 to 0.14, and 0.05 to 0.24 for VOL, MOT, CON, and ABN, respectively. The repeatability ranged from 0.29 to 0.53, 0.52 to 0.89, 0.26 to 0.42, and 0.50 to 0.80 for VOL, MOT, CON, and ABN, respectively. In Large White, the additive variance ranged from 918.31 to 1228.71, 6.45 to 21.62, 0.11 to 0.32, and 4.15 to 48.41 for VOL, MOT, CON, and ABN, respectively. The permanent environmental variance ranged from 658.85 to 2568.19, 26.50 to 212.22, 0.19 to 0.30, and 28.08 to 110.30 for VOL, MOT, CON, and ABN, respectively. The heritabilities ranged from 0.20 to 0.26, 0.07 to 0.15, 0.10 to 0.23, and 0.06 to 0.34 for VOL, MOT, CON, and ABN, respectively. The repeatabilities ranged from 0.40 to 0.61, 0.43 to 0.83, 0.32 to 0.44, and 0.47 to 0.81 for VOL, MOT, CON, and ABN, respectively.

The genetic correlation estimate of VOL, MOT, CON, and ABN at different ages of boar is shown in **Tables 4–7** respectively. Genetic correlations declined as the interval between ages increased in VOL and increased first and then increased in MOT, CON, and ABN. Genetic correlations of VOL, MOT, CON and ABN with the best random regression model from week 33 to week 150 range from 0.32 to 1, −0.52 to 1, −0.41 to 1, and 0.13 to 1 in Landrace. In Large White, the genetic correlation ranged from 0.06 to 1, 0.43 to 1, −0.33 to 1, and −0.13 to 1 for VOL, MOT, CON, and ABN in Landrace, respectively.

The interval between two subsequent semen collections on semen traits had a large effect (**Figure 6**). VOL increased when the interval was prolonged from 1 to 9 days; however, it decreases starting from the 10-day interval in the two breeds. In comparison, it is not an obvious effect in the interval between two subsequent semen collections on MOT, CON, and ABN. From the perspective of CON and ABN, the most suitable collective interval for Landrace is 2 days. When the semen collection interval is 9–10 days, the CON reaches the maximum in the two breeds.

DISCUSSION

It is very quick to obtain the result by using a repeatability model to analyze semen traits because of a substantially lower number of parameters (Gredler et al., 2007; Wolf, 2009a; Burren et al., 2019). However, the repeatability model assumes that two repeated measurements should have the same genetic correlations. Therefore, this model has some defects. First, the heritability of semen traits is different at different ages. Several researchers have reported that the heritability of daily milk yields is different from days in milk (Takma and Akbas, 2007). Second, genetic correlations among repeated

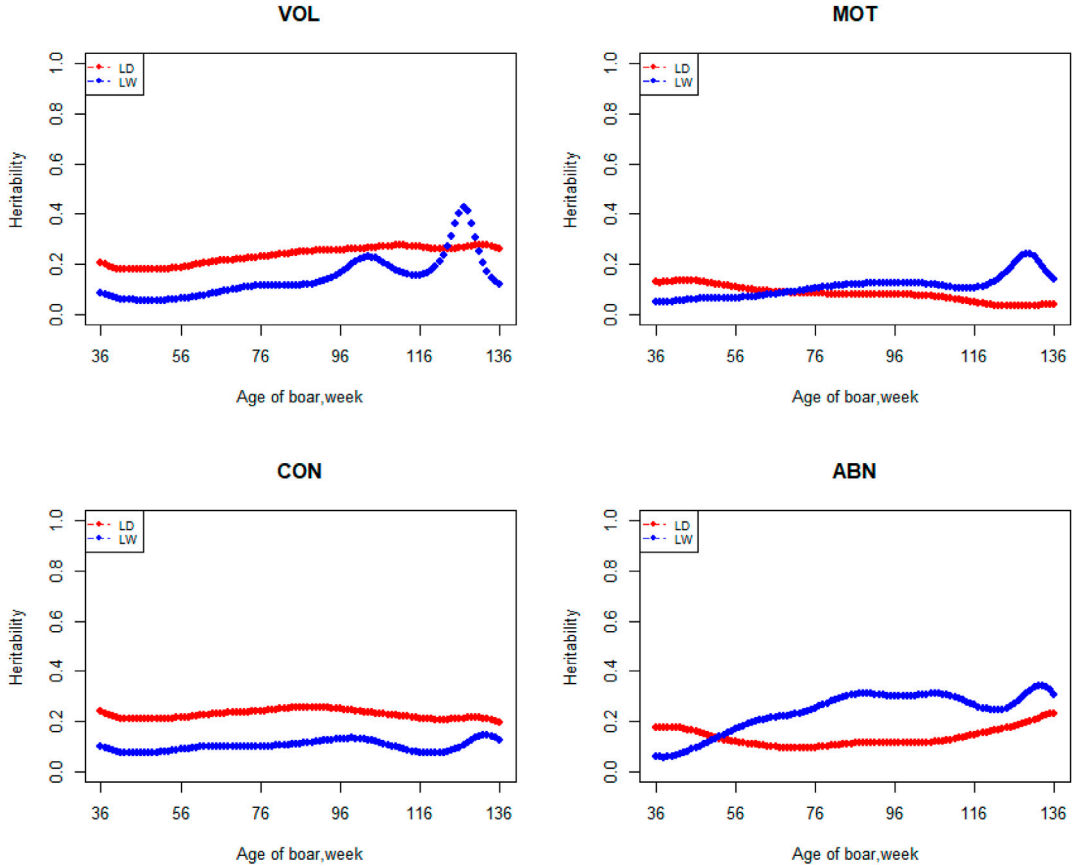


FIGURE 4 | Heritability from 36 weeks to the age of 136 weeks, estimated with a random regression animal model.

measurements usually tend to decrease as functions of time. In comparison with the repeatability model, the RR models allow for modeling variance components as time functions, although the more parameters estimated, the more reasonable results. Using Legendre polynomials to fit RR models needs to be carried out carefully when interpreting results in extreme cases of boar age (Oh et al., 2006; Carabano et al., 2007). Li's study shows that a random regression model with third-order of LP is suggested to be an appropriate model for genetic evaluation of milk yield in local Chinese Holstein populations (Li et al., 2020). However, the best model always has a higher order of permanent environment effect in the current study. Prakash discovered that the RR model with a lower-order polynomial for modeling additive genetic effect and higher-order polynomial for modeling animal permanent environmental effect is optimal for genetic evaluation (Prakash et al., 2017). In our study, a sudden increase in additive and permanent environmental variance relevant to the number of samples with over 136 weeks is small. Some studies reported that using splines to fit RR models was more robust against end of extreme time problems than polynomial models (Meyer, 2005; Bohmanova et al., 2008).

The best model indicated that VOL is a medium heritability trait with heritability ranging from 0.18 to 0.28 and from 0.20 to 0.26 in Landrace and Large White, respectively. These values strongly agreed with using the repeatability model for this trait reported previously by Wolf and Li (Wolf, 2009b; Li et al., 2019). Wolf estimated heritabilities of 0.19–0.25 for VOL in Czech Landrace, and Li estimated heritabilities of 0.25 ± 0.02, 0.21 ± 0.02, and 0.23 ± 0.02 for VOL in Duroc, Landrace, and Yorkshire, respectively. Heritability of VOL tends to increase and then decrease over time in Large White boars. Strathe reported a similar trend in heritability in the semen trait of pigs (Strathe et al., 2013). The heritability of sperm motility ranges from 0.06 to 0.43 and 0.07 to 0.15 in Landrace and Large White boars, respectively. Heritability changes suddenly after 96 weeks of age in the Landrace breed. It may due to insufficient data for Landrace boars, especially the data after 96 weeks. Marques reported that the heritabilities of MOT were 0.25 ± 0.05 and 0.08 ± 0.03 in Large White and Landrace boars, respectively. For the Large White, the heritability of MOT is higher than what we have studied, which may be due to the difference in population structure. Wolf estimated heritabilities of 0.06 ± 0.02 and 0.16 ± 0.03 for MOT in Czech Large White and Czech Landrace boars, respectively, and it is similar to our research. The heritabilities ranged from 0.03 to 0.14 and 0.10 to 0.23 for CON in Landrace and Large White boars, respectively. Grandjot estimated heritabilities of 0.17–0.26 (Grandjot et al., 1997) and Strathe

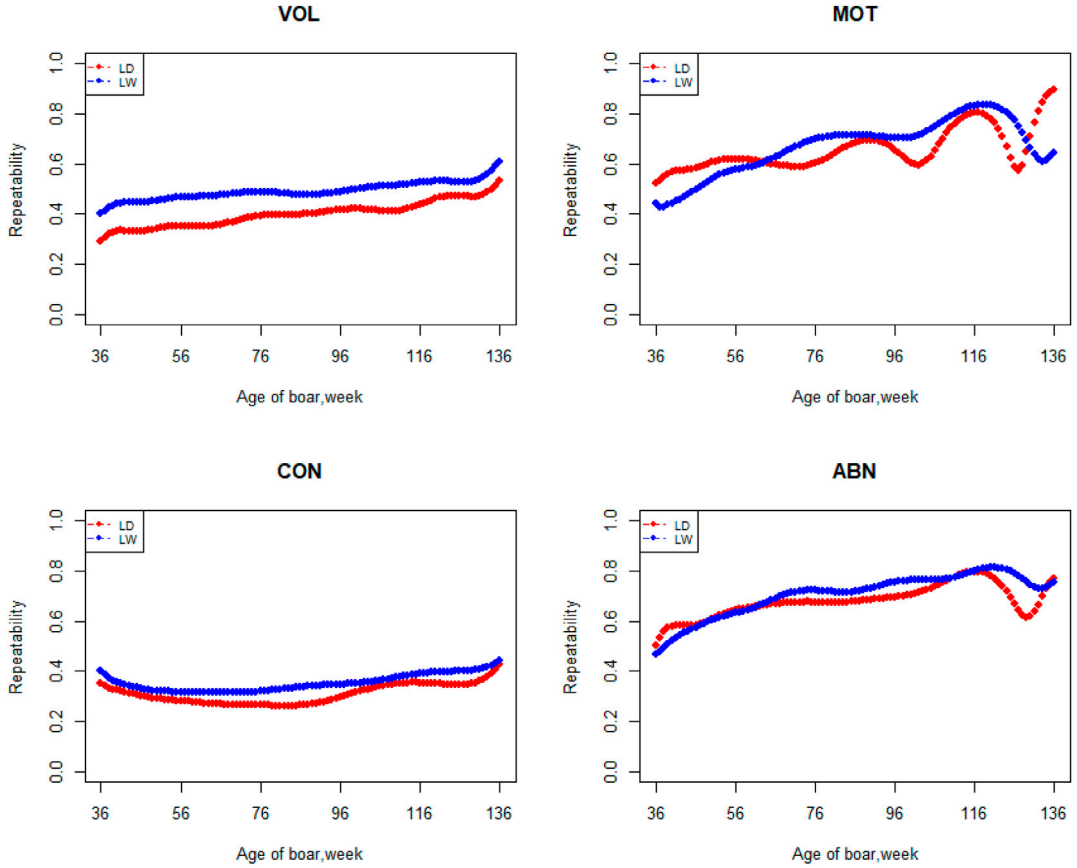


FIGURE 5 | Repeatability from 36 weeks to the age of 136 weeks, estimated with a random regression animal model.

estimated heritabilities of 0.23–0.26 for CON in Danish Landrace boars, which is slightly greater than the current estimates (Strathe et al., 2013). The heritabilities ranged from 0.05 to 0.24 and 0.06 to 0.34 for ABN in Landrace and Large White boars, respectively, which are well-consistent with those (0.15 ± 0.01 to 0.21 ± 0.02) obtained by Li and (0.15 ± 0.05 to 0.25 ± 0.06) estimated by Marques (Li et al., 2019; Marques et al.) using average value over all ejaculates for each boar and obtained estimates of 0.58, 0.38, 0.49, and 0.34 for VOL, MOT, CON, and ABN by Smital estimated, being substantially greater than the current estimates (Smital et al., 2005). This is reasonable because if the repeatability is less than 1, the heritability of the average number of records must be greater than that of a single record.

Apart from estimates of heritabilities, the estimates of the additive genetic variance and repeatability are also of particular interest to animal breeders. The additive variance directly determines the response to selection and the opportunities for genetic change by natural or artificial selection (Hill et al., 2008). For the VOL and ABN, the additive variance of Large White is obviously higher than that of Landrace in the first 100 weeks of age. The estimates of additive genetic variances for MOT and CON not differ greatly at the first 80 weeks of age. In the later stages of boar life, the additive variance usually varies a lot. It is determined by the character of models and the semen traits. In addition, fewer pigs survive as they age, leading to fewer records of high-frequency ejaculation. The repeatability of Large White is higher than that of Landrace in VOL and CON, but it showed fluctuation in the MOT and ABN.

As mentioned above, the estimate of heritability from four semen traits in Landrace and Large White indicated that selection for VOL could achieve reasonable rapid genetic gains. However, for the other three traits, the result indicated that the traditional selection will not gain genetic progress quickly because of low heritability. In addition, the boar semen traits are sex-limited traits, leading to the effect of traditional selection based on phenotype, and genealogical information is not obvious (Brigatti, 2021). Genome selection has outstanding advantages in complex traits and low heritability traits (Ibáñez-Escriche et al., 2014). How to estimate the breeding value of these traits and how to incorporate them into selectivity indicators will be considered in the next stage.

Genetic correlations between measurements at the age of 33 through 150 weeks are of great differences. Those results indicate that a repeatability model is an unacceptable approach to model variation for semen traits in this population. Genetic correlation decreases with age, which may also be due to limited data and

TABLE 4 | Genetic correlations between semen volume of all ages, estimated with a random regression model.

Age week	33	42	51	60	69	78	87	96	105	114	123	132	141	150
33	–	0.96	0.85	0.73	0.6	0.5	0.42	0.35	0.29	0.24	0.19	0.14	0.1	0.06
42	0.96	–	0.97	0.89	0.8	0.72	0.65	0.59	0.52	0.46	0.38	0.29	0.19	0.09
51	0.85	0.96	–	0.98	0.93	0.87	0.82	0.77	0.71	0.64	0.55	0.43	0.28	0.13
60	0.72	0.89	0.98	–	0.99	0.96	0.92	0.88	0.83	0.76	0.66	0.53	0.36	0.18
69	0.6	0.8	0.93	0.99	–	0.99	0.97	0.94	0.9	0.84	0.74	0.61	0.43	0.23
78	0.51	0.73	0.88	0.96	0.99	–	0.99	0.98	0.94	0.89	0.8	0.67	0.49	0.28
87	0.44	0.67	0.84	0.93	0.98	1	–	0.99	0.97	0.93	0.86	0.73	0.56	0.35
96	0.4	0.63	0.8	0.9	0.96	0.98	1	–	0.99	0.96	0.9	0.79	0.63	0.43
105	0.36	0.59	0.76	0.87	0.93	0.96	0.98	0.99	–	0.99	0.95	0.86	0.71	0.53
114	0.35	0.56	0.72	0.83	0.89	0.93	0.95	0.98	0.99	–	0.98	0.92	0.81	0.64
123	0.34	0.53	0.67	0.77	0.83	0.87	0.91	0.94	0.97	0.99	–	0.98	0.9	0.77
132	0.33	0.49	0.61	0.69	0.74	0.79	0.83	0.87	0.91	0.95	0.99	–	0.97	0.89
141	0.33	0.44	0.53	0.59	0.63	0.67	0.72	0.77	0.82	0.88	0.94	0.98	–	0.97
150	0.32	0.39	0.43	0.47	0.5	0.54	0.58	0.64	0.71	0.78	0.86	0.93	0.98	

Note: Genetic correlations between semen volume of all ages between 33 and 150 weeks, estimated with a random regression model. Genetic correlations on the diagonal were Large White boars and under the diagonal were Landrace boars.

TABLE 5 | Genetic correlations between sperm motility of all ages, estimated with a random regression model.

Age week	33	42	51	60	69	78	87	96	105	114	123	132	141	150
33	–	0.93	0.8	0.68	0.59	0.53	0.49	0.46	0.44	0.43	0.43	0.45	0.47	0.52
42	0.76	–	0.96	0.9	0.85	0.8	0.77	0.75	0.74	0.73	0.73	0.74	0.76	0.79
51	0.32	0.85	–	0.98	0.96	0.93	0.91	0.9	0.89	0.89	0.89	0.89	0.91	0.93
60	0	0.63	0.94	–	0.99	0.98	0.97	0.96	0.96	0.95	0.95	0.96	0.97	0.98
69	–0.22	0.43	0.82	0.97	–	1	0.99	0.99	0.98	0.98	0.98	0.98	0.99	1
78	–0.37	0.23	0.67	0.87	0.97	–	1	1	0.99	0.99	0.99	1	1	1
87	–0.47	0.05	0.49	0.74	0.88	0.97	–	1	1	1	1	1	1	1
96	–0.51	–0.1	0.32	0.59	0.77	0.9	0.98	–	1	1	1	1	1	1
105	–0.52	–0.2	0.18	0.45	0.65	0.81	0.93	0.99	–	1	1	1	1	1
114	–0.5	–0.26	0.09	0.35	0.57	0.74	0.88	0.96	0.99	–	1	1	1	0.99
123	–0.47	–0.27	0.05	0.31	0.52	0.71	0.85	0.94	0.98	1	–	1	1	1
132	–0.44	–0.24	0.07	0.31	0.52	0.7	0.85	0.94	0.98	1	1	–	1	1
141	–0.41	–0.18	0.14	0.38	0.58	0.75	0.88	0.95	0.99	1	1	1	–	1
150	–0.38	–0.07	0.27	0.5	0.68	0.82	0.92	0.97	0.99	0.98	0.98	0.98	0.99	–

Note: Genetic correlations between sperm motility of all ages between 33 and 150 weeks, estimated with a random regression model. Genetic correlations on the diagonal were Large White boars and under the diagonal were Landrace boars.

TABLE 6 | Genetic correlations between sperm concentration of all ages, estimated with a random regression model.

Age week	33	42	51	60	69	78	87	96	105	114	123	132	141	150
33	–	0.95	0.86	0.74	0.58	0.44	0.32	0.21	0.07	–0.07	–0.19	–0.29	–0.33	–0.21
42	0.81	–	0.97	0.85	0.66	0.5	0.4	0.32	0.24	0.15	0.05	–0.06	–0.19	–0.25
51	0.62	0.96	–	0.95	0.8	0.66	0.58	0.52	0.47	0.4	0.3	0.2	0.06	–0.08
60	0.6	0.95	1	–	0.95	0.87	0.81	0.77	0.71	0.62	0.51	0.43	0.36	0.24
69	0.67	0.96	0.98	0.99	–	0.98	0.95	0.91	0.84	0.73	0.61	0.56	0.58	0.53
78	0.77	0.95	0.93	0.94	0.98	–	0.99	0.96	0.89	0.77	0.66	0.63	0.7	0.69
87	0.82	0.9	0.83	0.85	0.92	0.98	–	0.99	0.93	0.83	0.73	0.71	0.78	0.76
96	0.81	0.83	0.75	0.77	0.86	0.94	0.99	–	0.98	0.9	0.82	0.81	0.86	0.76
105	0.72	0.76	0.71	0.74	0.82	0.91	0.97	0.99	–	0.97	0.93	0.91	0.91	0.7
114	0.52	0.66	0.67	0.71	0.79	0.86	0.9	0.92	0.97	–	0.99	0.97	0.92	0.6
123	0.23	0.48	0.56	0.62	0.67	0.71	0.72	0.76	0.84	0.95	–	0.99	0.91	0.54
132	–0.01	0.23	0.34	0.4	0.45	0.49	0.53	0.58	0.69	0.85	0.96	–	0.94	0.59
141	0.02	–0.09	–0.09	–0.03	0.08	0.21	0.35	0.47	0.58	0.68	0.74	0.84	–	0.83
150	0.26	–0.24	–0.41	–0.39	–0.25	–0.04	0.16	0.29	0.33	0.27	0.18	0.26	0.73	–

Note: Genetic correlations between sperm concentration of all ages between 33 and 150 weeks, estimated with a random regression model. Genetic correlations on the diagonal were Large White boars and under the diagonal were Landrace boars.

TABLE 7 | Genetic correlations between abnormal sperm percentage of all ages, estimated with a random regression model.

Age week	33	42	51	60	69	78	87	96	105	114	123	132	141	150
33	–	0.33	−0.04	−0.13	−0.13	−0.09	−0.05	−0.02	0.01	0.04	0.05	0.07	0.09	0.11
42	0.93	–	0.91	0.81	0.71	0.62	0.54	0.49	0.46	0.44	0.43	0.39	0.29	0.13
51	0.99	0.98	–	0.97	0.91	0.82	0.75	0.69	0.65	0.63	0.61	0.56	0.46	0.29
60	0.97	0.83	0.93	–	0.98	0.93	0.87	0.83	0.79	0.77	0.75	0.72	0.63	0.48
69	0.85	0.61	0.76	0.95	–	0.98	0.95	0.92	0.9	0.88	0.87	0.84	0.77	0.65
78	0.71	0.42	0.6	0.85	0.97	–	0.99	0.98	0.96	0.95	0.94	0.92	0.87	0.77
87	0.6	0.3	0.48	0.75	0.92	0.98	–	1	0.99	0.98	0.98	0.96	0.93	0.85
96	0.49	0.22	0.38	0.65	0.83	0.92	0.98	–	1	1	0.99	0.98	0.96	0.89
105	0.38	0.16	0.3	0.52	0.69	0.8	0.89	0.97	–	1	1	0.99	0.97	0.91
114	0.27	0.13	0.22	0.38	0.52	0.63	0.74	0.87	0.97	–	1	1	0.98	0.93
123	0.22	0.14	0.19	0.3	0.41	0.51	0.63	0.78	0.91	0.99	–	1	0.99	0.94
132	0.28	0.17	0.24	0.37	0.48	0.58	0.7	0.83	0.94	1	1	–	0.99	0.96
141	0.46	0.21	0.36	0.62	0.79	0.89	0.96	1	0.98	0.9	0.83	0.87	–	0.98
150	0.49	0.14	0.35	0.67	0.87	0.94	0.93	0.85	0.69	0.48	0.34	0.41	0.8	–

Note: Genetic correlations between abnormal sperm percentage of all ages between 33 and 150 weeks, estimated with a random regression model. Genetic correlations on the diagonal were Large White boars and under the diagonal were Landrace boars.

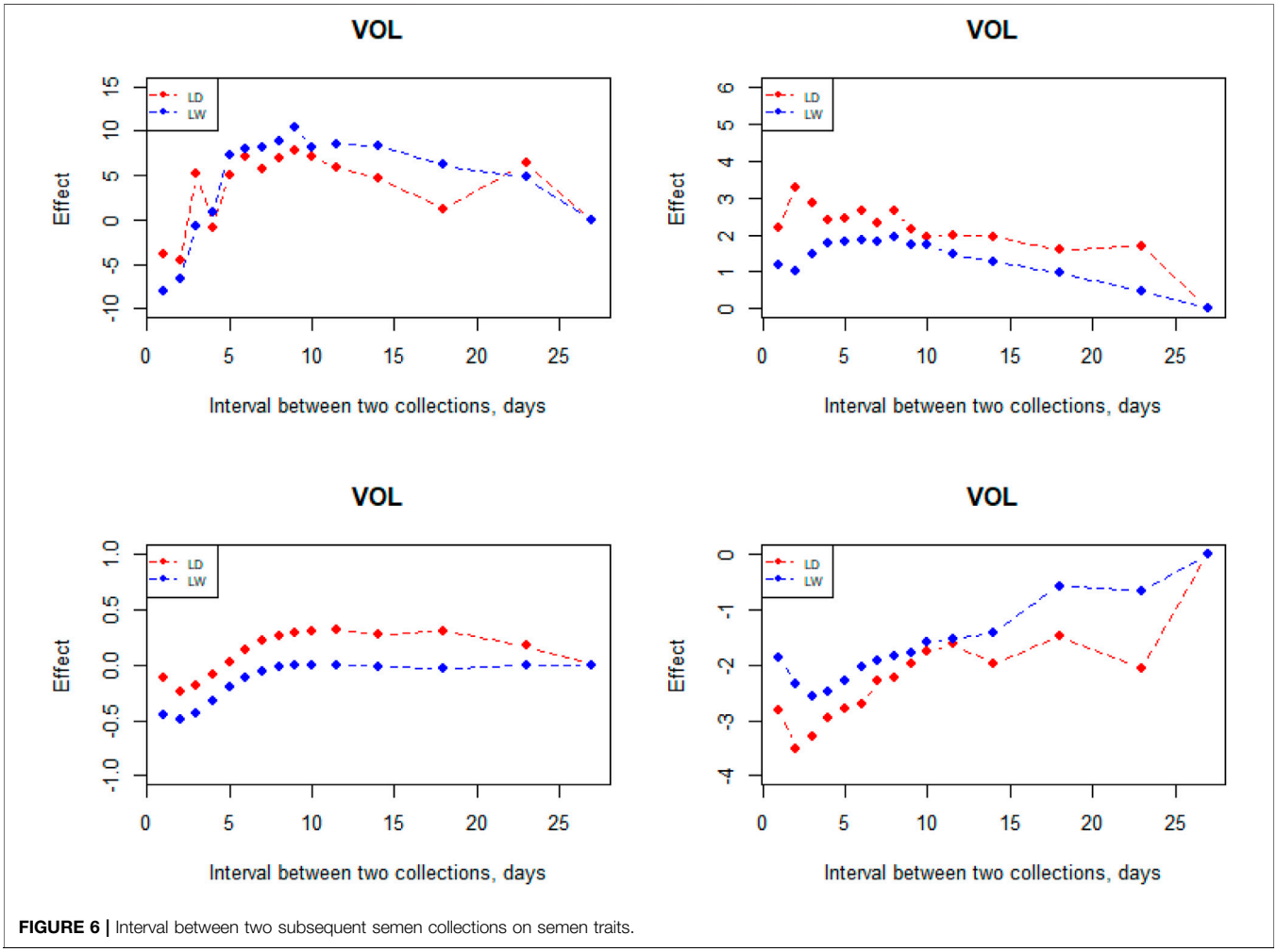


FIGURE 6 | Interval between two subsequent semen collections on semen traits.

selection of records in the prescribed age range. S.H.Oh estimated that genetic correlations were high between adjacent ages and decreased as the interval between ages increased in the sperm cell trait, and this result is consistent with our discovery (Oh et al., 2006). These results suggest that future performance may be harder to predict accurately from earlier records.

If the interval between ejaculation is too long, sperm function will be significantly reduced. However, if the

interval between ejaculation is too short, the VOL will be significantly reduced (Check et al., 1991; Knecht et al., 2017). Thus, it is important to control the interval between successive collections. Based on the result of interval effect, it is indicated that 8–10 days is a best choice to design the interval during days of successive collection for Landrace and Large White breeds. However, it is not a best choice for MOT, CON, and ABN. Wolf found that the time interval of 7–10 days seems to be a good choice for getting the values of all semen traits near optimum (Wolf and Smital, 2009). Rutten et al. (2000) investigated collection intervals from 1 to 10 days and found that the highest number of doses per collection can be generated for intervals from 7 to 10 days. These results are in good agreement. Bajena reported that ejaculate CON remained at a relatively high level when ejaculates were collected with a frequency of 3–7 days, but further shortening of the interval between the successive collections led to a drastic decrease in CON (Bajena et al., 2016).

CONCLUSION

We estimated the genetic parameters of VOL, MOT, CON, and ABN in different boar ages for two breeds. The higher-order polynomial of permanent environment effects and the lower-order polynomials for fixed effects and additive effects are the best orders to fit the random regression models. In addition, the best interval for semen collection is 8–10 days.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because the data used for this study were collected as part of routine data that are recorded in a commercial breeding program. Semen collections were conducted strictly in line with the Guidelines for the Care and Use of Experimental Animals established by the Ministry of Science and Technology of the People's Republic of China. All efforts were made to minimize animal suffering.

AUTHOR CONTRIBUTIONS

Conceptualization, ZW and CT; methodology, CT; validation, JY, CT; formal analysis, YH; investigation, GC, DW, and DL; resources, CT; data curation, YH and CT; writing–original draft preparation, YH and LY; writing–review and editing, YH and CT; supervision, ZW; project administration, ZW; funding acquisition, ZW All authors have read and agreed to the published version of the manuscript.

FUNDING

This study was supported by the Natural Science Foundation of Guangdong Province, China (Grant No.2018B030313011) and the Science and Technology Innovation Strategy projects of Guangdong Province, China (Grant No.2018B020203002).

ACKNOWLEDGMENTS

Thanks to all the staff at the pig core breeding farms of Wens Foodstuff Group Co., Ltd. (Guangdong, China) for record collection.

REFERENCES

Andersen, S., and Pedersen, B. (2010). Growth and Food Intake Curves for Group-Housed Gilts and Castrated Male Pigs. *Anim. Sci.* 63 (3), 457–464. doi:10.1017/s1357729800015356

Bajena, M., Kondracki, S., Iwanina, M., Wysokinska, A., and Adamiak, A. (2016). Physical Characteristics of Ejaculates Produced by Insemination Boars Depending on the Interval between Successive Ejaculate Collections. *Jcea* 17, 260–271. doi:10.5513/JCEA01/17.2.1699

Bohmanova, J., Miglior, F., Jamrozik, J., Misztal, I., and Sullivan, P. G. (2008). Comparison of Random Regression Models with Legendre Polynomials and Linear Splines for Production Traits and Somatic Cell Score of Canadian Holstein Cows. *J. Dairy Sci.* 91 (9), 3627–3638. doi:10.3168/jds.2007-0945

Brigatti, K. W. (2021). "Introduction to Concepts of Genetics and Genomics," in *Genomics of Rare Diseases*. Editors C. Gonzaga-Jauregui and J. R. Lupski (Academic Press), 1–15. doi:10.1016/b978-0-12-820140-4.00009-0

Burren, A., Joerg, H., Erbe, M., Gilmour, A. R., Witschi, U., and Schmitz-Hsu, F. (2019). Genetic Parameters for Semen Production Traits in Swiss Dairy Bulls. *Reprod. Dom Anim.* 54 (9), 1177–1181. doi:10.1111/rda.13492

Carabano, M. J., Díaz, C., Ugarte, C., and Serrano, M. (2007). Exploring the Use of Random Regression Models with Legendre Polynomials to Analyze Measures of Volume of Ejaculate in Holstein Bulls. *J. Dairy Sci.* 90 (2), 1044–1057. doi:10.3168/jds.S0022-0302(07)71591-6

Check, J. H., Epstein, R., and Long, R. (1991). Effect of Time Interval between Ejaculations on Semen Parameters. *Arch. Androl.* 27 (2), 93–95. doi:10.3109/01485019108987658

Grandjot, G., Brandt, H., and Glodek, P. J. A. T. (1997). Genetische und phänotypische Untersuchungen zu Eigenleistungs-, Sperma-und Fruchtbarkeitsmerkmalen von Besamungsebern. 1. *Mitteilung: Systematische Einflußfaktoren auf genetische Parameter* 40, 421–432.

Gredler, B., Fuerst, C., Fuerst-Waltl, B., Schwarzenbacher, H., and Sölkner, J. (2007). Genetic Parameters for Semen Production Traits in Austrian Dual-Purpose Simmental Bulls. *Reprod. Domest. Anim.* 42 (3), 326–328. doi:10.1111/j.1439-0531.2006.00778.x

Henderson, C. R., Jr. (1982). Analysis of Covariance in the Mixed Model: Higher-Level, Nonhomogeneous, and Random Regressions. *Biometrics* 38 (3), 623–640. doi:10.2307/2530044

Hill, W. G. (1999). *Advances in Methodology for Utilising Sequential Records*. Cattle Practice.

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and Theory point to Mainly Additive Genetic Variance for Complex Traits. *Plos Genet.* 4 (2), e1000008. doi:10.1371/journal.pgen.1000008

Ibáñez-Escriche, N., Forni, S., Noguera, J. L., and Varona, L. (2014). Genomic Information in Pig Breeding: Science Meets Industry Needs. *Livestock Sci.* 166, 94–100. doi:10.1016/j.livsci.2014.05.020

Knecht, D., Jankowska-Mąkosa, A., and Duziński, K. (2017). The Effect of Age, Interval Collection and Season on Selected Semen Parameters and Prediction of AI Boars Productivity. *Livestock Sci.* 201, 13–21. doi:10.1016/j.livsci.2017.04.013

Li, J., Gao, H., Madsen, P., Li, R., Liu, W., Bao, P., et al. (2020). Impact of the Order of Legendre Polynomials in Random Regression Model on Genetic Evaluation for Milk Yield in Dairy Cattle Population. *Front. Genet.* 11, 586155. doi:10.3389/fgene.2020.586155

Li, X., Jiang, B., Wang, X., Liu, X., Zhang, Q., and Chen, Y. (2019). Estimation of Genetic Parameters and Season Effects for Semen Traits in Three Pig Breeds of South China. *J. Anim. Breed. Genet.* 136 (3), 183–189. doi:10.1111/jbg.12393

Lopez Rodriguez, A., Van Soom, A., Arsenakis, I., and Maes, D. (2017). Boar Management and Semen Handling Factors Affect the Quality of Boar Extended Semen. *Porc Health Manag.* 3, 15. doi:10.1186/s40813-017-0062-5

Marques, D. B. D., Lopes, M. S., Broekhuijse, M. L. W. J., Guimarães, S. E. F., Knol, E. F., Bastiaansen, J. W. M., et al. (2017). Genetic Parameters for Semen Quality and Quantity Traits in Five Pig Lines1. *J. Anim. Sci.* 95 (10), 4251–4259. doi:10.2527/jas2017.1683

Meyer, B. K. (1999). Estimates of Genetic and Phenotypic Covariance Functions for Postweaning Growth and Mature Weight of Beef Cows. *J. Anim. Breed. Genet.* 116 (3), 181–205. doi:10.1046/j.1439-0388.1999.00193.x

Meyer, K. (2005). Random Regression Analyses Using B-Splines to Model Growth of Australian Angus Cattle. *Genet. Selection Evol.* 37 (5), 473–500. doi:10.1186/1297-9686-37-6-473

Meyer, K., and Hill, W. G. (1997). Estimation of Genetic and Phenotypic Covariance Functions for Longitudinal or "repeated" Records by Restricted Maximum Likelihood. *Livestock Prod. Sci.* 47 (3), 185–200. doi:10.1016/s0301-6226(96)01414-5

Neath, A. A., and Cavanaugh, J. E. (2011). The Bayesian Information Criterion: Background, Derivation, and Applications. *Wires Comp. Stat.* 4 (2), 199–203. doi:10.1002/wics.199

Oh, S. H., See, M. T., Long, T. E., and Galvin, J. M. (2006). Genetic Parameters for Various Random Regression Models to Describe Total Sperm Cells Per Ejaculate over the Reproductive Lifetime of Boars. *J. Anim. Sci.* 84 (3), 538–545. doi:10.2527/2006.843538x

Prakash, V., Gupta, A. K., Gupta, A., Gandhi, R. S., Singh, A., and Chakravarty, A. K. (2017). Random Regression Model with Heterogeneous Residual Variance Reduces Polynomial Order Fitted for Permanent Environmental Effect but Does Not Affect Genetic Parameters for Milk Production in Sahiwal Cattle. *Anim. Prod. Sci.* 57 (6), 1022. doi:10.1071/an15347

Rutten, S. C., Morrison, R. B., and Reicks, D. (2000). *Boar Stud Production Analysis*. Swine Health & production.

Sasaki, O., Aihara, M., Nishiura, A., and Takeda, H. (2017). Genetic Correlations between the Cumulative Pseudo-survival Rate, Milk Yield, and Somatic Cell Score during Lactation in Holstein Cattle in Japan Using a Random Regression Model. *J. Dairy Sci.* 100 (9), 7282–7294. doi:10.3168/jds.2016-12311

Schaeffer, L. R. (2004). Application of Random Regression Models in Animal Breeding. *Livestock Prod. Sci.* 86 (1-3), 35–45. doi:10.1016/s0301-6226(03)00151-9

Smital, J., Wolf, J., and Sousa, L. L. D. (2005). Estimation of Genetic Parameters of Semen Characteristics and Reproductive Traits in AI Boars. *Anim. Reprod. Sci.* 86 (1-2), 119–130. doi:10.1016/j.anireprosci.2004.05.023

Strathe, A. B., Velander, I. H., Mark, T., Ostersen, T., Hansen, C., and Kadarmideen, H. N. (2013). Genetic Parameters for Male Fertility and its Relationship to Skatole and Androstenone in Danish Landrace Boars. *J. Anim. Sci.* 91 (10), 4659–4668. doi:10.2527/jas.2013-6454

Takma, Ç., and Akbaş, Y. (2007). Estimates of Genetic Parameters for Test Day Milk Yields of a Holstein Friesian Herd in Turkey with Random Regression Models. *Arch. Anim. Breed.* 50, 327–336. doi:10.5194/aab-50-327-2007

Tesfay, H. H., Sun, Y., Li, Y., Shi, L., Fan, J., Wang, P., et al. (2020). Comparative Studies of Semen Quality Traits and Sperm Kinematic Parameters in Relation to Fertility Rate between 2 Genetic Groups of Breed Lines. *Poult. Sci.* 99 (11), 6139–6146. doi:10.1016/j.psj.2020.06.088

Wolf, J. (2009a). Genetic Correlations between Production and Semen Traits in Pig. *Animal* 3 (8), 1094–1099. doi:10.1017/S1751731109004686

Wolf, J. (2009b). Genetic Parameters for Semen Traits in AI Boars Estimated from Data on Individual Ejaculates. *Reprod. Domest. Anim.* 44 (2), 338–344. doi:10.1111/j.1439-0531.2008.01083.x

Wolf, J., and Smital, J. (2009). Effects in Genetic Evaluation for Semen Traits in Czech Large White and Czech Landrace Boars. *Czech J. Anim. ence* 54 (8), 349–358. doi:10.17221/1658-cjas

Conflict of Interest: YH, XH, DW, JY, GC, DL, ZW, and CT were employed by the company WENS Foodstuff Group Co., Ltd.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Hong, Yan, He, Wu, Ye, Cai, Liu, Wu and Tan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

REVIEW
published: 07 February 2022
doi: 10.3389/fcell.2021.803252

Strategies for High-Efficiency Mutation Using the CRISPR/Cas System

Shuying Feng[1†], Zilong Wang[1†], Aifang Li[1†], Xin Xie[1], Junjie Liu[1], Shuxuan Li[1], Yalan Li[1], Baiyan Wang[1], Lina Hu[1], Lianhe Yang[1] and Tao Guo[2]**

[1]Medical College, Henan University of Chinese Medicine, Zhengzhou, China, [2]Department of Pharmacy, Henan University of Chinese Medicine, Zhengzhou, China

Clustered regularly interspaced short palindromic repeats (CRISPR)-associated systems have revolutionized traditional gene-editing tools and are a significant tool for ameliorating gene defects. Characterized by high target specificity, extraordinary efficiency, and cost-effectiveness, CRISPR/Cas systems have displayed tremendous potential for genetic manipulation in almost any organism and cell type. Despite their numerous advantages, however, CRISPR/Cas systems have some inherent limitations, such as off-target effects, unsatisfactory efficiency of delivery, and unwanted adverse effects, thereby resulting in a desire to explore approaches to address these issues. Strategies for improving the efficiency of CRISPR/Cas-induced mutations, such as reducing off-target effects, improving the design and modification of sgRNA, optimizing the editing time and the temperature, choice of delivery system, and enrichment of sgRNA, are comprehensively described in this review. Additionally, several newly emerging approaches, including the use of Cas variants, anti-CRISPR proteins, and mutant enrichment, are discussed in detail. Furthermore, the authors provide a deep analysis of the current challenges in the utilization of CRISPR/Cas systems and the future applications of CRISPR/Cas systems in various scenarios. This review not only serves as a reference for improving the maturity of CRISPR/Cas systems but also supplies practical guidance for expanding the applicability of this technology.

Keywords: CRISPR/Cas system, optimized strategies, highly efficient, mutant, off-target effect

OPEN ACCESS

Edited by:
Dov Greenbaum,
Yale University, United States

Reviewed by:
Keiichiro Suzuki,
Osaka University, Japan
Rashmi Kaul,
International Centre for Genetic Engineering and Biotechnology, India
Mario Andrea Marchisio,
Tianjin University, China

****Correspondence:***
Lianhe Yang
13598420711@139.com
Tao Guo
gt010010@163.com

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Epigenomics and Epigenetics, a section of the journal Frontiers in Cell and Developmental Biology

Received: *27 October 2021*
Accepted: *22 December 2021*
Published: *07 February 2022*

Citation:
Feng S, Wang Z, Li A, Xie X, Liu J, Li S, Li Y, Wang B, Hu L, Yang L and Guo T (2022) Strategies for High-Efficiency Mutation Using the CRISPR/Cas System. Front. Cell Dev. Biol. 9:803252. doi: 10.3389/fcell.2021.803252

1 INTRODUCTION

In aquatic systems, host–pathogen interactions are meaningful in the ecology and evolution of all organisms. These interactions are often characterized by a strong immune defense between prokaryotic cells (archaea) and viruses, leading to their co-evolution (England and Whitaker, 2013). The strong immune defense mechanism utilized by these organisms is known as the clustered regulatory interspaced short palindromic repeats (CRISPR) system, which is used in prokaryotes to combat a viral infection. Earlier reports of CRISPR/Cas systems report three different types: I, II, and III (Makarova et al., 2011). Each type of system is characterized by a signature protein(s). The most common type, type II CRISPR/Cas9 system, mediates the immune response in three stages as follows: (1) adaption, (2) expression, and (3) interference (Makarova et al., 2020). In the adaption stage, DNA fragments of invading plasmids or phages (termed protospacers) are incorporated into the host CRISPR locus as spacers in the form of CRISPR RNA (crRNA) repeats. In the expression stage, the

precursor CRISPR RNA (pre-crRNA) molecules are processed by expressed Cas proteins and cofactors into short, mature crRNA. Next, in the interference stage, the Cas9 protein recognizes and targets the crRNA, silencing the foreign sequences (Gasiunas et al., 2012; Janik et al., 2020). Single-guide RNA (sgRNA) synthesized by crRNA and tracrRNA then guides the Cas protein to generate double-strand breaks (DSBs) three base pairs upstream from the protospacer adjacent motifs (PAM) (Jinek et al., 2012). Through this mechanism, CRISPR/Cas systems can also serve as a precise gene-editing tool for genetic manipulation.

So far, the CRISPR/Cas systems have been divided into six types (types I–VI), in which type II-A (CRISPR-Cas9), type V-A (CRISPR-Cas12a or Cpf1), and (CRISPR-Cas12b or C2c1) have been most widely studied (Adli, 2018; Yao et al., 2018a). More than 10 different CRISPR/Cas proteins have been repurposed for genome editing. Among them, some of the most recently discovered Cas proteins are hotspots for research, such as the Cas12a proteins from *Acidaminococcus* sp. (AsCas12a) and Lachnospiraceae bacteria (LbCas12a). Beyond Cas proteins, optimization of CRISPR systems has been thoroughly studied, including sgRNA design, cell enrichment, editing conditions, *etc*. With the rapid development and progress of gene editing technology, CRISPR systems have been shown to be powerful and highly efficient gene-editing tools in various fields. Through numerous experiments in model and non-model organisms (Oh et al., 2010), these systems have been utilized to reveal cancer mechanisms (Sottnik et al., 2021), define gene function and phenotypes (Johansen et al., 2017), and treat human diseases (Torre et al., 2021).

As to traditional editing tools, zinc finger nucleases and transcription activator-like effector nucleases (TALENs) have overwhelmingly contributed to developments in biomedical research and application (Urnov et al., 2005). Their application is greatly limited, however, due to limitations such as high cost, low efficiency, and low throughput targeting (Batool et al., 2021). In contrast, the CRISPR technology has some unique advantages, including targeted editing of multiple genomic sites (Zhang and Showalter, 2020), fast generation of mutants (Zhang and Showalter, 2020), and accessible sgRNA design (Xu et al., 2020b). These advantages have led to a surge in CRISPR applications in various fields, such as agriculture (Zheng et al., 2019), animal husbandry (Liu et al., 2020b), chemical fields (Liu et al., 2021b), materiology (Demirer et al., 2021), *etc*. Although the framework of the structures and functions of CRISPR/Cas systems has been built, there are still several challenges in this system (Wang et al., 2016), including off-target effects (Coelho et al., 2020), variable efficiency (Jin et al., 2020), requirement of PAM and sgRNA (Heussler et al., 2015; Cameron et al., 2017), and inactive mutants (Ren et al., 2019). This review proposes some strategies to overcome these issues by reducing off-target effects, improving the repair efficiency of the homology-directed repair (HDR) pathway, choosing the optimal delivery system, and utilizing variants of Cas proteins. Additionally, regulation of nuclease-dead mutants of Cas9, anti-CRISPR (Acrs) protein application, and enrichment of cells and sgRNA may be effective strategies for the efficacy of CRISPR/Cas systems.

2 STRATEGIES FOR REDUCING OFF-TARGET EFFECTS

Presently, off-target effects in CRISPR/Cas systems are a major issue for gene editing. Whether the Cas protein is off- or on-target to a PAM site is mainly determined by the sgRNA, Cas proteins, ribonucleoprotein (RNP) concentration, as well as other factors, such as editing temperature and action time. The off-target cleavage of CRISPR/Cas systems often originates from the unsuccessful design or modification of gRNA, low specificity of Cas proteins, or excessive and prolonged expression of CRISPR/Cas9. Accordingly, various strategies are proposed to overcome these issues. Additionally, methods for sgRNA selection with off-target predictions have been established, such as PEM-seq (Yin et al., 2019b), CRISPR-PLANT v2 (Minkenberg et al., 2019), and CRISPR-GE (Xie et al., 2017), which avoid a waste of manpower and material resources and improve editing efficiency.

2.1 Reasonable Design and Modification of sgRNA

In CRISPR/Cas systems, the binding of sgRNA to the PAM site is a critical step in gene editing (**Figure 1A**). An unsuccessful design of sgRNA will result in lower specificity and higher miss rate (Doench et al., 2016). To avoid this, sgRNA must be accurately designed using computational tools (Liu et al., 2020a), such as CRISPR-P 2.0 (Liu et al., 2017a), E-CRISP (Heigwer et al., 2014), and CasFinder (Abby et al., 2014). On the basis of rational design, further modification of sgRNA can improve the specificity of RNA-guided Cas9 by truncation or addition of nucleotides to the 5′ or 3′ end (Pattanayak et al., 2013; Lin et al., 2014b). The 5′ end-truncated sgRNAs (2-3 bp) considerably reduce off-target mutations, but with the same on-target mutation efficiency as the full-length sequence (Fu et al., 2014). By decreasing the binding affinity of the sgRNA, the binding stringency of Cas9 to the target sequence was increased, and the off-target effect was reduced (Fu et al., 2014). Since truncated sgRNAs can reduce the off-target effect of paired Cas9 nickases without compromising the efficiency of on-target genome editing, their combination results in a much greater target specificity (Guilinger et al., 2014). Contrarily, the 3′ end-truncated sgRNA or 5′ end-added sgRNA (-GG) can decrease the on-target activity. Meanwhile, if they consist of 16 nucleotides or fewer, truncated sgRNAs exhibit lower or undetectable activity compared to matched full-length sgRNAs. Thus, at least 17-nucleotide sgRNAs are required for the CRISPR/Cas9 system to be active during gene editing. Due to the disadvantages of the traditional enzymatic preparation of sgRNAs, such as complexity, time consumption, and safety concerns, the direct chemical synthesis of sgRNAs has been widely accepted, with high sgRNA stability and low off-target effect. Recently, a potential strategy has been reported to reduce off-target editing by DNA–RNA chimera (Yin et al., 2018a). Using the Cas9–sgRNA complex as a guide, the 5′- and 3′-DNA-replaced crRNA enables more efficient genome editing—for example, replacing the crRNA with 10 DNA nucleotides could provide the same level of off-target site indel formation as the truncated sgRNA. Additionally, the synthesis cost of DNA bases is much lower (10-fold cheaper) than

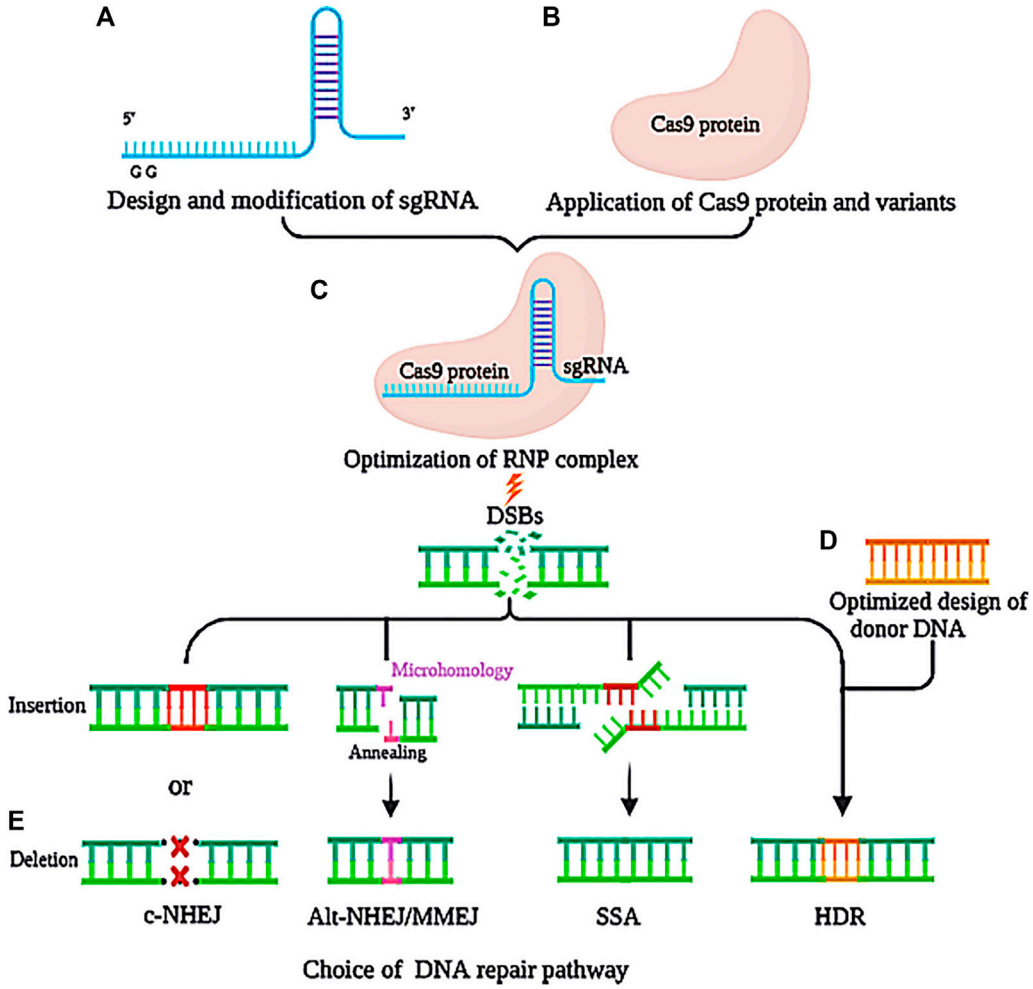


FIGURE 1 | Optimization strategies of different steps of the CRISPR/Cas system. **(A)** The sgRNA sequence was optimally designed and modified by truncation or addition of 3' or 5' end of sgRNA, DNA-RNA chimera, etc. **(B)** Cas9 protein was optimized for concentration, temperature, and time, as well as application of variants. **(C)** Optimization of the RNP complex was conducted for proportion, function conditions, and transformation methods. **(D)** Donor DNA was optimized for design of the DNA template, proximity to CRISPR components, and choice of high-efficient delivery system. **(E)** DNA repair pathway was optimized with inhibition of the NHEJ pathway, enhancement of the HDR pathway, and modification of other pathways.

that of native crRNA. In light of this, the DNA–RNA chimera could provide a novel approach to reduce the cost and off-target effect of CRISPR/Cas systems.

2.2 Cas Variant Application

So far, several highly specific Cas9 variants have been identified, including eSpCas9 (Slaymaker et al., 2016), SpCas9-HF1 (Kleinstiver et al., 2016), HypaCas9 (Chen et al., 2017), xCas9 (Legut et al., 2020), Sniper-Cas9 (Lee et al., 2018), evo Cas9 (evolved Cas9) (Casini et al., 2018), HiFiCas9 (Vakulskas et al., 2018), and HeFSpCas9 (Kulcsar et al., 2017). The main mechanism and characteristics of each variant are comprehensively summarized in **Table 1**. Among these mechanisms, the most common cause of alteration in Cas9 function is amino acid substitution of the critical domain. Due to the minimum binding energy required to introduce DSBs into the genome, non-specific interactions between Cas9 and target DNA were reduced by decreasing the excess energy of wild-type SpCas9. As shown in **Table 1**, although they exhibit greater target specificity, each variant has its own limitations, such as low activity (Slaymaker et al., 2016), scope limitation (Kleinstiver et al., 2016), strict dependency on a PAM site (Legut et al., 2020), *etc*. Future studies should be conducted to increase the efficiency of genome editing using Cas9 variants. For a given target sequence, the optimal variant should be selected based on a comparison of activity, specificity, and PAM compatibility. By comparing 13 SpCas9 variants, the results demonstrated that the overall activity order of high-fidelity variants could be ranked as SpCas9 ≥ Sniper-Cas9 > eSpCas9 (1.1) > SpCas9-HF1 > HypaCas9 ≈ xCas9 > evoCas9, whereas their overall specificity could be ranked as evoCas9 > HypaCas9 ≥ SpCas9-HF1 ≈ eSpCas9 (1.1) > xCas9 > Sniper-Cas9 > SpCas9 (Kim et al., 2020b). Using established computational models, these results provide guidance for the selection of Cas9 variants and offer a

TABLE 1 | The features of different Cas protein variants.

Cas variants	Description	Mechanisms	Target sequence	Average indel frequency	Advantages	Limitation	Reference
xCas9	Generation of xCas9 by "phage-assisted continuous evolution (PACE)" method	Closing to PAM or the DNA-sgRNA interface refines the DNA-RNA contact region	Refer to the three lentiviral libraries (Kim et al., 2020a)	32%	Improve the target specificity and extend the target range, present a higher DNA specificity and lower off-target activity	Profoundly diminished of xCas9 activity at target sites with NGH PAM	Nishimasu et al. (2018)
SpCas9-HF1	The quadruple substitution variant (N497A/R661A/Q695A/Q926A) of wild-type SpCas9	Reduce the rate of DNA cleavage but have no effect on the rate of DNA reversion and release	Refer to the three lentiviral libraries (Kim et al., 2020b)	34%	A high-fidelity variant retains on-target activities comparable to wild-type SpCas9 with >85% of sgRNAs	The unclear mechanism of target discrimination and fidelity needs to be further improved	Kleinstiver et al. (2016)
eSpCas9	SpCas9 mutants consisting of individual alanine substitutions at 32 positively charged residues within the nt-groove	Neutralization of positively charged residues within this non-target strand and then weaken non-target strand binding and encourage re-hybridization between the target and non-target DNA strands	Refer to the three lentiviral libraries (Kim et al., 2020a)	40%	Decrease the off-target activities and maintain efficient on-target editing	The unclear mechanism of target discrimination and fidelity needs to be further improved	Slaymaker et al. (2016)
HypaCas9	The quadruple substitution variant (N692A/M694A/Q695A/H698A) of wild-type SpCas9	The quadruple substitutions in the REC3 domain of wild-type SpCas9	Refer to the three lentiviral libraries (Kim et al., 2020b)	30%	Higher genome-wide fidelity without affecting the on-target genome editing	Not mentioned	Chen et al. (2017)
Cas9n	Inactivating of HNH or RuvC nuclease domains	Use dual-RNAs for site-specific DNA cleavage	Two human genes: C4BPB and CCR5	75 and 60%	Greater target specificity	Rational design of sgRNAs on the plus and minus strands within a limited distance	Trevino and Zhang (2014)
Sniper-Cas9	The quadruple substitution variant (F539S/M763I/K890N) of wild-type SpCas9	Weakening non-specific interactions between RNP and its substrate DNA	Refer to the three lentiviral libraries (Kim et al., 2020a)	46%	Retain WT-level on-target activity with diminished off-target effect	Not mentioned	Lee et al. (2018)
evoCas9	The quadruple substitution variant (M495V/Y515N/K526E/R661Q) of wild-type SpCas9	Weakening non-specific interactions between RNP and its substrate DNA	Refer to the three lentiviral libraries (Kim et al., 2020b)	15%	Retain WT level on-target activity with diminished off-target effect	Not mentioned	Casini et al. (2018)
HiFiCas9	The quadruple substitution variant (R691A) of wild-type SpCas9	Weakening non-specific interactions between RNP and its substrate DNA	Five human genes: HBB, IL2RG, CCR5, HEXB, and TRAC	Similar to WT Cas9	Retain WT level on-target activity with diminished off-target effect	Not mentioned	Vakulskas et al. (2018)
HeFSpCas9	The quadruple substitution variant (N497A/R661A/K846A/Q926A/K1003A/R1060A) of wild-type SpCas9	Combinations of mutation domain from both eSpCas9 and SpCas9-HF1	Not shown	Not shown	Retain WT level on-target activity with diminished off-target effect	Not mention	Kulcsar et al. (2017)

more effective exploration of variants for future research (**Figure 1B**).

2.3 Determination of the Optimal RNP Concentration

In general, the specificity and the activity of enzymes are often highly dependent on reaction conditions. RNP delivery produces at least twofold more colonies than plasmid transfection does (Kim et al., 2014). In the CRISPR/Cas9 system, RNP concentration plays a decisive role in both specificity and activity. After delivery to cells, RNPs almost immediately cleave chromosomal DNA and then degrade rapidly. With a high RNP concentration, the off-target effects of a CRISPR/Cas system may be amplified (**Figure 1C**). Meanwhile, a low RNP concentration leads to a reduction of on-target cleavage efficiency. Therefore, a suitable concentration of RNP is of paramount importance to

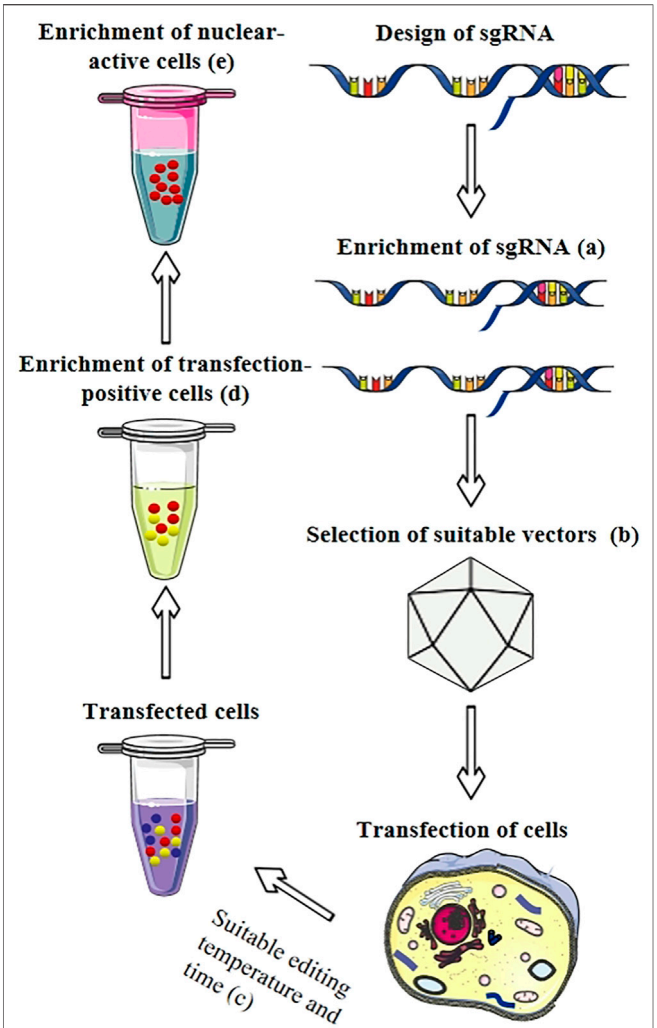


FIGURE 2 | Enrichment strategies for sgRNA and mutants in CRISPR/Cas systems. **(A)** The sgRNA was enriched with by PCR or CRISPR. **(B)** The optimal vector was selected according to the different host cell and other factors. **(C)** Optimization of editing time and temperature was conducted through application of nuclease-dead mutants of Cas9 and anti-CRISPR proteins, heat stress method. **(D)** Transfection-positive cells were enriched based on fluorescent proteins, antibiotic-resistance genes, and cell-surface antigens. **(E)** Nuclear-active cells were enriched using NHEJ-based and SSA-based surrogate reporters.

minimize nonspecific cleavage (**Figure 1E**). This can be achieved by using either low concentrations of plasmids or different promoters. The former method directly reduces RNP transcription, while the latter alters the 5′-untranslated region of the target sequence, ultimately affecting translation efficiency (Hsu et al., 2013). Therefore, extensive measurements should be performed with consideration of both Cas9 activity and specificity. Compared with typical RNP concentrations, on-target activity will inevitably be inhibited to some extent. By modifying Cas9 and sgRNA instead (**Figure 2A**), the intrinsic specificity of Cas9 can be improved without sacrificing cleavage efficiency.

2.4 Suitable Editing Time and Temperature

During gene editing, the efficacy, efficiency, and accuracy of CRISPR systems are often limited by temperature and time (**Figure 2C**). Studies have shown that a longer editing time of the Cas proteins in cells increases the off-target effects and negatively impacts outcomes (Ruan et al., 2017). Therefore, reducing the action time or overexpression of Cas proteins enhances the efficiency of gene editing. Early reports have utilized tissue culture-free systems (Manghwar et al., 2020), DNA-free systems (Kim et al., 2020a), and transient systems (Canto, 2016) to overcome these limitations. Now, however, anti-CRISPR proteins and nuclease-dead Cas proteins can be used to address these limitations. Temperature has been shown to affect Cas protein activity, but the findings are inconsistent. Hoyland-Kroghsbo *et al.* (2018) reported that a lower temperature is more effective than a higher one in *P. aeruginosa* PA14 due to the greater stability of the CRISPR/Cas complex. Additionally, low temperatures may enhance interference and adaptation by increasing the annealing efficiency of the crRNA to its target. Another report has shown that exposure to heat stress results in a greater amount of targeted mutations than with exposure to standard temperature (22°C) (LeBlanc et al., 2018). This is likely due to the fact that the activities of Cas9 and Cas12a at 37 and 34°C, respectively, are higher than at 28°C, and the expression level of sgRNA is raised at 39°C (Xiang et al., 2017). The mechanism of the effect of temperature on CRISPR/Cas systems is still unclear and should be further elucidated.

2.5 Application of Anti-CRISPR Proteins

In the course of long-term antagonism between bacteria and phages, the phages have evolved Acr proteins to evade CRISPR/Cas-mediated immunity. Up to now, a total of 44 Acr proteins have been identified and named (Zhang et al., 2019a). Within the CRISPR/Cas system subtypes, these Acr proteins are present in types I, II, and V, but not in other subtypes—for instance, FnCas9 (Green and Hu, 2017) and C2c2 (Zhang et al., 2019a) Acr proteins were not found in the subtype II-B CRISPR/Cas9 system and the type VI CRISPR/Cas13a system, respectively. In view of this, the identification and characterization of a novel Acr is a key focus for future studies. As natural inhibitors, Acrs protect the host genome from destruction by inhibiting Cas nuclease activity. This inhibition can be achieved through the following three mechanisms: (1) inhibition of Cas9 binding to DNA (Malone et al., 2020), (2) interference of Cas9 binding to gRNA (Harrington et al., 2017), and (3) blocking the activity of Cas9 (Harrington et al., 2017). Currently, only AcrIIA2 and AcrIIA4 have been utilized as tools to reduce off-target effects. Through competition with the PAM site and/or other Cas9 sequences, these Acr proteins block the cleavage activity of Cas9, preventing the excessive and prolonged expression of CRISPR/Cas9 and thus decreasing off-target effects (Hoffmann et al., 2019; Liu et al., 2019). Acrs appear to be a new agent to improve the accuracy and safety of CRISPR-based therapies. Other special functions of Acr proteins also deserve attention—for example, AcrII-C3 can precisely regulate gene expression with dCas9-based tools, which is very helpful for the development of versatile genome engineering modulators (Liu et al., 2018c). The optogenetic-controlled AcrIIA4 enables light-

mediated genome and epigenome editing. By inserting the AsLOV2 domain into the most C-terminal loop of AcrIIA4, the protein can switch the CRISPR-Cas9 activity according to light/dark conditions (Bubeck et al., 2018). The mechanisms of other Acrs, such as AcrIIA5-10 and AcrVA2-3, have yet to be described (Zhang et al., 2019a).

3 STRATEGIES TO IMPROVE THE EFFICIENCY OF THE HDR REPAIR PATHWAY

After Cas9 nuclease cleavage, DSBs can be repaired in a host through at least one of two different pathways: nonhomologous end joining (NHEJ)/canonical NHEJ (c-NHEJ) and HDR (Ghezraoui et al., 2014; Sander and Joung, 2014). While c-NHEJ is the predominant approach, due to its speed and high efficiency, it is also prone to error because of leading uncertain inserts or deletions (indels). Indels contribute to the generation of a targeted knockout during cell repair (Pawelczak et al., 2018). The HDR pathway enables accurate genome editing in a variety of manners, such as gene knock-in, knockout, replacement, and point mutations (Platt et al., 2014; Zuo et al., 2017; Vakulskas et al., 2018; Lu et al., 2020). However, due to competition with the NHEJ pathway, the HDR pathway tends to be less efficient (Liu et al., 2018a). Given this, different approaches have been established to improve the repair efficiency of the HDR pathway, including inhibition of the NHEJ pathway (Maruyama et al., 2015), regulation of HDR-related factors (Paulsen et al., 2017), cell cycle synchronization (Ferrari et al., 2020), optimal design of the donor DNA template (Renaud et al., 2016), and optimizing the proximity of the CRISPR component and donor DNA template (Ma et al., 2017a). These strategies are discussed in detail in the following paragraphs.

3.1 Inhibition of Nonhomologous End Joining Pathway

In theory, because of the competition between the two repair pathways, the efficiency of the HDR pathway can be boosted by inhibiting key factors of the NHEJ pathway. Among different inhibitors of the NHEJ pathway, SCR7 is a key factor that interferes with the affinity of DNA ligase IV to DSBs (Srivastava et al., 2012; Chu et al., 2015; Maruyama et al., 2015; Li et al., 2017; Shao et al., 2017; Hu et al., 2018). Maruyama *et al.* (2015) reported that using SCR7 increased the efficiency of HDR-mediated genome editing by up to 19-fold with the most significant enhancement effect, primarily due to co-injection of the CRISPR-Cas9 constructs with SCR7 into zygotes rather than other cells. The combination of SCR7 with other factors could significantly improve the efficiency of the HDR pathway by either downregulating KU expression (Chu et al., 2015), optimizing the donor template (Hu et al., 2018), or upregulating Rad52 expression and other small molecules (Li et al., 2017; Shao et al., 2017). Among these methods, the efficiency of the HDR pathway using Rad52 combined with SCR7 is the highest, reaching up to 40% (Shao et al., 2017). However, the effect of SCR7 in enhancing the HDR pathway remains controversial at present (Greco et al., 2016), with some reporting that embryonic stem cells tend to occur intrinsically HDR incident, suggesting that the effect of SCR7 is likely uncertain (Yang et al., 2020). Therefore, the effect of SCR7 in increasing HDR efficiency needs to be further explored. Additionally, the use of SCR7 should be extended to other fields as well instead of being limited to human- and mammalian-related diseases only, whether used alone or in combination with other medicines (Manjunath et al., 2021).

Aside from SCR7, other approaches also improved the efficiency of the HDR pathway—for instance, by downregulating KU protein expression *via* siRNA silencing, the frequency of the HDR pathway can be increased at least slightly (Li et al., 2018a). This method raised the affinity of downstream NHEJ enzymatic components by attaching to DNA termini upstream of the NHEJ pathway (Mateos-Gomez et al., 2017). The combination of multiple inhibitors can further increase the inhibitory effect on the NHEJ pathway—for example, M3814 combined with trichostatin A inhibits the NHEJ pathway 3-fold (Fu et al., 2021). NU7441 and KU-0060648, inhibitors of a key NHEJ pathway factor, DNA-PK, caused a 2-fold increase in HDR efficiency in HEK-293T cells (Robert et al., 2015). Due to the fact that the NHEJ pathway is crucial for the stabilization of the genome, excessive inhibition of the NHEJ pathway may eventually lead to the accumulation of unrepaired DSBs in cells, inducing cell death or embryonic lethality (Beumer et al., 2013). Therefore, it is worth noting that the safety of these inhibitors needs to be carefully evaluated in future work.

3.2 Regulation of HDR-Related Factors

Alternatively, compared with inhibition of the NHEJ pathway, direct regulation of HDR-related factors can result in more precise editing and improve the efficiency of the HDR pathway. Several HDR-related factors have been well studied, including Rad51 (Ma et al., 2020), Rad52 (Shao et al., 2017), C-terminal-binding protein interacting protein (CtIP) (Charpentier et al., 2018), and RS-1 (Song et al., 2016). All of these factors enhance a link in the HDR pathway, thereby boosting repair efficiency. Among these, the overexpression of Rad proteins and the application of CtIP result in superior outcomes. In the HDR pathway, Rad51 proteins search for a DNA donor template to perform reconstitution through the formation of filaments on the DNA (Symington, 2014). As a back-up for Rad51, Rad52 is involved in the formation of Rad51 (Lok and Powell, 2012). When Rad52 fuses with any other factor or factors, HDR efficiency can be boosted at least 2- to 7-folds and sometimes much more (Paulsen et al., 2017; Shao et al., 2017; Tran et al., 2019)—for example, Rad52 fused with Cas9 yielded an approximately 3-fold increase in the efficiency of the HDR pathway, while Rad52 combined with SCR7 increased the HDR efficiency by about 40% (Shao et al., 2017). In the early stage of the HDR pathway, a key protein, CtIP, initiates the resection process and then creates 3′ single-stranded overhangs with exonuclease 1 and bloom syndrome protein complex (Symington, 2014). When combined with Cas9 or the MS2 system, CtIP can shift the ratio of the HDR/NHEJ pathway activities by a factor of 14.9 (Tran et al., 2019). A minimal N-terminal fragment of CtIP can also be used as an HDR enhancer, which is sufficient to stimulate the HDR pathway

and improve repair efficiency by at least 2-fold (Charpentier et al., 2018). Other small molecules, such as L755507 and resveratrol, can also increase HDR efficiency 2- to 3-fold (Li et al., 2017). Almost all key factors of the HDR pathway were upregulated after treatment with the small molecules mentioned above, but a high concentration of resveratrol (more than 50 μM) resulted in severe cytotoxicity, significantly reducing cell viability and slightly upregulating the expression of the NHEJ factor. Therefore, the application of these factors needs to be further explored to improve the efficiency of the HDR pathway.

The selection of the NHEJ or HDR pathway plays a pivotal role in DNA repair. As an inhibitor of the HDR pathway, the tumor suppressor p53-binding protein 1 (53BP1) blocks DNA end resection and simultaneously inhibits BRCA1 recruitment to DSB sites (Panier and Boulton, 2014; Zimmermann and de Lange, 2014). By means of fusing, inhibiting, and binding 53BP1, HDR repair frequency can be increased from 20 to 86% (Paulsen et al., 2017; Canny et al., 2018; Jayavaradhan et al., 2019). Compared with a control, the correction frequency of the HDR pathway increased by nearly 20% when RAD52 was ectopically co-expressed with dominant-negative 53BP1 (dn53BP1) (Paulsen et al., 2017). Interestingly, dn53BP1 itself did not improve HDR efficiency unless combined with Rad52, suggesting that this fragment may not effectively promote the HDR pathway. To avoid the unwanted effects of global NHEJ inhibition, the fusion of DN1S and dn53BP1 significantly blocked NHEJ events locally while improving the correction frequency of HDR up to 86% (Jayavaradhan et al., 2019). This effect was likely due to the restrictively suppressive effect of dn53BP1 around the DSB site, which makes the CRISPR-Cas9-DN1S system a far more effective and stable approach in clinical treatments with high HDR frequency. Inhibition of 53BP1 is another indispensable strategy for regulating different repair pathways—for instance, utilization of 53BP1 inhibitor (i53) resulted in up to a 5.6-fold gene conversion and demonstrated an effective impact on the conversion mediated by single-stranded oligodeoxynucleotides (ssODN) compared to canonical modality double-stranded DNA (dsDNA) donors (Canny et al., 2018). However, the unknown toxicities or adverse incidents of i53 need to be carefully considered before its use. Moreover, the direct binding of related proteins to 53BP1 is a new target for enhancing the HDR pathway. A typical example is the TIRR protein, which acts similar to i53 and provides strong suppression by binding directly to 53BP1 (Anuchina et al., 2020). Since the function of TIRR is unclear, more studies should focus on its mechanism and the potential effects on the HDR pathway. TIRR may hold promise as a new target for enhancing the HDR pathway in genome editing.

3.3 Synchronization of Cas9 Activity and HDR-Active Cell Cycle

Since HDR repair activity is restricted to the S and G2 phases of the cell cycle, synchronizing cells in these phases can significantly enhance the repair activity. In terms of a single aspect of cell cycle synchronization, chemicals that maintain the cell cycle in the S and G2 phases containing nocodazole (Yiangou et al., 2019), ABT-751 (Yang et al., 2016), and RO-3306 (Sunada et al., 2021) have been commonly used in CRISPR systems and increased the HDR efficiency by a factor of 1.38–6 (Lin et al., 2014a; Yang et al., 2016; Wienert et al., 2020). ABT-751 and nocodazole arrest the cells in the G2/M phase by inhibiting microtubule polymerization (Vasquez et al., 1997; Hande et al., 2006). Meanwhile, RO-3306 can transiently arrest cells at the transition from G2 to M phase by inhibiting the CDK1 function, enriching the number of cells in the S and G2 phases (Vassilev, 2006). Recently, Lomova *et al.* reported that the transient suppression of Cas9 activity and synchronization of the HDR-active cell cycle may have a prominent effect on the HDR pathway. RO-3306 combined with Cas9, which nuclease activity is reduced in the G1 phase, can improve the HDR/NHEJ ratio 20-fold, thereby limiting unnecessary NHEJ events (Lomova et al., 2019). In addition, the timed delivery of pre-assembled Cas9 RNP and chemical synchronization agents can also enhance the HDR rates by up to 38% (Lin et al., 2014a). Thus, synchronizing the cell cycle paired with controlled timing of Cas9 activity might be more practical and safer than synchronizing the cell cycle alone. In conclusion, more efficient synchronization agents for *in vitro* application need to be further studied. Agents with lower toxicity should also be explored for *in vivo* application in subsequent research. More importantly, there is no doubt that the combination of multiple technologies, such as controlled timing of Cas9 activity and cell cycle synchronization, will result in better outcomes.

3.4 Increasing the Proximity of CRISPR Components to the Donor DNA Template

At the time of DNA cleavage, if the donor DNA template is in closer proximity to the CRISPR components or has a higher concentration in the nucleus, the efficiency of the HDR pathway can be significantly increased (Devkota, 2018). Based on this hypothesis, many studies have demonstrated its feasibility and potential value in clinical applications (Liang et al., 2017). By attaching the donor DNA template to modified sgRNA, a S1mplex strategy improves the enhancement of the HDR pathway. In this system, the modified S1m-sgRNA adds an aptamer, which binds the streptavidin protein. Biotinylated ssODN linked to the streptavidin then increases proximity. Through this powerful S1mplex strategy, the ratio of HDR increased 18-fold compared with the unlinked components (Carlson-Stevermer et al., 2017). By virtue of the affinity between avidin and biotin, Ma *et al.* devised a Cas9 variant that was fused to avidin *via* a flexible linker and bridged with biotin-modified ssDNA to increase the proximity. This system achieved ~20% HDR frequency in mouse embryos (Ma et al., 2017a). HUH endonuclease is a bridge that is also capable of forming robust covalent bindings with unmodified donor DNA templates (Lovendahl et al., 2017). Utilizing this convenient technique could create a stable Cas9 RNP–ssODN complex (Aird et al., 2018). Additionally, Natasa *et al.* linked ssODN to Cas9 through SNAP-tag technology, allowing O6-benzylguanine-labeled ssODN to covalently bind to SNAP-tag fusion proteins (Savic et al., 2018). Both SNAP-tag and HUH-tag enable the

spatio-temporal co-localization of the donor templates at DSBs, thus achieving 24- and 30-fold enhancement of HDR efficiency, respectively. In addition to ssODN attachment to sgRNA in the S1mplex system, other experiments use a variety of ways to attach ssODN to Cas9 protein. Among them, the HUH-tag strategy displayed a more promising prospect due to the superior ease of use and lower cost associated with modifying donor DNA. As noted above, Cas9 RNP complexes can connect with donor DNA templates through a variety of chemical modifications, all of which significantly enhance the transient expression of the HDR pathway. Furthermore, Cas9 and sgRNA delivered in the RNP format also exhibited a lower frequency of unwanted mutations and off-target effects (Svitashev et al., 2016), suggesting that the RNP format may be a promising approach in the broad field of gene editing.

3.5 Selection of the Donor DNA Template

To a considerable extent, the repair efficiency of the HDR pathway depends on the selection of donor DNA, including DNA modality, length, and flanking homologous sequences (Song and Stieger, 2017; Renaud et al., 2016). Generally, the modality of a DNA template can be divided into three forms: plasmid, ssODN, and linear dsDNA (**Figure 1D**). As the most common form of genetic material, circular plasmids are widely used in gene editing but will not be discussed in this review (Bosch et al., 2020; Sondergaard et al., 2020; Yoshimi et al., 2021). Compared with other donor DNA templates, ssODNs have the advantages of ease of design, lower time investment, less illegitimate random integration (introducing less than 200 nucleotides), and site-specific modification for precise editing (Yang et al., 2013; Miura et al., 2018). By comparing the modification efficiency of CRISPR/mRNA and CRISPR/nuclease for a target gene, the results have shown that the 36-nt length of ssODN with the CRISPR/nuclease form achieved the optimal condition for modification of the target gene, with a slight advantage over the CRISPR/mRNA approach (Kumita et al., 2019). Currently, ssODNs have become a routine editing tool both *in vitro* and *in vivo*, especially in multigene knock-in experiments (Yoshimi et al., 2016; Miura et al., 2018; Lim et al., 2020). For large sequence DNA modifications, linear dsDNA templates (up to 11 kb) were commonly used for CRISPR systems with homology arms of 500–800 bp (Yang et al., 2013). The targeted integration of linearized dsDNA–CRISPR can increase the knock-in efficiency 12-fold by injecting PCR-amplified donor DNA, Cas9 mRNA, and sgRNA (Yao et al., 2018b). Therefore, optimal editing outcomes can be obtained by selecting the suitable donor DNA modality according to experimental requirements.

4 SELECTION OF A HIGHLY EFFICIENT DELIVERY SYSTEM

So far, numerous delivery systems to deliver drugs and genes have been developed (**Figure 2B**). In this section, we have selected the current delivery systems with high delivery efficiency, potential for development, high biological safety, and strong tissue specificity for overview. According to their biological characteristics, they can be classified as either bioactive or abiotic. In bioactive systems, common CRISPR delivery systems contain viral vectors (Jarrett et al., 2018; Boucher et al., 2020; Lee et al., 2021), extracellular vehicles (Yao et al., 2021), cell-penetrating peptides (CPPs) (Ramakrishna et al., 2014b), or lipid nanoparticles (Cheng et al., 2020). In abiotic systems, gold nanomaterials (Wang et al., 2018), polymers (Lv et al., 2018), and graphene oxide (Yue et al., 2018) had a better effect on CRISPR system delivery. Several prominent reviews have comprehensively described the mechanisms, efficiency, challenges, and future directions for each of these systems (Yin et al., 2017; Li et al., 2018b; Glass et al., 2018; Wilbie et al., 2019; Zhang et al., 2021). The current status of these delivery systems will be exposited in the following paragraphs. Conventional physical delivery methods, such as electroporation (Shi et al., 2018), microfluidics (DiTommaso et al., 2018), and microinjection (Xu, 2019), possess unique advantages, including high local tissue transfection efficiency and extensive cellular adaptability (Mashel et al., 2020). They are not good candidates for this review, however, as they can also cause cell damage and potentially substantial cell death. Therefore, we did not repeat their descriptions in this article.

4.1 Bioactive Delivery Systems

4.1.1 Viral Vectors

In recent years, viruses have been represented as an essential and powerful tool for CRISPR due to their efficient gene delivery and long-term stable transgenic expression (Heckl et al., 2014). The most commonly utilized viral vectors are derived from adeno-associated virus (AAV) (Jarrett et al., 2018), lentivirus (LV) (Lee et al., 2021), adenovirus (Boucher et al., 2020), and baculovirus (Yin et al., 2021). These viral vectors have been widely used to deliver CRISPR/Cas9 elements for remedying genetic defects, like hearing loss (Omichi et al., 2019), neurological disorders (Pena et al., 2020), muscular dystrophies (Crudele and Chamberlain, 2019), and cystic fibrosis lung disease (Wold and Toth, 2013; Hart and Harrison, 2017). Several excellent reviews concerning different aspects of viral vectors for CRISPR-based genome editing have been published, covering topics such as viral mechanisms (Xu et al., 2019), viral vector application (Song et al., 2021), and viral vector progress (DiCarlo et al., 2017). Although highly efficient, viral vectors are presently hindered by their inherent disadvantages of carcinogenesis, insertion size limitation, immune response, genotoxicity, cytotoxicity, and difficulties of large-scale production (Matrai et al., 2010; Kotterman et al., 2015; Chen and Goncalves, 2016; Chen et al., 2020a; Shirley et al., 2020). These viral vectors have been improved in other aspects, such as pseudotyped LV and dual-AAV systems. When delivering cargo into cells, LVs need to interact with a cellular receptor to trigger the fusion of the viral envelope with the cell membrane. The envelope glycoprotein on the LV surface is exchanged with a heterologous glycoprotein in a process known as pseudotyping. Pseudotyped LVs consist of virus particles bearing glycoproteins derived from other enveloped viruses. Thus far, a variety of viral glycoproteins, including vesicular stomatitis virus (Liu et al., 2017c; Sena-

Esteves and Gao, 2018; Rust et al., 2020), baboon endogenous retrovirus (Belot et al., 2019), and feline endogenous retrovirus (Zucchelli et al., 2017), have been incorporated into LVs to improve their infectivity and confer a more selective tropism. The versatile tropism of pseudotyped LVs has been utilized in the treatment of tumors (Lee et al., 2021) and gene modification (Gutierrez-Guerrero et al., 2020). AAV vectors are hindered by their relatively low packaging capacity (Wu et al., 2010), with a packaging range of no more than 5 kb, making them inappropriate for the delivery of larger Cas9 variants (Mali et al., 2013). To address this issue, dual-AAV systems have been explored, in which one encodes Cas9 and another encodes gRNA, resulting in a large target gene transfer (Zhi et al., 2022). It needs to be pointed out, however, that the disadvantages of this system limit its clinical application, such as low probability of delivering both viral vectors to the same cell and insufficient expression efficiency.

4.1.2 Extracellular Vesicles

Whether *in vitro* or *in vivo*, extracellular vesicles (EVs) have been widely used to efficiently deliver genes or drugs (Choi et al., 2016; Montagna et al., 2018; Campbell et al., 2019; Mangeot et al., 2019; Gee et al., 2020). As natural cell-derived membrane vesicles, EVs serve the function of cell-to-cell communication with outstanding biocompatibility and immune-privileged characteristics. EVs are also hardly cleared by the immune system, avoiding the occurrence of hypersensitivity reactions (Zhang et al., 2014). Since EVs do not contain viral genomes, they have significant biosafety without the risk of endogenous virus recombination (Fuenmayor et al., 2017). Additionally, EVs transmit Cas9 with transient exposure, reducing the off-target chance triggered by Cas9 overexpression (Wu et al., 2014). All these advantages demonstrate an excellent potential for EVs as endogenous nano-vehicles in various fields. However, a major obstacle for EVs is the lack of robust tissue-specific delivery to specific cells. Targeted ligand modification on the surface of EVs is a promising avenue to ameliorate this weakness (Mathieu et al., 2019)—for instance, valency-controlled tetrahedral DNA nanostructures (TDNs) conjugated with DNA aptamers can be anchored on the EV surface *via* cholesterol, improving cell-specific delivery (Zhuang et al., 2020). The 3D tetrahedral steric superiority of TDN DNA aptamers can minimize lateral interactions among DNA, resulting in increased receptor–ligand binding and greatly enhancing tissue specificity. Compared with a control group, the TDN1-EVs-RNP group maximally restrained tumor growth in terms of tumor weight, volume, and percentage of tumor cells, demonstrating that the modified group accomplished a 2-fold increase in indel rate (up to 30%). Recently, EVs have been used in chimeric-antigen receptor (CAR) T-cell therapy to deliver CRISPR components to target cells precisely. By expressing chimeric-antigen receptors on vesicles derived from T cells, the anti-CD19-CAR-EVs preferentially accumulated in tumors compared to the liver, kidney, and other healthy tissues. Nevertheless, normal EVs were more evenly distributed throughout the body (Xu et al., 2020a). In addition to delivering CRISPR/Cas9 components, EVs also show great potential for drug delivery (Mateescu et al., 2017; Yang et al., 2018), anticancer therapy (Pascucci et al., 2014; Saari et al., 2015), and antigen delivery for vaccine development (Rabu et al., 2019).

4.1.3 Lipid Nanoparticles

Lipid nanoparticles (LNP) as CRISPR delivery vehicles have attracted the interest of scientists (Yin et al., 2014; Kulkarni et al., 2019). They not only help CRISPR components cross cell membranes but also protect them from enzymatic degradation and immune responses (Liu et al., 2018b; Noll et al., 2018). Due to the advantages of excellent controlled release, targeting, and high stability, LNPs have been widely used as a CRISPR delivery vector for all kinds of cargo modality, such as plasmid DNA, mRNA, and RNP complexes (Li et al., 2018c; Li et al., 2019). Theoretically, endocytosis is considered to be the key to cell internalization for almost all common LNP materials. To improve tissue specificity and delivery efficacy, several new strategies have been reported in recent years. Firstly, based on the hypothesis that charge adjustment can mediate tissue-specific delivery, a new strategy termed selective organ targeting (SORT) has been established. By adding DOTAP (a permanently cationic lipid) and constantly regulating its proportion to the original composition of LNP, we can control the charge for tissue-specific delivery (Cheng et al., 2020). The results show that this SORT strategy can achieve high organ selectivity for CRISPR cargos delivered in the lung, spleen, liver, and other organs. Among these organs, delivery to hepatocytes has the highest specificity at 93%. Secondly, ultrasound has been reported to facilitate the delivery of CRISPR components (Shen et al., 2016; Yoon et al., 2017). Ultrasound at specific locations can cause microbubbles to create local membrane deformations and pore formation in response to acoustic energy (Taniyama et al., 2011; Zhou et al., 2012). LNP released by microbubbles can then be transferred directly into the cytoplasm by diffusion. The results show that LNP incorporated with microbubbles can effectively facilitate cargo to the target site for RNP delivery, and the editing efficiency of Cas9 RNP was improved by 71.6% (Ryu et al., 2020). Thirdly, under optimized synthetic conditions, microfluidic device-designed lipid nanoparticles achieved intracellular RNP delivery with 97% gene disruption and 23% base substitution without any apparent cytotoxicity (Suzuki et al., 2020). In short, optimizing the formulation of LNP or integrating other technologies into the delivery system will be a crucial direction for achieving tissue-specific and efficient systems. Lipid-based formulations, however, do have some disadvantages. Once nanoparticles pass through the surface of cells, they are typically encased within an endosome. The encased contents then enter the lysosomal pathway directly and are eventually degraded. Therefore, coating polymers on the LNP surface or developing other unique chemical modifications to facilitate cellular uptake and disrupting endosomal membranes are promising directions that could prompt endosomal escape and avoid detection by the immune system.

4.1.4 Cell-Penetrating Peptides

As short stretches of amino acids, CPPs are polycationic, amphipathic, or non-polar in nature and possess an intrinsic

ability to translocate across cell membranes (Suresh et al., 2017). Owing to the advantages of low cytotoxicity, better biological tolerance, less off-target effect, and no chemical reagent, CPPs have been exploited to deliver different cargos into cells *in vitro* and *in vivo* (Liu et al., 2014; Gagat et al., 2017). When delivering RNP complexes, CPPs conjugated with RNP to form CPPs–RNP, which can improve cellular uptake and/or fusion. However, few studies have been reported on CPP-mediated CRISPR component delivery at present. Moreover, both delivery efficiency and subsequent editing efficiency were usually at a low level of just 10–20% (Ramakrishna et al., 2014b; Yin et al., 2018b; Del'Guidice et al., 2018; Yin et al., 2019a). This likely stems from the indefinite mechanism of CPP internalization and requirement for extensive optimization for targeting each type of cargo and cell. As the major CPP cargo is trapped in endosomes, they end up being recycled or degraded in a targeted manner instead of releasing cargo to the specific destination. Thus, enhancing endosomal escape would be a potential approach to improve the efficiency of delivery and editing (LeCher et al., 2017).

4.2 Abiotic Delivery Systems

As an alternative, abiotic vectors may offer tantalizing possibilities for CRISPR/Cas9 delivery systems due to their low immunogenicity, larger delivery gene payload, ease of large-scale production (Li et al., 2015), and absence of endogenous virus recombination. Many excellent delivery systems with new properties have been established in various fields, such as gold nanomaterials (Wang et al., 2018), polymers (Lv et al., 2018), and other systems. The characteristics of each material are described in detail in the following sections.

4.2.1 Gold Nanomaterials

Due to their tunable surface functionalization, non-toxic nature, favorable size, optical properties, biocompatibility, and photothermal effect, inorganic gold nanocarriers have proved to be a promising platform for systemic gene delivery (Ghosh et al., 2008; Ma et al., 2017b). They are mainly characterized by their photothermal effect and ease of functionalization for delivering CRISPR components. As photothermal transducers, gold nanomaterials can regulate the conditional control of Cas9 activity through different optical means (Nihongaki et al., 2015). In locally specific tissues, heat converted by the second near-infrared optical window (1,000 to 1,700 nm) induces endonuclear transformation of the heat-shock factor (HSF) from an inactive monomer to an active trimer. Under the action of active HSF, the combined transfection of a cationic polymer-coated Au nanorod, Cas9 plasmid, and a heat-inducible promoter HSP70 can result in 90% GFP-positive cells, which is much higher than that of Lipofectamine 2000 or 25-kDa polyethyleneimine (Chen et al., 2020b). In the LACM system, the protective DNA-modified gold nanorod hybridizes with the target binding domain of sgRNA to protect sgRNA. Upon NIR laser irradiation, heat subsequently denatures the hybridized DNA and sgRNA, accomplishing the controlled release of sgRNA into cells (Peng et al., 2020). Thus, gold nanomaterials act as an optogenetic switch to regulate the expression and activity of Cas9 proteins with high spatial specificity.

Tunable surface functionalization is another outstanding feature of gold nanocarriers that accelerates the entry of foreign genes into cells. Various biomolecules, such as proteins, DNA, peptides, and polymers, can endow gold nanomaterials with tremendous functions for surface bioengineering (Miao et al., 2018). Protamine, as a natural protein that originates from sperm, has intrinsic cell-penetrating properties and nucleus-targeting abilities and can be used for the efficient delivery of the Cas9–sgRNA plasmid. Protamine can form a compact structure with anionic DNA and then deliver the DNA to the egg nucleus (Biju et al., 2012; Priya et al., 2014). Nanocomplexes of Cas9-gRNA$_{EGFP}$ and protamine-functionalized gold nanoclusters disrupt the EGFP gene effectively and convert approximately 30% of the EGFP-positive transformants to EGFP-negative cells (Tao et al., 2021). Meanwhile, AuNCs can be functionalized by electrostatic action to control the self-assembly process. In a highly pH-dependent manner, AuNCs assembled with Cas9 protein (SpCas9–AuNCs) can deliver SpCas9 into the cell and nucleus in physiological conditions (Ju et al., 2019). The self-assembled SpCas9–AuNC nanoparticles effectively transfect HPV18 E6 sgRNA into cervical cancer cells, knocking out the E6 oncogene at a rate of 34%. More importantly, self-assembled SpCas9–AuNCs had little effect on normal cells, showing a considerable potential for clinical application. However, concerning the application of gold nanomaterials, cytokine production, the extensive modification requirement, fewer *in vivo* experiments, and potential toxicity need to be fully considered (Dykman and Khlebtsov, 2017). Gold nanomaterials are potentially an excellent delivery system and a bright prospect for improving CRISPR systems. Additionally, they can be extensively applied to bioimaging, optical and electrochemical sensing, and medical diagnostics (Chen et al., 2016). The multifunctional integrated gold nanomaterial platform may make great contributions to biological research in the future.

4.2.2 Polymers

Polymers can also be used to deliver RNP complexes to target sites with many distinct advantages, such as ease of synthesis, structural and component flexibility, functionalization, and degradability (Chen et al., 2016). Their significant flexibility is the most fascinating feature, resulting in multifunctionality by the reasonable and convenient design of the chemical structure (Hsu and Uludag, 2012; Zhang et al., 2019b). Currently, commonly used polymers to deliver drugs or RNP include polylysine, chitosan nanoparticles, poly-(β-amino ester)s, and dendrimers. The first two kinds are commonly used for drug delivery, while the latter two are mostly used for RNP delivery. Studies of drug delivery with polymers have been described in detail in other reviews (Huo et al., 2017; Hasheminejad et al., 2019). For a wide range of unmet therapeutic needs and personalized medicine, poly-(β-amino esters), as a class of amphiphilic and pH-sensitive polymers, can efficiently bind to cargo proteins to facilitate efficient intracellular RNP delivery *via* hydrogen bonding as well as hydrophobic and ionic interactions (Dwivedi et al., 2012). This characteristic allows them to be customized specifically to overcome delivery barriers in varied applications (Karlsson et al., 2020). Dendrimers are a class of synthetic polymer with a spherical and hyperbranched structure, whose surface is functionalized with a high

density of phenylboronic acid moieties to ensure that RNPs are efficiently bound to the dendrimer scaffold and transmit RNP to specific cells (Dixit et al., 2014). As a novel therapeutic tool for genetic disorders, dendrimers allow the efficient delivery of RNP targeting multiple genetic loci in different cell lines, proving to be a useful material for the delivery of genome-editing tools with broad biomedical applications (Taharabaru et al., 2020). Several issues exist with RNP delivery using polymers, however, such as low efficiency, high cytotoxicity, and narrow application range, which need to be overcome in the future.

5 CRISPR REGULATION WITH NUCLEASE-DEAD CAS PROTEINS

Through the same mechanism mentioned above, sgRNA-directed dCas9 binds to specific DNA sequences. When dCas9 binds specifically to a genomic locus, it can sterically block or activate RNP progression to downstream genes. These two dCas9-based strategies are called CRISPR interference (CRISPRi) (Ji et al., 2020) and CRISPR activation (CRISPRa), respectively (Larson et al., 2013). Both strategies can precisely regulate the expression of the sgRNA module or dCas9 *via* an inducible expression system. As of yet, several dCas9-based CRISPRa methods have been established, including dCas9-P65AD (Gilbert et al., 2013), dCas9-VPR (Chavez et al., 2015), dCas9-p300 (Hilton et al., 2015), and dCas9-TET (Xu et al., 2018). Some CRISPRi methods have also been reported, including dCas9-KRAB (Abudayyeh et al., 2017), dCas9-LSD1/KDM1A (Gilbert et al., 2013), and dCas13-YTHDF2 (Rauch et al., 2018). Several excellent reviews concerning different dCas-based CRISPRi and CRISPRa strategies describe their mechanism and principle in detail (Kampmann, 2018; Xu et al., 2020b). Currently, they are utilized to screen cellular genomes, including for cell survival/proliferation, sensitivity to drugs or toxins, fluorescent reporters, and single-cell transcriptomes (Kampmann, 2018). They are expected to precisely regulate editing time to reduce off-target effects.

6 ENRICHMENT OF MUTANTS

Due to off-target effects, not all genetically modified cells are equipped with positive mutants *in vitro*. The selection of mutants from original gene-edited cells is still a challenge at present (Ren et al., 2015). Thus, new strategies need to be investigated for enrichment and selection (**Figure 2C**). The most common selection markers to enrich positive cells are fluorescent proteins, antibiotic resistance genes, cell surface antigens, and so forth. Due to the merits of visualization, time saving, and decreased labor, fluorescent proteins are widely utilized in CRISPR/Cas systems (Ren et al., 2019). For a variety of cellular and environmental contexts, the variety of fluorescent genes gives scientists immense flexibility in choosing tailored reporters, such as green fluorescent protein, red fluorescent proteins (Liu et al., 2021a), and fluorescent proteins (Cao et al., 2019). Nevertheless, isolated cells are easily damaged by the solid lasers and hydrostatic pressure of flow cytometry. Compared to fluorescent proteins, the antibiotic-based method offers an alternative strategy that does not require expensive equipment but needs more time (Moriarity et al., 2014; Liesche et al., 2016). Although numerous antibiotic resistance genes have been applied in various fields, such as hygromycin (Moriarity et al., 2014), neomycin (Gu et al., 2021), zeocin (Kobayashi et al., 2019), gentamicin (Mulsant et al., 1988) and puromycin (Pandey et al., 2021), marker-free strategies are the preferred method, ameliorating public concerns for the biosafety of antibiotic resistance genes. Another non-fluorescence activated cell sorting-based enrichment method is antigen gene H-2K^k, which has a high enrichment efficiency with magnetic bead separation (Wei et al., 2001). However, when insertions or deletions are generated at the target sequences, these reporter systems express H-2K^k and hygromycin resistance protein, respectively, enabling the efficient enrichment of mutants without flow cytometry (Kim et al., 2013).

However, no matter what efficient strategies are used to select mutants, mutant enrichment alone cannot classify all stable and highly expressed mutants (**Figure 2D**). Thus, to select nuclear-active mutants, two surrogate reporters based on the NHEJ and single-strand annealing (SSA) have been published (Pattanayak et al., 2013). Nuclease triggers a DSB on the target sequence within the surrogate reporter construct, resulting in the formation of small random indels by the error-prone NHEJ repair pathway and leading to the correction of reporter genes with 1/3 frequency. Compared with unsorted cells, the enrichment efficiency of mutants can be increased up to 8.6- and 18-fold with the first and second generation of NHEJ-based surrogate reporters, respectively (Ramakrishna et al., 2014a). The second surrogate reporter has the capacity to identify more nuclease-positive cells *via* SSA. Due to its higher sensitivity, this reporter significantly increases the possibility of obtaining the desired genetically modified cell clones (Yasuda et al., 2016). Although DNA repair pathways are influenced by cell type and the nature of broken DNA ends, genomic modification within mutants may be independent of repair pathways in surrogate reporters (Ren et al., 2015). On the basis of transfection-positive cells, these two surrogate reporter strategies can produce highly efficient, nuclease-active cells.

7 CONCLUSION AND FUTURE PROSPECTS

Aside from the above-mentioned approaches, other strategies can also significantly improve the editing efficiency of CRISPR/Cas systems. Firstly, owing to the fact that the nucleosome poses a strong barrier to Cas9, restoring Cas9 access to nucleosomes through the chromatin remodeling enzyme yChd1 therefore results in high efficiency editing. Nucleosome organization represents only one aspect of eukaryotic chromatin, however; thus, future research on how chromatin affects Cas9 activity needs to be done (Horlbeck et al., 2016). Secondly, cytosine base editors (CBE) and adenine base editors (ABE) have been utilized to change C/G to T/A and A/T to G/C. CBE deaminates cytosine to uracil, which is recognized by the cell replication machinery as thymine, resulting in a C/G to T/A transition. ABE-mediated

DNA editing operates under a similar mechanism as that of CBE (Koblan et al., 2018; Richter et al., 2020). Despite efforts to improve DNA base editors, base editing is confined to transition mutations (incapable of transversion mutation) and is not capable of inducing indel mutations. Next, by combining reverse transcriptase with prime editors gRNA and Cas-nickase nuclease, prime editing technology can edit or "search and replace" bases in a genome (Anzalone et al., 2019). It can also be used as an alternative genome editing tool to investigate various challenges, such as editing large genes, targeting autosomal dominant diseases, and editing premature stop codons and splice-site variants (Kantor et al., 2020). When prime editors are undesirable and the base editing window is well defined, base editors are typically more efficient than prime editors. On the contrary, when prime editors are acceptable and multiple editable bases are within a defined editing window, prime editors offer unsurmountable advantages.

In the last few years, we have seen the extraordinary growth and expansion of gene editing, particularly in the field of gene therapy. Based on CRISPR technology, a series of highly efficient and targeted transcription factor components has been developed and used to construct intelligent gene circuits, making tumor gene therapy possible (Zhou et al., 2019). In cardiovascular medicine, CRISPR-based tools have multiple applications, with a primary focus on direct therapeutic interventions to treat inherited cardiac disorders (Vermersch et al., 2020). CRISPR also represents a breakthrough advance in genetically engineered immune cells (Huang et al., 2020), personalized cancer medicine (Li and Kasinski, 2020), and modification of human embryos (Tang et al., 2017). Even in the current novel coronavirus (COVID-19) outbreak, CRISPR-based technology has shown strong application value. All-in-one dual CRISPR-Cas12a is instrumental in the detection of COVID-19, offering the advantages of being instrument-free, rapid, sensitive, one-pot, and point-of-care (Ding et al., 2020). Applications in microbiology are still being newly discovered and improved, specifically in the identification and modification of industrial-related lactobacilli and streptococci as well as foodborne pathogens, including *E. coli* (Altenbuchner, 2016), *Saccharomyces cerevisiae* (Biot-Pelletier and Martin, 2016), and thermophilic fungi (Liu et al., 2017b). As a new generation of precision gene editing tools, the great success of CRISPR/Cas systems in various fields shows that these have a wide range of application and wonderful prospects.

Collectively, knowledge and technologies of genome editing are ceaselessly developing in intricately interwoven fields and are creating huge synergies. With the recent developments in CRISPR/Cas systems, they are becoming increasingly accurate, efficient, and reliable. Although massive advances have been achieved, the CRISPR/Cas systems are far from their optimal state. Among various challenges, off-target effects are still the foremost barrier in CRISPR/Cas systems. We have listed above several strategies for reducing off-target effects. Among them, special attention should be paid to optimizing time and temperature, which are often inadvertently neglected. The CRISPR/Cas systems have other limitations, including inactive mutants, variable efficiency, requirement of PAM and sgRNA, fault-prone programmed DNA repair pathways, and the lack of an efficient and safe delivery system. Apart from these, future research will involve the enhancement of Cas9 activity, application of ACR proteins, and determination of the optimal Cas9 and sgRNA ratio so as to further improve the efficiency of CRISPR systems. Simultaneously, continuous optimization of external measures, including dCas9 regulation, delivery vector development, mutant enrichment, *etc.*, will help to further improve the efficiency. Although we are far from eliminating off-target effects completely, we are confident that CRISPR technology will continue to be perfected to meet the demands of different fields by adopting the aforementioned strategies.

AUTHOR CONTRIBUTIONS

All authors took part in writing, reviewing, and editing the manuscript. SF, ZW, and AL wrote the manuscript. XX and SL prepared the figures. YL and BW collected and organized the literature. JL and LH created the table. LY and TG modified the paper. All authors reviewed the manuscript and approved it for publication.

FUNDING

This work was supported by the National Natural Science Foundation of China (nos. U1804112 and 31571289) and the Zhongjing Core Scholar's Research Initial Fund of Henan University of Chinese Medicine (no. 00104311-2021).

REFERENCES

Abby, S. S., Néron, B., Ménager, H., Touchon, M., and Rocha, E. P. C. (2014). MacSyFinder: a Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS One* 9 (10), e110726. doi:10.1371/journal.pone.0110726

Abudayyeh, O. O., Gootenberg, J. S., Essletzbichler, P., Han, S., Joung, J., Belanto, J. J., et al. (2017). RNA Targeting with CRISPR-Cas13. *Nature* 550 (7675), 280–284. doi:10.1038/nature24049

Adli, M. (2018). The CRISPR Tool Kit for Genome Editing and beyond. *Nat. Commun.* 9 (1), 1911. doi:10.1038/s41467-018-04252-2

Aird, E. J., Lovendahl, K. N., St. Martin, A., Harris, R. S., and Gordon, W. R. (2018). Increasing Cas9-Mediated Homology-Directed Repair Efficiency through Covalent Tethering of DNA Repair Template. *Commun. Biol.* 1, 54. doi:10.1038/s42003-018-0054-2

Altenbuchner, J. (2016). Editing of the Bacillus Subtilis Genome by the CRISPR-Cas9 System. *Appl. Environ. Microbiol.* 82 (17), 5421–5427. doi:10.1128/aem.01453-16

Anuchina, A. A., Lavrov, A. V., and Smirnikhina, S. A. (2020). TIRR: a Potential Front Runner in HDR Race–hypotheses and Perspectives. *Mol. Biol. Rep.* 47 (3), 2371–2379. doi:10.1007/s12033-020-05285-x

Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., et al. (2019). Search-and-replace Genome Editing without Double-Strand Breaks or Donor DNA. *Nature* 576 (7785), 149–157. doi:10.1038/s41586-019-1711-4

Batool, A., Malik, F., and Andrabi, K. I. (2021). Expansion of the CRISPR/Cas Genome-Sculpting Toolbox: Innovations, Applications and Challenges. *Mol. Diagn. Ther.* 25 (1), 41–57. doi:10.1007/s40291-020-00500-8

Belot, L., Albertini, A., and Gaudin, Y. (2019). Structural and Cellular Biology of Rhabdovirus Entry. *Adv. Virus. Res.* 104, 147–183. doi:10.1016/bs.aivir.2019.05.003

Beumer, K. J., Trautman, J. K., Mukherjee, K., and Carroll, D. (2013). Donor DNA Utilization during Gene Targeting with Zinc-Finger Nucleases. *G3 (Bethesda)* 3 (4), 657–664. doi:10.1534/g3.112.005439

Biju, V., Anas, A., Akita, H., Shibu, E. S., Itoh, T., Harashima, H., et al. (2012). FRET from Quantum Dots to Photodecompose Undesired Acceptors and Report the Condensation and Decondensation of Plasmid DNA. *ACS Nano* 6 (5), 3776–3788. doi:10.1021/nn2048608

Biot-Pelletier, D., and Martin, V. J. J. (2016). Seamless Site-Directed Mutagenesis of the *Saccharomyces cerevisiae* Genome Using CRISPR-Cas9. *J. Biol. Eng.* 10, 6. doi:10.1186/s13036-016-0028-1

Bosch, J. A., Colbeth, R., Zirin, J., and Perrimon, N. (2020). Gene Knock-Ins in Drosophila Using Homology-independent Insertion of Universal Donor Plasmids. *Genetics* 214 (1), 75–89. doi:10.1534/genetics.119.302819

Boucher, P., Cui, X., and Curiel, D. T. (2020). Adenoviral Vectors for *In Vivo* Delivery of CRISPR-Cas Gene Editors. *J. Controlled Release* 327, 788–800. doi:10.1016/j.jconrel.2020.09.003

Bubeck, F., Hoffmann, M. D., Harteveld, Z., Aschenbrenner, S., Bietz, A., Waldhauer, M. C., et al. (2018). Engineered Anti-CRISPR Proteins for Optogenetic Control of CRISPR-Cas9. *Nat. Methods* 15 (11), 924–927. doi:10.1038/s41592-018-0178-9

Cameron, P., Fuller, C. K., Donohoue, P. D., Jones, B. N., Thompson, M. S., Carter, M. M., et al. (2017). Mapping the Genomic Landscape of CRISPR-Cas9 Cleavage. *Nat. Methods* 14 (6), 600–606. doi:10.1038/nmeth.4284

Campbell, L. A., Coke, L. M., Richie, C. T., Fortuno, L. V., Park, A. Y., and Harvey, B. K. (2019). Gesicle-Mediated Delivery of CRISPR/Cas9 Ribonucleoprotein Complex for Inactivating the HIV Provirus. *Mol. Ther.* 27 (1), 151–163. doi:10.1016/j.ymthe.2018.10.002

Canny, M. D., Moatti, N., Wan, L. C. K., Fradet-Turcotte, A., Krasner, D., Mateos-Gomez, P. A., et al. (2018). Inhibition of 53BP1 Favors Homology-dependent DNA Repair and Increases CRISPR-Cas9 Genome-Editing Efficiency. *Nat. Biotechnol.* 36 (1), 95–102. doi:10.1038/nbt.4021

Canto, T. (2016). Transient Expression Systems in Plants: Potentialities and Constraints. *Adv. Exp. Med. Biol.* 896, 287–301. doi:10.1007/978-3-319-27216-0_18

Cao, X., Zhang, C., Gao, Z., Liu, Y., Zhao, Y., Yang, Y., et al. (2019). Ultrafast Internal Conversion Dynamics of Bilirubin Bound to UnaG and its N57A Mutant. *Phys. Chem. Chem. Phys.* 21 (5), 2365–2371. doi:10.1039/c8cp07553k

Carlson-Stevermer, J., Abdeen, A. A., Kohlenberg, L., Goedland, M., Molugu, K., Lou, M., et al. (2017). Assembly of CRISPR Ribonucleoproteins with Biotinylated Oligonucleotides via an RNA Aptamer for Precise Gene Editing. *Nat. Commun.* 8 (1), 1711. doi:10.1038/s41467-017-01875-9

Casini, A., Olivieri, M., Petris, G., Montagna, C., Reginato, G., Maule, G., et al. (2018). A Highly Specific SpCas9 Variant Is Identified by *In Vivo* Screening in Yeast. *Nat. Biotechnol.* 36 (3), 265–271. doi:10.1038/nbt.4066

Charpentier, M., Khedher, A. H. Y., Menoret, S., Brion, A., Lamribet, K., Dardillac, E., et al. (2018). CtIP Fusion to Cas9 Enhances Transgene Integration by Homology-dependent Repair. *Nat. Commun.* 9 (1), 1133. doi:10.1038/s41467-018-03475-7

Chavez, A., Scheiman, J., Vora, S., Pruitt, B. W., Tuttle, M., P R Iyer, E., et al. (2015). Highly Efficient Cas9-Mediated Transcriptional Programming. *Nat. Methods* 12 (4), 326–328. doi:10.1038/nmeth.3312

Chen, F., Alphonse, M., and Liu, Q. (2020a). Strategies for Nonviral Nanoparticle-Based Delivery of CRISPR/Cas9 Therapeutics. *Wiley Interdiscip. Rev. Nanomed Nanobiotechnol* 12 (3), e1609. doi:10.1002/wnan.1609

Chen, J., Guo, Z., Tian, H., and Chen, X. (2016). Production and Clinical Development of Nanoparticles for Gene Delivery. *Mol. Ther. - Methods Clin. Develop.* 3, 16023. doi:10.1038/mtm.2016.23

Chen, J. S., Dagdas, Y. S., Kleinstiver, B. P., Welch, M. M., Sousa, A. A., Harrington, L. B., et al. (2017). Enhanced Proofreading Governs CRISPR-Cas9 Targeting Accuracy. *Nature* 550 (7676), 407–410. doi:10.1038/nature24268

Chen, X., Chen, Y., Xin, H., Wan, T., and Ping, Y. (2020b). Near-infrared Optogenetic Engineering of Photothermal nanoCRISPR for Programmable Genome Editing. *Proc. Natl. Acad. Sci. USA* 117 (5), 2395–2405. doi:10.1073/pnas.1912220117

Chen, X., and Gonçalves, M. A. F. V. (2016). Engineered Viruses as Genome Editing Devices. *Mol. Ther.* 24 (3), 447–457. doi:10.1038/mt.2015.164

Cheng, Q., Wei, T., Farbiak, L., Johnson, L. T., Dilliard, S. A., and Siegwart, D. J. (2020). Selective Organ Targeting (SORT) Nanoparticles for Tissue-specific mRNA Delivery and CRISPR-Cas Gene Editing. *Nat. Nanotechnol.* 15 (4), 313–320. doi:10.1038/s41565-020-0669-6

Choi, J. G., Dang, Y., Abraham, S., Ma, H., Zhang, J., Guo, H., et al. (2016). Lentivirus Pre-packed with Cas9 Protein for Safer Gene Editing. *Gene Ther.* 23 (7), 627–633. doi:10.1038/gt.2016.27

Chu, V. T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., et al. (2015). Increasing the Efficiency of Homology-Directed Repair for CRISPR-Cas9-Induced Precise Gene Editing in Mammalian Cells. *Nat. Biotechnol.* 33 (5), 543–548. doi:10.1038/nbt.3198

Coelho, M. A., De Braekeleer, E., Firth, M., Bista, M., Lukasiak, S., Cuomo, M. E., et al. (2020). CRISPR GUARD Protects Off-Target Sites from Cas9 Nuclease Activity Using Short Guide RNAs. *Nat. Commun.* 11 (1), 4132. doi:10.1038/s41467-020-17952-5

Crudele, J. M., and Chamberlain, J. S. (2019). AAV-based Gene Therapies for the Muscular Dystrophies. *Hum. Mol. Genet.* 28 (R1), R102–R107. doi:10.1093/hmg/ddz128

Del'Guidice, T., Lepetit-Stoffaes, J. P., Bordeleau, L. J., Roberge, J., Théberge, V., Lauvaux, C., et al. (2018). Membrane Permeabilizing Amphiphilic Peptide Delivers Recombinant Transcription Factor and CRISPR-Cas9/Cpf1 Ribonucleoproteins in Hard-To-Modify Cells. *PLoS One* 13 (4), e0195558. doi:10.1371/journal.pone.0195558

Demirer, G. S., Silva, T. N., Jackson, C. T., Thomas, J. B., Ehrhardt, D, W., Rhee, S. Y., et al. (2021). Nanotechnology to advance CRISPR-Cas Genetic Engineering of Plants. *Nat. Nanotechnol.* 16 (3), 243–250. doi:10.1038/s41565-021-00854-y

Devkota, S. (2018). The Road Less Traveled: Strategies to Enhance the Frequency of Homology-Directed Repair (HDR) for Increased Efficiency of CRISPR/Cas-mediated Transgenesis. *BMB Rep.* 51 (9), 437–443. doi:10.5483/bmbrep.2018.51.9.187

DiCarlo, J. E., Deeconda, A., and Tsang, S. H. (2017). Viral Vectors, Engineered Cells and the CRISPR Revolution. *Adv. Exp. Med. Biol.* 1016, 3–27. doi:10.1007/978-3-319-63904-8_1

Ding, X., Yin, K., Li, Z., Lalla, R. V., Ballesteros, E., Sfeir, M. M., et al. (2020). Ultrasensitive and Visual Detection of SARS-CoV-2 Using All-In-One Dual CRISPR-Cas12a Assay. *Nat. Commun.* 11 (1), 4711. doi:10.1038/s41467-020-18575-6

DiTommaso, T., Cole, J. M., Cassereau, L., Buggé, J. A., Hanson, J. L. S., Bridgen, D. T., et al. (2018). Cell Engineering with Microfluidic Squeezing Preserves Functionality of Primary Immune Cells *In Vivo*. *Proc. Natl. Acad. Sci. USA* 115 (46), E10907–E10914. doi:10.1073/pnas.1809671115

Dixit, S., Singh, S. R., Yilma, A. N., Agee, R. D., 2nd, Taha, M., and Dennis, V. A. (2014). Poly(lactic Acid)-Poly(ethylene Glycol) Nanoparticles Provide Sustained Delivery of a *Chlamydia trachomatis* Recombinant MOMP Peptide and Potentiate Systemic Adaptive Immune Responses in Mice. *Nanomedicine: Nanotechnology, Biol. Med.* 10 (6), 1311–1321. doi:10.1016/j.nano.2014.02.009

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., et al. (2016). Optimized sgRNA Design to Maximize Activity and Minimize Off-Target Effects of CRISPR-Cas9. *Nat. Biotechnol.* 34 (2), 184–191. doi:10.1038/nbt.3437

Dwivedi, V., Manickam, C., Binjawadagi, B., Joyappa, D., and Renukaradhya, G. J. (2012). Biodegradable Nanoparticle-Entrapped Vaccine Induces Cross-Protective Immune Response against a Virulent Heterologous Respiratory Viral Infection in Pigs. *PLoS One* 7 (12), e51794. doi:10.1371/journal.pone.0051794

Dykman, L. A., and Khlebtsov, N. G. (2017). Immunological Properties of Gold Nanoparticles. *Chem. Sci.* 8 (3), 1719–1735. doi:10.1039/c6sc03631g

England, W. E., and Whitaker, R. J. (2013). Evolutionary Causes and Consequences of Diversified CRISPR Immune Profiles in Natural Populations. *Biochem. Soc. Trans.* 41 (6), 1431–1436. doi:10.1042/bst20130243

Ferrari, S., Jacob, A., Beretta, S., Unali, G., Albano, L., Vavassori, V., et al. (2020). Efficient Gene Editing of Human Long-Term Hematopoietic Stem Cells Validated by Clonal Tracking. *Nat. Biotechnol.* 38 (11), 1298–1308. doi:10.1038/s41587-020-0551-y

Fu, Y.-W., Dai, X.-Y., Wang, W.-T., Yang, Z.-X., Zhao, J.-J., Zhang, J.-P., et al. (2021). Dynamics and Competition of CRISPR-Cas9 Ribonucleoproteins and AAV Donor-Mediated NHEJ, MMEJ and HDR Editing. *Nucleic Acids Res.* 49 (2), 969–985. doi:10.1093/nar/gkaa1251

Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M., and Joung, J. K. (2014). Improving CRISPR-Cas Nuclease Specificity Using Truncated Guide RNAs. *Nat. Biotechnol.* 32 (3), 279–284. doi:10.1038/nbt.2808

Fuenmayor, J., Gòdia, F., and Cervera, L. (2017). Production of Virus-like Particles for Vaccines. *New Biotechnol.* 39 (Pt B), 174–180. doi:10.1016/j.nbt.2017.07.010

Gagat, M., Zielińska, W., and Grzanka, A. (2017). Cell-penetrating Peptides and Their Utility in Genome Function Modifications (Review). *Int. J. Mol. Med.* 40 (6), 1615–1623. doi:10.3892/ijmm.2017.3172

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA Ribonucleoprotein Complex Mediates Specific DNA Cleavage for Adaptive Immunity in Bacteria. *Proc. Natl. Acad. Sci.* 109 (39), E2579–E2586. doi:10.1073/pnas.1208507109

Gee, P., Lung, M. S. Y., Okuzaki, Y., Sasakawa, N., Iguchi, T., Makita, Y., et al. (2020). Extracellular Nanovesicles for Packaging of CRISPR-Cas9 Protein and sgRNA to Induce Therapeutic Exon Skipping. *Nat. Commun.* 11 (1), 1334. doi:10.1038/s41467-020-14957-y

Ghezraoui, H., Piganeau, M., Renouf, B., Renaud, J.-B., Sallmyr, A., Ruis, B., et al. (2014). Chromosomal Translocations in Human Cells Are Generated by Canonical Nonhomologous End-Joining. *Mol. Cel* 55 (6), 829–842. doi:10.1016/j.molcel.2014.08.002

Ghosh, P., Han, G., De, M., Kim, C., and Rotello, V. (2008). Gold Nanoparticles in Delivery Applications☆. *Adv. Drug Deliv. Rev.* 60 (11), 1307–1315. doi:10.1016/j.addr.2008.03.016

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., et al. (2013). CRISPR-mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* 154 (2), 442–451. doi:10.1016/j.cell.2013.06.044

Glass, Z., Lee, M., Li, Y., and Xu, Q. (2018). Engineering the Delivery System for CRISPR-Based Genome Editing. *Trends Biotechnol.* 36 (2), 173–185. doi:10.1016/j.tibtech.2017.11.006

Greco, G. E., Matsumoto, Y., Brooks, R. C., Lu, Z., Lieber, M. R., and Tomkinson, A. E. (2016). SCR7 Is Neither a Selective Nor a Potent Inhibitor of Human DNA Ligase IV. *DNA Repair* 43, 18–23. doi:10.1016/j.dnarep.2016.04.004

Green, J. C., and Hu, J. S. (2017). Editing Plants for Virus Resistance Using CRISPR-Cas. *Acta Virol.* 61 (2), 138–142. doi:10.4149/av_2017_02_02

Gu, X., Wang, D., Xu, Z., Wang, J., Guo, L., Chai, R., et al. (2021). Prevention of Acquired Sensorineural Hearing Loss in Mice by *In Vivo* Htra2 Gene Editing. *Genome Biol.* 22 (1), 86. doi:10.1186/s13059-021-02311-4

Guilinger, J. P., Thompson, D. B., and Liu, D. R. (2014). Fusion of Catalytically Inactive Cas9 to FokI Nuclease Improves the Specificity of Genome Modification. *Nat. Biotechnol.* 32 (6), 577–582. doi:10.1038/nbt.2909

Gutierrez-Guerrero, A., Cosset, F. L., and Verhoeyen, E. (2020). Lentiviral Vector Pseudotypes: Precious Tools to Improve Gene Modification of Hematopoietic Cells for Research and Gene Therapy. *Viruses* 12 (9), 1016. doi:10.3390/v12091016

Hande, K. R., Hagey, A., Berlin, J., Cai, Y., Meek, K., Kobayashi, H., et al. (2006). The Pharmacokinetics and Safety of ABT-751, a Novel, Orally Bioavailable Sulfonamide Antimitotic Agent: Results of a Phase 1 Study. *Clin. Cancer Res.* 12 (9), 2834–2840. doi:10.1158/1078-0432.ccr-05-2159

Harrington, L. B., Doxzen, K. W., Ma, E., Liu, J.-J., Knott, G. J., Edraki, A., et al. (2017). A Broad-Spectrum Inhibitor of CRISPR-Cas9. *Cell* 170 (6), 1224–1233. e15. doi:10.1016/j.cell.2017.07.037

Hart, S. L., and Harrison, P. T. (2017). Genetic Therapies for Cystic Fibrosis Lung Disease. *Curr. Opin. Pharmacol.* 34, 119–124. doi:10.1016/j.coph.2017.10.006

Hasheminejad, N., Khodaiyan, F., and Safari, M. (2019). Improving the Antifungal Activity of Clove Essential Oil Encapsulated by Chitosan Nanoparticles. *Food Chem.* 275, 113–122. doi:10.1016/j.foodchem.2018.09.085

Heckl, D., Kowalczyk, M. S., Yudovich, D., Belizaire, R., Puram, R. V., McConkey, M. E., et al. (2014). Generation of Mouse Models of Myeloid Malignancy with Combinatorial Genetic Lesions Using CRISPR-Cas9 Genome Editing. *Nat. Biotechnol.* 32 (9), 941–946. doi:10.1038/nbt.2951

Heigwer, F., Kerr, G., and Boutros, M. (2014). E-CRISP: Fast CRISPR Target Site Identification. *Nat. Methods* 11 (2), 122–123. doi:10.1038/nmeth.2812

Heussler, G. E., Cady, K. C., Koeppen, K., Bhuju, S., Stanton, B. A., and O'Toole, G. A. (2015). Clustered Regularly Interspaced Short Palindromic Repeat-dependent, Biofilm-specific Death of *Pseudomonas aeruginosa* Mediated by Increased Expression of Phage-Related Genes. *mBio* 6 (3), e00129–15. doi:10.1128/mBio.00129-15

Hilton, I. B., D'Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., et al. (2015). Epigenome Editing by a CRISPR-Cas9-Based Acetyltransferase Activates Genes from Promoters and Enhancers. *Nat. Biotechnol.* 33 (5), 510–517. doi:10.1038/nbt.3199

Hoffmann, M. D., Aschenbrenner, S., Grosse, S., Rapti, K., Domenger, C., Fakhiri, J., et al. (2019). Cell-specific CRISPR-Cas9 Activation by microRNA-dependent Expression of Anti-CRISPR Proteins. *Nucleic Acids Res.* 47 (13), e75. doi:10.1093/nar/gkz271

Horlbeck, M. A., Witkowsky, L. B., Guglielmi, B., Replogle, J. M., Gilbert, L. A., Villalta, J. E., et al. (2016). Nucleosomes Impede Cas9 Access to DNA *In Vivo* and *In Vitro*. *Elife* 5, e12677. doi:10.7554/elife.12677

Høyland-Kroghsbo, N. M., Muñoz, K. A., and Bassler, B. L. (2018). Temperature, by Controlling Growth Rate, Regulates CRISPR-Cas Activity in *Pseudomonas aeruginosa*. *mBio* 9 (6), e02184–18. doi:10.1128/mBio.02184-18

Hsu, C. Y. M., and Uludağ, H. (2012). A Simple and Rapid Nonviral Approach to Efficiently Transfect Primary Tissue-Derived Cells Using Polyethylenimine. *Nat. Protoc.* 7 (5), 935–945. doi:10.1038/nprot.2012.038

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., et al. (2013). DNA Targeting Specificity of RNA-Guided Cas9 Nucleases. *Nat. Biotechnol.* 31 (9), 827–832. doi:10.1038/nbt.2647

Hu, Z., Shi, Z., Guo, X., Jiang, B., Wang, G., Luo, D., et al. (2018). Ligase IV Inhibitor SCR7 Enhances Gene Editing Directed by CRISPR-Cas9 and ssODN in Human Cancer Cells. *Cell Biosci* 8, 12. doi:10.1186/s13578-018-0200-z

Huang, D., Miller, M., Ashok, B., Jain, S., and Peppas, N. A. (2020). CRISPR/Cas Systems to Overcome Challenges in Developing the Next Generation of T Cells for Cancer Therapy. *Adv. Drug Deliv. Rev.* 158, 17–35. doi:10.1016/j.addr.2020.07.015

Huo, Q., Zhu, J., Niu, Y., Shi, H., Gong, Y., Li, Y., et al. (2017). pH-Triggered Surface Charge-Switchable Polymer Micelles for the Co-delivery of Paclitaxel/disulfiram and Overcoming Multidrug Resistance in Cancer. *Int. J. Nanomedicine* 12, 8631–8647. doi:10.2147/ijn.s144452

Janik, E., Niemcewicz, M., Ceremuga, M., Krzowski, L., Saluk-Bijak, J., and Bijak, M. (2020). Various Aspects of a Gene Editing System-CRISPR-Cas9. *Int. J. Mol. Sci.* 21 (24), 9604. doi:10.3390/ijms21249604

Jarrett, K. E., Lee, C., De Giorgi, M., Hurley, A., Gillard, B. K., Doerfler, A. M., et al. (2018). Somatic Editing of Ldlr with Adeno-Associated Viral-CRISPR Is an Efficient Tool for Atherosclerosis Research. *Arterioscler. Thromb. Vasc. Biol.* 38 (9), 1997–2006. doi:10.1161/atvbaha.118.311221

Jayavaradhan, R., Pillis, D. M., Goodman, M., Zhang, F., Zhang, Y., Andreassen, P. R., et al. (2019). CRISPR-Cas9 Fusion to Dominant-Negative 53BP1 Enhances HDR and Inhibits NHEJ Specifically at Cas9 Target Sites. *Nat. Commun.* 10 (1), 2866. doi:10.1038/s41467-019-10735-7

Ji, X., Zhao, H., Zhu, H., Zhu, K., Tang, S.-Y., and Lou, C. (2020). CRISPRi/dCpf1-mediated Dynamic Metabolic Switch to Enhance Butenoic Acid Production in *Escherichia coli*. *Appl. Microbiol. Biotechnol.* 104 (12), 5385–5393. doi:10.1007/s00253-020-10610-2

Jin, J., Jia, B., and Yuan, Y.-J. (2020). Yeast Chromosomal Engineering to Improve Industrially-Relevant Phenotypes. *Curr. Opin. Biotechnol.* 66, 165–170. doi:10.1016/j.copbio.2020.07.003

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* 337 (6096), 816–821. doi:10.1126/science.1225829

Johansen, A. K., Molenaar, B., Versteeg, D., Leitoguinho, A. R., Demkes, C., Spanjaard, B., et al. (2017). Postnatal Cardiac Gene Editing Using CRISPR/Cas9 with AAV9-Mediated Delivery of Short Guide RNAs Results in Mosaic Gene Disruption. *Circ. Res.* 121 (10), 1168–1181. doi:10.1161/circresaha.116.310370

Ju, E., Li, T., Ramos da Silva, S., and Gao, S.-J. (2019). Gold Nanocluster-Mediated Efficient Delivery of Cas9 Protein through pH-Induced Assembly-Disassembly for Inactivation of Virus Oncogenes. *ACS Appl. Mater. Inter.* 11 (38), 34717–34724. doi:10.1021/acsami.9b12335

Kampmann, M. (2018). CRISPRi and CRISPRa Screens in Mammalian Cells for Precision Biology and Medicine. *ACS Chem. Biol.* 13 (2), 406–416. doi:10.1021/acschembio.7b00657

Kantor, A., McClements, M. E., and MacLaren, R. E. (2020). CRISPR-Cas9 DNA Base-Editing and Prime-Editing. *Int. J. Mol. Sci.* 21 (17), 6240. doi:10.3390/ijms21176240

Karlsson, J., Rhodes, K. R., Green, J. J., and Tzeng, S. Y. (2020). Poly(beta-amino Ester)s as Gene Delivery Vehicles: Challenges and Opportunities. *Expert Opin. Drug Deliv.* 17 (10), 1395–1410. doi:10.1080/17425247.2020.1796628

Kim, H., Choi, J., and Won, K.-H. (2020a). A Stable DNA-free Screening System for CRISPR/RNPs-mediated Gene Editing in Hot and Sweet Cultivars of Capsicum Annuum. *BMC Plant Biol.* 20 (1), 449. doi:10.1186/s12870-020-02665-0

Kim, H., Kim, M.-S., Wee, G., Lee, C.-i., Kim, H., and Kim, J.-S. (2013). Magnetic Separation and Antibiotics Selection Enable Enrichment of Cells with ZFN/TALEN-induced Mutations. *PLoS One* 8 (2), e56476. doi:10.1371/journal.pone.0056476

Kim, N., Kim, H. K., Lee, S., Seo, J. H., Choi, J. W., Park, J., et al. (2020b). Prediction of the Sequence-specific Cleavage Activity of Cas9 Variants. *Nat. Biotechnol.* 38 (11), 1328–1336. doi:10.1038/s41587-020-0537-9

Kim, S., Kim, D., Cho, S. W., Kim, J., and Kim, J.-S. (2014). Highly Efficient RNA-Guided Genome Editing in Human Cells via Delivery of Purified Cas9 Ribonucleoproteins. *Genome Res.* 24 (6), 1012–1019. doi:10.1101/gr.171322.113

Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z., et al. (2016). High-fidelity CRISPR-Cas9 Nucleases with No Detectable Genome-wide Off-Target Effects. *Nature* 529 (7587), 490–495. doi:10.1038/nature16526

Kobayashi, K., Tsukiyama, T., Nakaya, M., Kageyama, S., Tomita, K., Murai, R., et al. (2019). Generation of an OCT3/4 Reporter Cynomolgus Monkey ES Cell Line Using CRISPR/Cas9. *Stem Cel Res.* 37, 101439. doi:10.1016/j.scr.2019.101439

Koblan, L. W., Doman, J. L., Wilson, C., Levy, J. M., Tay, T., Newby, G. A., et al. (2018). Improving Cytidine and Adenine Base Editors by Expression Optimization and Ancestral Reconstruction. *Nat. Biotechnol.* 36 (9), 843–846. doi:10.1038/nbt.4172

Kotterman, M. A., Chalberg, T. W., and Schaffer, D. V. (2015). Viral Vectors for Gene Therapy: Translational and Clinical Outlook. *Annu. Rev. Biomed. Eng.* 17, 63–89. doi:10.1146/annurev-bioeng-071813-104938

Kulcsár, P. I., Tálas, A., Huszár, K., Ligeti, Z., Tóth, E., Weinhardt, N., et al. (2017). Crossing Enhanced and High Fidelity SpCas9 Nucleases to Optimize Specificity and Cleavage. *Genome Biol.* 18 (1), 190. doi:10.1186/s13059-017-1318-8

Kulkarni, J. A., Witzigmann, D., Chen, S., Cullis, P. R., and van der Meel, R. (2019). Lipid Nanoparticle Technology for Clinical Translation of siRNA Therapeutics. *Acc. Chem. Res.* 52 (9), 2435–2444. doi:10.1021/acs.accounts.9b00368

Kumita, W., Sato, K., Suzuki, Y., Kurotaki, Y., Harada, T., Zhou, Y., et al. (2019). Efficient Generation of Knock-in/Knock-Out Marmoset Embryo via CRISPR/Cas9 Gene Editing. *Sci. Rep.* 9 (1), 12719. doi:10.1038/s41598-019-49110-3

Larson, M. H., Gilbert, L. A., Wang, X., Lim, W. A., Weissman, J. S., and Qi, L. S. (2013). CRISPR Interference (CRISPRi) for Sequence-specific Control of Gene Expression. *Nat. Protoc.* 8 (11), 2180–2196. doi:10.1038/nprot.2013.132

LeBlanc, C., Zhang, F., Mendez, J., Lozano, Y., Chatpar, K., Irish, V. F., et al. (2018). Increased Efficiency of Targeted Mutagenesis by CRISPR/Cas9 in Plants Using Heat Stress. *Plant J.* 93 (2), 377–386. doi:10.1111/tpj.13782

LeCher, J. C., Nowak, S. J., and McMurry, J. L. (2017). Breaking in and Busting Out: Cell-Penetrating Peptides and the Endosomal Escape Problem. *Biomol. Concepts* 8 (3-4), 131–141. doi:10.1515/bmc-2017-0023

Lee, J. K., Jeong, E., Lee, J., Jung, M., Shin, E., Kim, Y.-h., et al. (2018). Directed Evolution of CRISPR-Cas9 to Increase its Specificity. *Nat. Commun.* 9 (1), 3048. doi:10.1038/s41467-018-05477-x

Lee, S., Kim, Y.-Y., and Ahn, H. J. (2021). Systemic Delivery of CRISPR/Cas9 to Hepatic Tumors for Cancer Treatment Using Altered Tropism of Lentiviral Vector. *Biomaterials* 272, 120793. doi:10.1016/j.biomaterials.2021.120793

Legut, M., Daniloski, Z., Xue, X., McKenzie, D., Guo, X., Wessels, H.-H., et al. (2020). High-Throughput Screens of PAM-Flexible Cas9 Variants for Gene Knockout and Transcriptional Modulation. *Cel Rep.* 30 (9), 2859–2868. doi:10.1016/j.celrep.2020.02.010

Li, C., and Kasinski, A. L. (2020). *In Vivo* Cancer-Based Functional Genomics. *Trends Cancer* 6 (12), 1002–1017. doi:10.1016/j.trecan.2020.07.004

Li, G., Liu, D., Zhang, X., Quan, R., Zhong, C., Mo, J., et al. (2018a). Suppressing Ku70/Ku80 Expression Elevates Homology-Directed Repair Efficiency in Primary Fibroblasts. *Int. J. Biochem. Cel Biol.* 99, 154–160. doi:10.1016/j.biocel.2018.04.011

Li, G., Zhang, X., Zhong, C., Mo, J., Quan, R., Yang, J., et al. (2017). Small Molecules Enhance CRISPR/Cas9-mediated Homology-Directed Genome Editing in Primary Cells. *Sci. Rep.* 7 (1), 8943. doi:10.1038/s41598-017-09306-x

Li, L., He, Z.-Y., Wei, X.-W., Gao, G.-P., and Wei, Y.-Q. (2015). Challenges in CRISPR/CAS9 Delivery: Potential Roles of Nonviral Vectors. *Hum. Gene Ther.* 26 (7), 452–462. doi:10.1089/hum.2015.069

Li, L., Hu, S., and Chen, X. (2018b). Non-viral Delivery Systems for CRISPR/Cas9-based Genome Editing: Challenges and Opportunities. *Biomaterials* 171, 207–218. doi:10.1016/j.biomaterials.2018.04.031

Li, Y., Bolinger, J., Yu, Y., Glass, Z., Shi, N., Yang, L., et al. (2019). Intracellular Delivery and Biodistribution Study of CRISPR/Cas9 Ribonucleoprotein Loaded Bioreducible Lipidoid Nanoparticles. *Biomater. Sci.* 7 (2), 596–606. doi:10.1039/c8bm00637g

Li, Y., Yang, T., Yu, Y., Shi, N., Yang, L., Glass, Z., et al. (2018c). Combinatorial Library of Chalcogen-Containing Lipidoids for Intracellular Delivery of Genome-Editing Proteins. *Biomaterials* 178, 652–662. doi:10.1016/j.biomaterials.2018.03.011

Liang, X., Potter, J., Kumar, S., Ravinder, N., and Chesnut, J. D. (2017). Enhanced CRISPR/Cas9-mediated Precise Genome Editing by Improved Design and Delivery of gRNA, Cas9 Nuclease, and Donor DNA. *J. Biotechnol.* 241, 136–146. doi:10.1016/j.jbiotec.2016.11.011

Liesche, C., Venkatraman, L., Aschenbrenner, S., Grosse, S., Grimm, D., Eils, R., et al. (2016). Death Receptor-Based Enrichment of Cas9-Expressing Cells. *BMC Biotechnol.* 16, 17. doi:10.1186/s12896-016-0250-4

Lim, D., Sreekanth, V., Cox, K. J., Law, B. K., Wagner, B. K., Karp, J. M., et al. (2020). Engineering Designer Beta Cells with a CRISPR-Cas9 Conjugation Platform. *Nat. Commun.* 11 (1), 4043. doi:10.1038/s41467-020-17725-0

Lin, S., Staahl, B. T., Alla, R. K., and Doudna, J. A. (2014a). Enhanced Homology-Directed Human Genome Engineering by Controlled Timing of CRISPR/Cas9 Delivery. *Elife* 3, e04766. doi:10.7554/eLife.04766

Lin, Y., Cradick, T. J., Brown, M. T., Deshmukh, H., Ranjan, P., Sarode, N., et al. (2014b). CRISPR/Cas9 Systems Have Off-Target Activity with Insertions or Deletions between Target DNA and Guide RNA Sequences. *Nucleic Acids Res.* 42 (11), 7473–7485. doi:10.1093/nar/gku402

Liu, A., Huang, X., He, W., Xue, F., Yang, Y., Liu, J., et al. (2021a). pHmScarlet Is a pH-Sensitive Red Fluorescent Protein to Monitor Exocytosis Docking and Fusion Steps. *Nat. Commun.* 12 (1), 1413. doi:10.1038/s41467-021-21666-7

Liu, G., Zhang, Y., and Zhang, T. (2020a). Computational Approaches for Effective CRISPR Guide RNA Design and Evaluation. *Comput. Struct. Biotechnol. J.* 18, 35–44. doi:10.1016/j.csbj.2019.11.006

Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K., and Chen, L.-L. (2017a). CRISPR-P 2.0: An Improved CRISPR-Cas9 Tool for Genome Editing in Plants. *Mol. Plant* 10 (3), 530–532. doi:10.1016/j.molp.2017.01.003

Liu, J., Gaj, T., Patterson, J. T., Sirk, S. J., and Barbas III, C. F., 3rd (2014). Cell-penetrating Peptide-Mediated Delivery of TALEN Proteins via Bioconjugation for Genome Engineering. *PLoS One* 9 (1), e85755. doi:10.1371/journal.pone.0085755

Liu, L., Yin, M., Wang, M., and Wang, Y. (2019). Phage AcrIIA2 DNA Mimicry: Structural Basis of the CRISPR and Anti-CRISPR Arms Race. *Mol. Cel* 73 (3), 611–620. e3. doi:10.1016/j.molcel.2018.11.011

Liu, M., Rehman, S., Tang, X., Gu, K., Fan, Q., Chen, D., et al. (2018a). Methodologies for Improving HDR Efficiency. *Front. Genet.* 9, 691. doi:10.3389/fgene.2018.00691

Liu, P.-F., Zhao, K.-R., Liu, Z.-J., Wang, L., Ye, S.-Y., and Liang, G.-X. (2021b). Cas12a-based Electrochemiluminescence Biosensor for Target Amplification-free DNA Detection. *Biosens. Bioelectron.* 176, 112954. doi:10.1016/j.bios.2020.112954

Liu, Q., Chen, F., Hou, L., Shen, L., Zhang, X., Wang, D., et al. (2018b). Nanocarrier-Mediated Chemo-Immunotherapy Arrested Cancer Progression and Induced Tumor Dormancy in Desmoplastic Melanoma. *ACS Nano* 12 (8), 7812–7825. doi:10.1021/acsnano.8b01890

Liu, Q., Gao, R., Li, J., Lin, L., Zhao, J., Sun, W., et al. (2017b). Development of a Genome-Editing CRISPR/Cas9 System in Thermophilic Fungal Myceliophthora Species and its Application to Hyper-Cellulase Production Strain Engineering. *Biotechnol. Biofuels* 10, 1. doi:10.1186/s13068-016-0693-9

Liu, X., Dong, Y., Wang, J., Li, L., Zhong, Z., Li, Y.-p., et al. (2017c). VSV-G Viral Envelope Glycoprotein Prepared from Pichia pastoris Enhances Transfection of

DNA into Animal Cells. *J. Microbiol. Biotechnol.* 27 (6), 1098–1105. doi:10.4014/jmb.1611.11082

Liu, X., Qiu, S., Mei, L., Jing, H., Lin, X., and Wang, Q. (2020b). A High-Resolution Melting Analysis with an Unlabeled Probe for CRISPR/Cas9-Induced ZBED6 Knockout Pigs Detection. *J. AOAC Int.* 104, 541–545. doi:10.1093/jaoacint/qsaa161

Liu, X. S., Wu, H., Krzisch, M., Wu, X., Graef, J., Muffat, J., et al. (2018c). Rescue of Fragile X Syndrome Neurons by DNA Methylation Editing of the FMR1 Gene. *Cell* 172 (5), 979–992. doi:10.1016/j.cell.2018.01.012

Lok, B. H., and Powell, S. N. (2012). Molecular Pathways: Understanding the Role of Rad52 in Homologous Recombination for Therapeutic Advancement. *Clin. Cancer Res.* 18 (23), 6400–6406. doi:10.1158/1078-0432.ccr-11-3150

Lomova, A., Clark, D. N., Campo-Fernandez, B., Flores-Bjurström, C., Kaufman, M. L., Fitz-Gibbon, S., et al. (2019). Improving Gene Editing Outcomes in Human Hematopoietic Stem and Progenitor Cells by Temporal Control of DNA Repair. *Stem Cells* 37 (2), 284–294. doi:10.1002/stem.2935

Lovendahl, K. N., Hayward, A. N., and Gordon, W. R. (2017). Sequence-Directed Covalent Protein-DNA Linkages in a Single Step Using HUH-Tags. *J. Am. Chem. Soc.* 139 (20), 7030–7035. doi:10.1021/jacs.7b02572

Lu, Y., Tian, Y., Shen, R., Yao, Q., Wang, M., Chen, M., et al. (2020). Targeted, Efficient Sequence Insertion and Replacement in rice. *Nat. Biotechnol.* 38 (12), 1402–1407. doi:10.1038/s41587-020-0581-5

Lv, J., He, B., Yu, J., Wang, Y., Wang, C., Zhang, S., et al. (2018). Fluoropolymers for Intracellular and *In Vivo* Protein Delivery. *Biomaterials* 182, 167–175. doi:10.1016/j.biomaterials.2018.08.023

Ma, L., Ruan, J., Song, J., Wen, L., Yang, D., Zhao, J., et al. (2020). MiCas9 Increases Large Size Gene Knock-In Rates and Reduces Undesirable On-Target and Off-Target Indel Edits. *Nat. Commun.* 11 (1), 6082. doi:10.1038/s41467-020-19842-2

Ma, M., Zhuang, F., Hu, X., Wang, B., Wen, X.-Z., Ji, J.-F., et al. (2017a). Efficient Generation of Mice Carrying Homozygous Double-Floxp Alleles Using the Cas9-Avidin/Biotin-Donor DNA System. *Cell Res* 27 (4), 578–581. doi:10.1038/cr.2017.29

Ma, X., Hartmann, R., Jimenez de Aberasturi, D., Yang, F., Soenen, S. J. H., Manshian, B. B., et al. (2017b). Colloidal Gold Nanoparticles Induce Changes in Cellular and Subcellular Morphology. *ACS Nano* 11 (8), 7807–7820. doi:10.1021/acsnano.7b01760

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., et al. (2011). Evolution and Classification of the CRISPR-Cas Systems. *Nat. Rev. Microbiol.* 9 (6), 467–477. doi:10.1038/nrmicro2577

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J. J., et al. (2020). Evolutionary Classification of CRISPR-Cas Systems: a Burst of Class 2 and Derived Variants. *Nat. Rev. Microbiol.* 18 (2), 67–83. doi:10.1038/s41579-019-0299-x

Mali, P., Esvelt, K. M., and Church, G. M. (2013). Cas9 as a Versatile Tool for Engineering Biology. *Nat. Methods* 10 (10), 957–963. doi:10.1038/nmeth.2649

Malone, L. M., Birkholz, N., and Fineran, P. C. (2020). Conquering CRISPR: How Phages Overcome Bacterial Adaptive Immunity. *Curr. Opin. Biotechnol.* 68, 30–36. doi:10.1016/j.copbio.2020.09.008

Mangeot, P. E., Risson, V., Fusil, F., Marnef, A., Laurent, E., Blin, J., et al. (2019). Genome Editing in Primary Cells and *In Vivo* Using Viral-Derived Nanoblades Loaded with Cas9-sgRNA Ribonucleoproteins. *Nat. Commun.* 10 (1), 45. doi:10.1038/s41467-018-07845-z

Manghwar, H., Li, B., Ding, X., Hussain, A., Lindsey, K., Zhang, X., et al. (2020). CRISPR/Cas Systems in Genome Editing: Methodologies and Tools for sgRNA Design, Off-Target Evaluation, and Strategies to Mitigate Off-Target Effects. *Adv. Sci.* 7 (6), 1902312. doi:10.1002/advs.201902312

Manjunath, M., Choudhary, B., and Raghavan, S. C. (2021). SCR7, a Potent Cancer Therapeutic Agent and a Biochemical Inhibitor of Nonhomologous DNA End-joining. *Cancer Rep.* 4, e1341. doi:10.1002/cnr2.1341

Maruyama, T., Dougan, S. K., Truttmann, M. C., Bilate, A. M., Ingram, J. R., and Ploegh, H. L. (2015). Increasing the Efficiency of Precise Genome Editing with CRISPR-Cas9 by Inhibition of Nonhomologous End Joining. *Nat. Biotechnol.* 33 (5), 538–542. doi:10.1038/nbt.3190

Mashel, T. V., Tarakanchikova, Y. V., Muslimov, A. R., Zyuzin, M. V., Timin, A. S., Lepik, K. V., et al. (2020). Overcoming the Delivery Problem for Therapeutic Genome Editing: Current Status and Perspective of Non-viral Methods. *Biomaterials* 258, 120282. doi:10.1016/j.biomaterials.2020.120282

Mateescu, B., Kowal, E. J. K., van Balkom, B. W. M., Bartel, S., Bhattacharyya, S. N., Buzás, E. I., et al. (2017). Obstacles and Opportunities in the Functional Analysis of Extracellular Vesicle RNA - an ISEV Position Paper. *J. Extracellular Vesicles* 6 (1), 1286095. doi:10.1080/20013078.2017.1286095

Mateos-Gomez, P. A., Kent, T., Deng, S. K., McDevitt, S., Kashkina, E., Hoang, T. M., et al. (2017). The Helicase Domain of Polθ Counteracts RPA to Promote Alt-NHEJ. *Nat. Struct. Mol. Biol.* 24 (12), 1116–1123. doi:10.1038/nsmb.3494

Mathieu, M., Martin-Jaular, L., Lavieu, G., and Théry, C. (2019). Specificities of Secretion and Uptake of Exosomes and Other Extracellular Vesicles for Cell-To-Cell Communication. *Nat. Cel Biol* 21 (1), 9–17. doi:10.1038/s41556-018-0250-9

Mátrai, J., Chuah, M. K., and VandenDriessche, T. (2010). Recent Advances in Lentiviral Vector Development and Applications. *Mol. Ther.* 18 (3), 477–490. doi:10.1038/mt.2009.319

Miao, Z., Gao, Z., Chen, R., Yu, X., Su, Z., and Wei, G. (2018). Surface-bioengineered Gold Nanoparticles for Biomedical Applications. *Cmc* 25 (16), 1920–1944. doi:10.2174/0929867325666180117111404

Minkenberg, B., Zhang, J., Xie, K., and Yang, Y. (2019). CRISPR - PLANT V2: an Online Resource for Highly Specific Guide RNA Spacers Based on Improved Off-target Analysis. *Plant Biotechnol. J.* 17 (1), 5–8. doi:10.1111/pbi.13025

Miura, H., Quadros, R. M., Gurumurthy, C. B., and Ohtsuka, M. (2018). Easi-CRISPR for Creating Knock-In and Conditional Knockout Mouse Models Using Long ssDNA Donors. *Nat. Protoc.* 13 (1), 195–215. doi:10.1038/nprot.2017.153

Montagna, C., Petris, G., Casini, A., Maule, G., Franceschini, G. M., Zanella, I., et al. (2018). VSV-G-Enveloped Vesicles for Traceless Delivery of CRISPR-Cas9. *Mol. Ther. - Nucleic Acids* 12, 453–462. doi:10.1016/j.omtn.2018.05.010

Moriarity, B. S., Rahrmann, E. P., Beckmann, D. A., Conboy, C. B., Watson, A. L., Carlson, D. F., et al. (2014). Simple and Efficient Methods for Enrichment and Isolation of Endonuclease Modified Cells. *PLoS One* 9 (5), e96114. doi:10.1371/journal.pone.0096114

Mulsant, P., Gatignol, A., Dalens, M., and Tiraby, G. r. (1988). Phleomycin Resistance as a Dominant Selectable Marker in CHO Cells. *Somat Cel Mol Genet* 14 (3), 243–252. doi:10.1007/bf01534585

Nishimasu, S., Shi, X., Ishiguro, S., Gao, L., Hirano, S., Okazaki, S., et al. (2018). Engineered CRISPR-Cas9 Nuclease with Expanded Targeting Space. *Science* 361, 1259–1262. doi:10.1126/science.aas9129

Nihongaki, Y., Kawano, F., Nakajima, T., and Sato, M. (2015). Photoactivatable CRISPR-Cas9 for Optogenetic Genome Editing. *Nat. Biotechnol.* 33 (7), 755–760. doi:10.1038/nbt.3245

Noll, L. W., Chall, R., Shridhar, P. B., Liu, X., Bai, J., Delannoy, S., et al. (2018). Validation and Application of a Real-Time PCR Assay Based on the CRISPR Array for Serotype-specific Detection and Quantification of Enterohemorrhagic *Escherichia coli* O157:H7 in Cattle Feces†. *J. Food Prot.* 81 (7), 1157–1164. doi:10.4315/0362-028x.jfp-18-049

Oh, J., Fung, E., Price, M. N., Dehal, P. S., Davis, R. W., Giaever, G., et al. (2010). A Universal TagModule Collection for Parallel Genetic Analysis of Microorganisms. *Nucleic Acids Res.* 38 (14), e146. doi:10.1093/nar/gkq419

Omichi, R., Shibata, S. B., Morton, C. C., and Smith, R. J. H. (2019). Gene Therapy for Hearing Loss. *Hum. Mol. Genet.* 28 (R1), R65–R79. doi:10.1093/hmg/ddz129

Pandey, P., Zhang, N., Curtis, B. R., Newman, P. J., and Denomme, G. A. (2021). Generation of 'designer Erythroblasts' Lacking One or More Blood Group Systems from CRISPR/Cas9 Gene-edited Human-induced Pluripotent Stem Cells. *J. Cel Mol Med* 25, 9340–9349. doi:10.1111/jcmm.16872

Panier, S., and Boulton, S. J. (2014). Double-strand Break Repair: 53BP1 Comes into Focus. *Nat. Rev. Mol. Cel Biol* 15 (1), 7–18. doi:10.1038/nrm3719

Pascucci, L., Coccè, V., Bonomi, A., Ami, D., Ceccarelli, P., Ciusani, E., et al. (2014). Paclitaxel Is Incorporated by Mesenchymal Stromal Cells and Released in Exosomes that Inhibit *In Vitro* Tumor Growth: a New Approach for Drug Delivery. *J. Controlled Release* 192, 262–270. doi:10.1016/j.jconrel.2014.07.042

Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. A., and Liu, D. R. (2013). High-throughput Profiling of Off-Target DNA Cleavage Reveals RNA-Programmed Cas9 Nuclease Specificity. *Nat. Biotechnol.* 31 (9), 839–843. doi:10.1038/nbt.2673

Paulsen, B. S., Mandal, P. K., Frock, R. L., Boyraz, B., Yadav, R., Upadhyayula, S., et al. (2017). Ectopic Expression of RAD52 and dn53BP1 Improves Homology-

Directed Repair during CRISPR-Cas9 Genome Editing. *Nat. Biomed. Eng.* 1 (11), 878–888. doi:10.1038/s41551-017-0145-2

Pawelczak, K. S., Gavande, N. S., VanderVere-Carozza, P. S., and Turchi, J. J. (2018). Modulating DNA Repair Pathways to Improve Precision Genome Engineering. *ACS Chem. Biol.* 13 (2), 389–396. doi:10.1021/acschembio.7b00777

Pena, S. A., Iyengar, R., Eshraghi, R. S., Bencie, N., Mittal, J., Aljohani, A., et al. (2020). Gene Therapy for Neurological Disorders: Challenges and Recent Advancements. *J. Drug Target.* 28 (2), 111–128. doi:10.1080/1061186x.2019.1630415

Peng, H., Le, C., Wu, J., Li, X.-F., Zhang, H., and Le, X. C. (2020). A Genome-Editing Nanomachine Constructed with a Clustered Regularly Interspaced Short Palindromic Repeats System and Activated by Near-Infrared Illumination. *ACS Nano* 14 (3), 2817–2826. doi:10.1021/acsnano.9b05276

Platt, R. J., Chen, S., Zhou, Y., Yim, M. J., Swiech, L., Kempton, H. R., et al. (2014). CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling. *Cell* 159 (2), 440–455. doi:10.1016/j.cell.2014.09.014

Priya, S. S., Rekha, M. R., and Sharma, C. P. (2014). Pullulan-protamine as Efficient Haemocompatible Gene Delivery Vector: Synthesis and *In Vitro* Characterization. *Carbohydr. Polym.* 102, 207–215. doi:10.1016/j.carbpol.2013.11.024

Rabu, C., Rangan, L., Florenceau, L., Fortun, A., Charpentier, M., Dupré, E., et al. (2019). Cancer Vaccines: Designing Artificial Synthetic Long Peptides to Improve Presentation of Class I and Class II T Cell Epitopes by Dendritic Cells. *Oncoimmunology* 8 (4), e1560919. doi:10.1080/2162402x.2018.1560919

Ramakrishna, S., Cho, S. W., Kim, S., Song, M., Gopalappa, R., Kim, J.-S., et al. (2014a). Surrogate Reporter-Based Enrichment of Cells Containing RNA-Guided Cas9 Nuclease-Induced Mutations. *Nat. Commun.* 5, 3378. doi:10.1038/ncomms4378

Ramakrishna, S., Kwaku Dad, A.-B., Beloor, J., Gopalappa, R., Lee, S.-K., and Kim, H. (2014b). Gene Disruption by Cell-Penetrating Peptide-Mediated Delivery of Cas9 Protein and Guide RNA. *Genome Res.* 24 (6), 1020–1027. doi:10.1101/gr.171264.113

Rauch, S., He, C., and Dickinson, B. C. (2018). Targeted m6A Reader Proteins to Study Epitranscriptomic Regulation of Single RNAs. *J. Am. Chem. Soc.* 140 (38), 11974–11981. doi:10.1021/jacs.8b05012

Ren, C., Xu, K., Liu, Z., Shen, J., Han, F., Chen, Z., et al. (2015). Dual-reporter Surrogate Systems for Efficient Enrichment of Genetically Modified Cells. *Cell. Mol. Life Sci.* 72 (14), 2763–2772. doi:10.1007/s00018-015-1874-6

Ren, C., Xu, K., Segal, D. J., and Zhang, Z. (2019). Strategies for the Enrichment and Selection of Genetically Modified Cells. *Trends Biotechnol.* 37 (1), 56–71. doi:10.1016/j.tibtech.2018.07.017

Renaud, J.-B., Boix, C., Charpentier, M., De Cian, A., Cochennec, J., Duvernois-Berthet, E., et al. (2016). Improved Genome Editing Efficiency and Flexibility Using Modified Oligonucleotides with TALEN and CRISPR-Cas9 Nucleases. *Cel Rep.* 14 (9), 2263–2272. doi:10.1016/j.celrep.2016.02.018

Richter, M. F., Zhao, K. T., Eton, E., Lapinaite, A., Newby, G. A., Thuronyi, B. W., et al. (2020). Phage-assisted Evolution of an Adenine Base Editor with Improved Cas Domain Compatibility and Activity. *Nat. Biotechnol.* 38 (7), 883–891. doi:10.1038/s41587-020-0453-z

Robert, F., Barbeau, M., Éthier, S., Dostie, J., and Pelletier, J. (2015). Pharmacological Inhibition of DNA-PK Stimulates Cas9-Mediated Genome Editing. *Genome Med.* 7, 93. doi:10.1186/s13073-015-0215-6

Ruan, G.-X., Barry, E., Yu, D., Lukason, M., Cheng, S. H., and Scaria, A. (2017). CRISPR/Cas9-Mediated Genome Editing as a Therapeutic Approach for Leber Congenital Amaurosis 10. *Mol. Ther.* 25 (2), 331–341. doi:10.1016/j.ymthe.2016.12.006

Rust, B. J., Becker, P. S., Chandrasekaran, D., Kubek, S. P., Peterson, C. W., Adair, J. E., et al. (2020). Envelope-Specific Adaptive Immunity Following Transplantation of Hematopoietic Stem Cells Modified with VSV-G Lentivirus. *Mol. Ther. - Methods Clin. Develop.* 19, 438–446. doi:10.1016/j.omtm.2020.10.002

Ryu, J.-Y., Won, E.-J., Lee, H. A. R., Kim, J. H., Hui, E., Kim, H. P., et al. (2020). Ultrasound-activated Particles as CRISPR/Cas9 Delivery System for Androgenic Alopecia Therapy. *Biomaterials* 232, 119736. doi:10.1016/j.biomaterials.2019.119736

Saari, H., Lázaro-Ibáñez, E., Viitala, T., Vuorimaa-Laukkanen, E., Siljander, P., and Yliperttula, M. (2015). Microvesicle- and Exosome-Mediated Drug Delivery Enhances the Cytotoxicity of Paclitaxel in Autologous Prostate Cancer Cells. *J. Controlled Release* 220 (Pt B), 727–737. doi:10.1016/j.jconrel.2015.09.031

Sander, J. D., and Joung, J. K. (2014). CRISPR-cas Systems for Editing, Regulating and Targeting Genomes. *Nat. Biotechnol.* 32 (4), 347–355. doi:10.1038/nbt.2842

Savic, N., Ringnalda, F. C., Lindsay, H., Berk, C., Bargsten, K., Li, Y., et al. (2018). Covalent Linkage of the DNA Repair Template to the CRISPR-Cas9 Nuclease Enhances Homology-Directed Repair. *Elife* 7, e33761. doi:10.7554/elife.33761

Sena-Esteves, M., and Gao, G. (2018). Titration of Lentivirus Vectors. *Cold Spring Harb Protoc.* 2018 (4), pdb.prot095695. doi:10.1101/pdb.prot095695

Shao, S., Ren, C., Liu, Z., Bai, Y., Chen, Z., Wei, Z., et al. (2017). Enhancing CRISPR/Cas9-mediated Homology-Directed Repair in Mammalian Cells by Expressing *Saccharomyces cerevisiae* Rad52. *Int. J. Biochem. Cel Biol.* 92, 43–52. doi:10.1016/j.biocel.2017.09.012

Shen, Z. Y., Xia, G. L., Wu, M. F., Ji, L. Y., and Li, Y. J. (2016). The Effects of Percutaneous Ethanol Injection Followed by 20-kHz Ultrasound and Microbubbles on Rabbit Hepatic Tumors. *J. Cancer Res. Clin. Oncol.* 142 (2), 373–378. doi:10.1007/s00432-015-2034-y

Shi, J., Ma, Y., Zhu, J., Chen, Y., Sun, Y., Yao, Y., et al. (2018). A Review on Electroporation-Based Intracellular Delivery. *Molecules* 23 (11), 3044. doi:10.3390/molecules23113044

Shirley, J. L., de Jong, Y. P., Terhorst, C., and Herzog, R. W. (2020). Immune Responses to Viral Gene Therapy Vectors. *Mol. Ther.* 28 (3), 709–722. doi:10.1016/j.ymthe.2020.01.001

Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X., and Zhang, F. (2016). Rationally Engineered Cas9 Nucleases with Improved Specificity. *Science* 351 (6268), 84–88. doi:10.1126/science.aad5227

Søndergaard, J. N., Geng, K., Sommerauer, C., Atanasoai, I., Yin, X., and Kutter, C. (2020). Successful Delivery of Large-Size CRISPR/Cas9 Vectors in Hard-To-Transfect Human Cells Using Small Plasmids. *Commun. Biol.* 3 (1), 319. doi:10.1038/s42003-020-1045-7

Song, F., and Stieger, K. (2017). Optimizing the DNA Donor Template for Homology-Directed Repair of Double-Strand Breaks. *Mol. Ther. - Nucleic Acids* 7, 53–60. doi:10.1016/j.omtn.2017.02.006

Song, J., Yang, D., Xu, J., Zhu, T., Chen, Y. E., and Zhang, J. (2016). RS-1 Enhances CRISPR/Cas9- and TALEN-Mediated Knock-In Efficiency. *Nat. Commun.* 7, 10548. doi:10.1038/ncomms10548

Song, X., Liu, C., Wang, N., Huang, H., He, S., Gong, C., et al. (2021). Delivery of CRISPR/Cas Systems for Cancer Gene Therapy and Immunotherapy. *Adv. Drug Deliv. Rev.* 168, 158–180. doi:10.1016/j.addr.2020.04.010

Sottnik, J. L., Vanderlinden, L., Joshi, M., Chauca-Diaz, A., Owens, C., Hansel, D. E., et al. (2021). Androgen Receptor Regulates CD44 Expression in Bladder Cancer. *Cancer Res.* 81, 2833–2846. doi:10.1158/0008-5472.CAN-20-3095

Srivastava, M., Nambiar, M., Sharma, S., Karki, S. S., Goldsmith, G., Hegde, M., et al. (2012). An Inhibitor of Nonhomologous End-Joining Abrogates Double-Strand Break Repair and Impedes Cancer Progression. *Cell* 151 (7), 1474–1487. doi:10.1016/j.cell.2012.11.054

Sunada, S., Saito, H., Zhang, D., Xu, Z., and Miki, Y. (2021). CDK1 Inhibitor Controls G2/M Phase Transition and Reverses DNA Damage Sensitivity. *Biochem. Biophysical Res. Commun.* 550, 56–61. doi:10.1016/j.bbrc.2021.02.117

Suresh, B., Ramakrishna, S., and Kim, H. (2017). Cell-Penetrating Peptide-Mediated Delivery of Cas9 Protein and Guide RNA for Genome Editing. *Methods Mol. Biol.* 1507, 81–94. doi:10.1007/978-1-4939-6518-2_7

Suzuki, Y., Onuma, H., Sato, R., Sato, Y., Hashiba, A., Maeki, M., et al. (2021). Lipid Nanoparticles Loaded with Ribonucleoprotein-Oligonucleotide Complexes Synthesized Using a Microfluidic Device Exhibit Robust Genome Editing and Hepatitis B Virus Inhibition. *J. Control. Release* 330, 61–71. doi:10.1016/j.jconrel.2020.12.013

Svitashev, S., Schwartz, C., Lenderts, B., Young, J. K., and Mark Cigan, A. (2016). Genome Editing in maize Directed by CRISPR-Cas9 Ribonucleoprotein Complexes. *Nat. Commun.* 7, 13274. doi:10.1038/ncomms13274

Symington, L. S. (2014). End Resection at Double-Strand Breaks: Mechanism and Regulation. *Cold Spring Harb Perspect. Biol.* 6 (8), a016436. doi:10.1101/cshperspect.a016436

Taharabaru, T., Yokoyama, R., Higashi, T., Mohammed, A. F. A., Inoue, M., Maeda, Y., et al. (2020). Genome Editing in a Wide Area of the Brain Using Dendrimer-Based Ternary Polyplexes of Cas9 Ribonucleoprotein. *ACS Appl. Mater. Inter.* 12 (19), 21386–21397. doi:10.1021/acsami.9b21667

Tang, L., Zeng, Y., Du, H., Gong, M., Peng, J., Zhang, B., et al. (2017). CRISPR/Cas9-mediated Gene Editing in Human Zygotes Using Cas9 Protein. *Mol. Genet. Genomics* 292 (3), 525–533. doi:10.1007/s00438-017-1299-z

Taniyama, Y., Azuma, J., Rakugi, H., and Morishita, R. (2011). Plasmid DNA-Based Gene Transfer with Ultrasound and Microbubbles. *Curr. Gene Ther.* 11 (6), 485–490. doi:10.2174/156652311798192851

Tao, Y., Yi, K., Hu, H., Shao, D., and Li, M. (2021). Coassembly of Nucleus-Targeting Gold Nanoclusters with CRISPR/Cas9 for Simultaneous Bioimaging and Therapeutic Genome Editing. *J. Mater. Chem. B* 9 (1), 94–100. doi:10.1039/d0tb01925a

Torre, E. A., Arai, E., Bayatpour, S., Jiang, C. L., Beck, L. E., Emert, B. L., et al. (2021). Genetic Screening for Single-Cell Variability Modulators Driving Therapy Resistance. *Nat. Genet.* 53 (1), 76–85. doi:10.1038/s41588-020-00749-z

Tran, N.-T., Bashir, S., Li, X., Rossius, J., Chu, V. T., Rajewsky, K., et al. (2019). Enhancement of Precise Gene Editing by the Association of Cas9 with Homologous Recombination Factors. *Front. Genet.* 10, 365. doi:10.3389/fgene.2019.00365

Urnov, F. D., Miller, J. C., Lee, Y.-L., Beausejour, C. M., Rock, J. M., Augustus, S., et al. (2005). Highly Efficient Endogenous Human Gene Correction Using Designed Zinc-finger Nucleases. *Nature* 435 (7042), 646–651. doi:10.1038/nature03556

Vakulskas, C. A., Dever, D. P., Rettig, G. R., Turk, R., Jacobi, A. M., Collingwood, M. A., et al. (2018). A High-Fidelity Cas9 Mutant Delivered as a Ribonucleoprotein Complex Enables Efficient Gene Editing in Human Hematopoietic Stem and Progenitor Cells. *Nat. Med.* 24 (8), 1216–1224. doi:10.1038/s41591-018-0137-0

Vasquez, R. J., Howell, B., Yvon, A. M., Wadsworth, P., and Cassimeris, L. (1997). Nanomolar Concentrations of Nocodazole Alter Microtubule Dynamic Instability *In Vivo* and *In Vitro*. *Mol. Biol. Cel* 8 (6), 973–985. doi:10.1091/mbc.8.6.973

Vassilev, L. T. (2006). Cell Cycle Synchronization at the G2/M Phase Border by Reversible Inhibition of CDK1. *Cell Cycle* 5 (22), 2555–2556. doi:10.4161/cc.5.22.3463

Vermersch, E., Jouve, C., and Hulot, J.-S. (2020). CRISPR/Cas9 Gene-Editing Strategies in Cardiovascular Cells. *Cardiovasc. Res.* 116 (5), 894–907. doi:10.1093/cvr/cvz250

Wang, H., La Russa, M., and Qi, L. S. (2016). CRISPR/Cas9 in Genome Editing and beyond. *Annu. Rev. Biochem.* 85, 227–264. doi:10.1146/annurev-biochem-060815-014607

Wang, P., Zhang, L., Zheng, W., Cong, L., Guo, Z., Xie, Y., et al. (2018). Thermo-triggered Release of CRISPR-Cas9 System by Lipid-Encapsulated Gold Nanoparticles for Tumor Therapy. *Angew. Chem. Int. Ed.* 57 (6), 1491–1496. doi:10.1002/anie.201708689

Wei, Q., Croy, B. A., and Etches, R. J. (2001). Selection of Genetically Modified Chicken Blastodermal Cells by Magnetic-Activated Cell Sorting. *Poult. Sci.* 80 (12), 1671–1678. doi:10.1093/ps/80.12.1671

Wienert, B., Nguyen, D. N., Guenther, A., Feng, S. J., Locke, M. N., Wyman, S. K., et al. (2020). Timed Inhibition of CDC7 Increases CRISPR-Cas9 Mediated Templated Repair. *Nat. Commun.* 11 (1), 2109. doi:10.1038/s41467-020-15845-1

Wilbie, D., Walther, J., and Mastrobattista, E. (2019). Delivery Aspects of CRISPR/Cas for *In Vivo* Genome Editing. *Acc. Chem. Res.* 52 (6), 1555–1564. doi:10.1021/acs.accounts.9b00106

Wold, W. S., and Toth, K. (2013). Adenovirus Vectors for Gene Therapy, Vaccination and Cancer Gene Therapy. *Curr. Gene Ther.* 13 (6), 421–433. doi:10.2174/15665232113666131125095046

Wu, X., Kriz, A. J., and Sharp, P. A. (2014). Target Specificity of the CRISPR-Cas9 System. *Quant Biol.* 2 (2), 59–70. doi:10.1007/s40484-014-0030-x

Wu, Z., Yang, H., and Colosi, P. (2010). Effect of Genome Size on AAV Vector Packaging. *Mol. Ther.* 18 (1), 80–86. doi:10.1038/mt.2009.255

Xiang, G., Zhang, X., An, C., Cheng, C., and Wang, H. (2017). Temperature Effect on CRISPR-Cas9 Mediated Genome Editing. *J. Genet. Genomics* 44 (4), 199–205. doi:10.1016/j.jgg.2017.03.004

Xie, X., Ma, X., Zhu, Q., Zeng, D., Li, G., and Liu, Y.-G. (2017). CRISPR-GE: A Convenient Software Toolkit for CRISPR-Based Genome Editing. *Mol. Plant* 10 (9), 1246–1249. doi:10.1016/j.molp.2017.06.004

Xu, C. L., Ruan, M. Z. C., Mahajan, V. B., and Tsang, S. H. (2019). Viral Delivery Systems for CRISPR. *Viruses* 11 (1). doi:10.3390/v11010028

Xu, Q., Zhang, Z., Zhao, L., Qin, Y., Cai, H., Geng, Z., et al. (2020a). Tropism-facilitated Delivery of CRISPR/Cas9 System with Chimeric Antigen Receptor-Extracellular Vesicles against B-Cell Malignancies. *J. Controlled Release* 326, 455–467. doi:10.1016/j.jconrel.2020.07.033

Xu, W. (2019). Microinjection and Micromanipulation: A Historical Perspective. *Methods Mol. Biol.* 1874, 1–16. doi:10.1007/978-1-4939-8831-0_1

Xu, X., Hulshoff, M. S., Tan, X., Zeisberg, M., and Zeisberg, E. M. (2020b). CRISPR/Cas Derivatives as Novel Gene Modulating Tools: Possibilities and *In Vivo* Applications. *Int. J. Mol. Sci.* 21 (9), 3038. doi:10.3390/ijms21093038

Xu, X., Tan, X., Tampe, B., Wilhelmi, T., Hulshoff, M. S., Saito, S., et al. (2018). High-fidelity CRISPR/Cas9- Based Gene-specific Hydroxymethylation Rescues Gene Expression and Attenuates Renal Fibrosis. *Nat. Commun.* 9 (1), 3509. doi:10.1038/s41467-018-05766-5

Yang, D., Scavuzzo, M. A., Chmielowiec, J., Sharp, R., Bajic, A., and Borowiak, M. (2016). Enrichment of G2/M Cell Cycle Phase in Human Pluripotent Stem Cells Enhances HDR-Mediated Gene Repair with Customizable Endonucleases. *Sci. Rep.* 6, 21264. doi:10.1038/srep21264

Yang, H., Ren, S., Yu, S., Pan, H., Li, T., Ge, S., et al. (2020). Methods Favoring Homology-Directed Repair Choice in Response to CRISPR/Cas9 Induced-Double Strand Breaks. *Int. J. Mol. Sci.* 21 (18), 6461. doi:10.3390/ijms21186461

Yang, H., Wang, H., Shivalila, C. S., Cheng, A. W., Shi, L., and Jaenisch, R. (2013). One-step Generation of Mice Carrying Reporter and Conditional Alleles by CRISPR/Cas-mediated Genome Engineering. *Cell* 154 (6), 1370–1379. doi:10.1016/j.cell.2013.08.022

Yang, X., Shi, G., Guo, J., Wang, C., and He, Y. (2018). Exosome-encapsulated Antibiotic against Intracellular Infections of Methicillin-Resistant *Staphylococcus aureus*. *Int. J. Nanomedicine* 13, 8095–8104. doi:10.2147/ijn.s179380

Yao, R., Liu, D., Jia, X., Zheng, Y., Liu, W., and Xiao, Y. (2018a). CRISPR-Cas9/Cas12a Biotechnology and Application in Bacteria. *Synth. Syst. Biotechnol.* 3 (3), 135–149. doi:10.1016/j.synbio.2018.09.004

Yao, X., Lyu, P., Yoo, K., Yadav, M. K., Singh, R., Atala, A., et al. (2021). Engineered Extracellular Vesicles as Versatile Ribonucleoprotein Delivery Vehicles for Efficient and Safe CRISPR Genome Editing. *J. Extracell Vesicles* 10 (5), e12076. doi:10.1002/jev2.12076

Yao, X., Zhang, M., Wang, X., Ying, W., Hu, X., Dai, P., et al. (2018b). Tild-CRISPR Allows for Efficient and Precise Gene Knockin in Mouse and Human Cells. *Develop. Cel* 45 (4), 526–536. doi:10.1016/j.devcel.2018.04.021

Yasuda, H., Kim, E., Reza, A. M. M. T., and Kim, J.-H. (2016). A Highly Efficient Method for Enriching TALEN or CRISPR/Cas9-edited Mutant Cells. *J. Genet. Genomics* 43 (12), 705–708. doi:10.1016/j.jgg.2016.10.006

Yiangou, L., Grandy, R. A., Morell, C. M., Tomaz, R. A., Osnato, A., Kadiwala, J., et al. (2019). Method to Synchronize Cell Cycle of Human Pluripotent Stem Cells without Affecting Their Fundamental Characteristics. *Stem Cel Rep.* 12 (1), 165–179. doi:10.1016/j.stemcr.2018.11.020

Yin, D., Ling, S., Wang, D., Dai, Y., Jiang, H., Zhou, X., et al. (2021). Targeting Herpes Simplex Virus with CRISPR-Cas9 Cures Herpetic Stromal Keratitis in Mice. *Nat. Biotechnol.* 39, 567–577. doi:10.1038/s41587-020-00781-8

Yin, H., Kanasty, R. L., Eltoukhy, A. A., Vegas, A. J., Dorkin, J. R., and Anderson, D. G. (2014). Non-viral Vectors for Gene-Based Therapy. *Nat. Rev. Genet.* 15 (8), 541–555. doi:10.1038/nrg3763

Yin, H., Kauffman, K. J., and Anderson, D. G. (2017). Delivery Technologies for Genome Editing. *Nat. Rev. Drug Discov.* 16 (6), 387–399. doi:10.1038/nrd.2016.280

Yin, H., Song, C.-Q., Suresh, S., Kwan, S.-Y., Wu, Q., Walsh, S., et al. (2018a). Partial DNA-Guided Cas9 Enables Genome Editing with Reduced Off-Target Activity. *Nat. Chem. Biol.* 14 (3), 311–316. doi:10.1038/nchembio.2559

Yin, J., Hou, S., Wang, Q., Bao, L., Liu, D., Yue, Y., et al. (2019a). Microenvironment-Responsive Delivery of the Cas9 RNA-Guided Endonuclease for Efficient Genome Editing. *Bioconjug. Chem.* 30 (3), 898–906. doi:10.1021/acs.bioconjchem.9b00022

Yin, J., Liu, M., Liu, Y., Wu, J., Gan, T., Zhang, W., et al. (2019b). Optimizing Genome Editing Strategy by Primer-Extension-Mediated Sequencing. *Cell Discov* 5, 18. doi:10.1038/s41421-019-0088-8

Yin, J., Wang, Q., Hou, S., Bao, L., Yao, W., and Gao, X. (2018b). Potent Protein Delivery into Mammalian Cells via a Supercharged Polypeptide. *J. Am. Chem. Soc.* 140 (49), 17234–17240. doi:10.1021/jacs.8b10299

Yoon, S., Wang, P., Peng, Q., Wang, Y., and Shung, K. K. (2017). Acoustic-transfection for Genomic Manipulation of Single-Cells Using High Frequency Ultrasound. *Sci. Rep.* 7 (1), 5275. doi:10.1038/s41598-017-05722-1

Yoshimi, K., Kunihiro, Y., Kaneko, T., Nagahora, H., Voigt, B., and Mashimo, T. (2016). ssODN-mediated Knock-In with CRISPR-Cas for Large Genomic Regions in Zygotes. *Nat. Commun.* 7, 10431. doi:10.1038/ncomms10431

Yoshimi, K., Oka, Y., Miyasaka, Y., Kotani, Y., Yasumura, M., Uno, Y., et al. (2021). Combi-CRISPR: Combination of NHEJ and HDR Provides Efficient and Precise Plasmid-Based Knock-Ins in Mice and Rats. *Hum. Genet.* 140 (2), 277–287. doi:10.1007/s00439-020-02198-4

Yue, H., Zhou, X., Cheng, M., and Xing, D. (2018). Graphene Oxide-Mediated Cas9/sgRNA Delivery for Efficient Genome Editing. *Nanoscale* 10 (3), 1063–1071. doi:10.1039/c7nr07999k

Zhang, B., Yin, Y., Lai, R. C., Tan, S. S., Choo, A. B. H., and Lim, S. K. (2014). Mesenchymal Stem Cells Secrete Immunologically Active Exosomes. *Stem Cell Develop.* 23 (11), 1233–1244. doi:10.1089/scd.2013.0479

Zhang, F., Song, G., and Tian, Y. (2019a). Anti-CRISPRs: The Natural Inhibitors for CRISPR-Cas Systems. *Anim. Model Exp Med* 2 (2), 69–75. doi:10.1002/ame2.12069

Zhang, S., Shen, J., Li, D., and Cheng, Y. (2021). Strategies in the Delivery of Cas9 Ribonucleoprotein for CRISPR/Cas9 Genome Editing. *Theranostics* 11 (2), 614–648. doi:10.7150/thno.47007

Zhang, Y., and Showalter, A. M. (2020). CRISPR/Cas9 Genome Editing Technology: A Valuable Tool for Understanding Plant Cell Wall Biosynthesis and Function. *Front. Plant Sci.* 11, 589517. doi:10.3389/fpls.2020.589517

Zhang, Z., Wan, T., Chen, Y., Chen, Y., Sun, H., Cao, T., et al. (2019b). Cationic Polymer-Mediated CRISPR/Cas9 Plasmid Delivery for Genome Editing. *Macromol Rapid Commun.* 40 (5), e1800068. doi:10.1002/marc.201800068

Zheng, W., Li, Q., Sun, H., Ali, M. W., and Zhang, H. (2019). Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR-associated 9-mediated Mutagenesis of the Multiple Edematous Wings Gene Induces Muscle Weakness and Flightlessness in Bactrocera Dorsalis (Diptera: Tephritidae). *Insect Mol. Biol.* 28 (2), 222–234. doi:10.1111/imb.12540

Zhi, S., Chen, Y., Wu, G., Wen, J., Wu, J., Liu, Q., et al. (2022). Dual-AAV Delivering Split Prime Editor System for *In Vivo* Genome Editing. *Mol. Ther.* 30 (1), 283–294. doi:10.1016/j.ymthe.2021.07.011

Zhou, Q., Zhan, H., Liao, X., Fang, L., Liu, Y., Xie, H., et al. (2019). A Revolutionary Tool: CRISPR Technology Plays an Important Role in Construction of Intelligentized Gene Circuits. *Cell Prolif* 52 (2), e12552. doi:10.1111/cpr.12552

Zhou, S. J., Li, S. W., Wang, J. J., Liu, Z. J., Yin, G. B., Gong, J. P., et al. (2012). High-intensity Focused Ultrasound Combined with Herpes Simplex Virus Thymidine Kinase Gene-Loaded Ultrasound-Targeted Microbubbles Improved the Survival of Rabbits with VX_2 Liver Tumor. *J. Gene Med.* 14 (9-10), 570–579. doi:10.1002/jgm.2668

Zhuang, J., Tan, J., Wu, C., Zhang, J., Liu, T., Fan, C., et al. (2020). Extracellular Vesicles Engineered with Valency-Controlled DNA Nanostructures Deliver CRISPR/Cas9 System for Gene Therapy. *Nucleic Acids Res.* 48 (16), 8870–8882. doi:10.1093/nar/gkaa683

Zimmermann, M., and de Lange, T. (2014). 53BP1: Pro Choice in DNA Repair. *Trends Cel Biol.* 24 (2), 108–117. doi:10.1016/j.tcb.2013.09.003

Zucchelli, E., Pema, M., Stornaiuolo, A., Piovan, C., Scavullo, C., Giuliani, E., et al. (2017). Codon Optimization Leads to Functional Impairment of RD114-TR Envelope Glycoprotein. *Mol. Ther. - Methods Clin. Develop.* 4, 102–114. doi:10.1016/j.omtm.2017.01.002

Zuo, E., Cai, Y.-J., Li, K., Wei, Y., Wang, B.-A., Sun, Y., et al. (2017). One-step Generation of Complete Gene Knockout Mice and Monkeys by CRISPR/Cas9-mediated Gene Editing with Multiple sgRNAs. *Cel Res* 27 (7), 933–945. doi:10.1038/cr.2017.81

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Copyright © 2022 Feng, Wang, Li, Xie, Liu, Li, Li, Wang, Hu, Yang and Guo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GLOSSORY

CRISPR clustered regulatory interspaced short palindromic repeats
ZFNs zinc finger nucleases
TALENs transcription activator-like effector nucleases crRNA
PAM proto-spacer adjacent motifs
sgRNA single guide RNA
DSBs double-strand breaks
PCR Polymerase chain reaction
Acr anti-CRISPR
NHEJ nonhomologous end joining
c-NHEJ canonical NHEJ
HDR homology-directed repair
53BP1 tumor suppressor p53-binding protein 1
CtIP CtIP C-terminal-binding protein interacting protein
dn53BP1 dominant-negative 53BP1
ssODN single-stranded oligodeoxynucleotides
dsDNA double-stranded DNA
EVs extracellular vehicles
CPPs cell-penetrating peptides
AAV adeno-associated virus
LV lentivirus
AdV adenovirus
GO graphene oxide
TDNs tetrahedral DNA nanostructures
CAR chimeric-antigen receptor
LNP lipid nanoparticles
NIR near-infrared
HSF heat-shock factor
AuNR gold nanorods
AuNCs gold nanoclusters
CRISPRi CRISPR interference
CRISPRa CRISPR activation
eGFP enhanced green fluorescent proteins
RFP red fluorescent proteins
FPs fluorescent proteins
CBE cytosine base editors
ABE adenine base edito

frontiers in Genetics

REVIEW
published: 08 February 2022
doi: 10.3389/fgene.2022.831866

A Final Frontier in Environment-Genome Interactions? Integrated, Multi-Omic Approaches to Predictions of Non-Communicable Disease Risk

*Alexandra J. Noble[1], Rachel V. Purcell[2], Alex T. Adams[1], Ying K. Lam[1], Paulina M. Ring[3], Jessica R. Anderson[3] and Amy J. Osborne[3]**

[1]Translational Gastroenterology Unit, Nuffield Department of Experimental Medicine, University of Oxford, Oxford, United Kingdom, [2]Department of Surgery, University of Otago Christchurch, Christchurch, New Zealand, [3]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand

Epidemiological and associative research from humans and animals identifies correlations between the environment and health impacts. The environment—health inter-relationship is effected through an individual's underlying genetic variation and mediated by mechanisms that include the changes to gene regulation that are associated with the diversity of phenotypes we exhibit. However, the causal relationships have yet to be established, in part because the associations are reduced to individual interactions and the combinatorial effects are rarely studied. This problem is exacerbated by the fact that our genomes are highly dynamic; they integrate information across multiple levels (from linear sequence, to structural organisation, to temporal variation) each of which is open to and responds to environmental influence. To unravel the complexities of the genomic basis of human disease, and in particular non-communicable diseases that are also influenced by the environment (e.g., obesity, type II diabetes, cancer, multiple sclerosis, some neurodegenerative diseases, inflammatory bowel disease, rheumatoid arthritis) it is imperative that we fully integrate multiple layers of genomic data. Here we review current progress in integrated genomic data analysis, and discuss cases where data integration would lead to significant advances in our ability to predict how the environment may impact on our health. We also outline limitations which should form the basis of future research questions. In so doing, this review will lay the foundations for future research into the impact of the environment on our health.

Keywords: multi-omics, integrated analyses, non-communicable diseases (NCD), gene-environment inreractions, health outcomes, disease risk prediction

OPEN ACCESS

Edited by:
Mehdi Pirooznia, Johnson & Johnson, United States

Reviewed by:
Satish Sati, University of Pennsylvania, United States
Adam de Smith, University of Southern California, United States

****Correspondence:***
Amy J. Osborne
amy.osborne@canterbury.ac.nz

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *09 December 2021*
Accepted: *19 January 2022*
Published: *08 February 2022*

Citation:
Noble AJ, Purcell RV, Adams AT, Lam YK, Ring PM, Anderson JR and Osborne AJ (2022) A Final Frontier in Environment-Genome Interactions? Integrated, Multi-Omic Approaches to Predictions of Non-Communicable Disease Risk. Front. Genet. 13:831866. doi: 10.3389/fgene.2022.831866

INTRODUCTION

Genomics is revolutionising our understanding of the biological basis of disease, and it is undisputed that our individual genotypes, in combination with lifetime exposures (our environments), are critical determinants of non-communicable disease (NCD) risk. For example, it is well established that the prenatal and early life environments strongly influence the risk of non-communicable

disease in later life [the Developmental Origins of Health and Disease, DOHaD (Barouki et al., 2012)], This is exemplified in a recent study into the role of genetics, early life nutrition and their interaction on adult health, which demonstrated that high genetic risk for non-communicable disease can be mitigated by an environmental intervention [i.e., longer duration of breastfeeding (Wang et al., 2021)]. However, definitively connecting environmental exposures to specific health outcomes remains stubbornly challenging, and it is generally unclear whether the associations we see are causal, partly causal, or simply correlative. While techniques such as causal inference and Mendelian randomisation have improved our ability to determine putative causal relationships between risk factors and disease [e.g., (Timpson et al., 2005; Pingault et al., 2018)], there is currently no way to predict whether, or how, a particular environmental exposure may help or harm human health, in a way that gives us the ability to assign cause and effect. Thus, mitigating the impacts of the environment on our health largely remains an elusive goal.

One factor that contributes to our lack of power to detect cause and effect for environment-phenotype interactions, is that the impact of our environment on our genome is governed by much more than our genotypes (Manolio et al., 2009). For example, there are complex layers of information embedded in the structure of the genome and epigenome that impact gene expression. Due to the complex nature of the interactions between these information levels and the other nuclear and cellular processes that emerge within the complex system, we still do not fully understand what happens at the genomic level to translate environmental signals into phenotypes, and very rarely do we have the power to draw conclusions, because, while the genetic code is largely static, the multiple, dynamic, and interacting layers of genome structure and organisation are open to environmental influence. Thus, while it is clear that there is a distinct sequence of events that must happen in order to translate environment into phenotype, exploration of these events, and any attempt to predict the effects of the environment on our health, will require integration of information from across multiple biological and information levels. For example, quantifying DNA methylation in response to a particular environmental variable will only provide an indication that something biologically meaningful might be happening at a few loci in the genome—it may well mark biologically relevant pathways, but it cannot tell us the impact that methylation has on gene expression. Therefore, we struggle to predict the phenotypic impact of DNA methylation alone. Many complex biological questions, such as understanding the biological basis of environment-driven health inequalities, have not been addressable with one-dimensional reductionist approaches, and advances in our ability to predict the impact of our environment on our health therefore will require integration of multi-layered genomic data in a way that accounts for interactions between and across the biological layers.

Developments in 'omics techniques and technologies, and environmental electronic data (e.g., from wearables) means that we are at a point in our endeavours where we can explore an integrated, multi-omics approach to health and wellbeing. Here, we first describe the major and most well studied layers of genome regulation, and focus on the application of these to NCDs and complex disease, reviewing current efforts to integrate multi-omic data in disease. We describe new and emerging technologies that will improve our ability to assign a phenotypic impact to an environmental exposure. In doing so, we argue that progress in this field will be dependent on our ability to undertake integrated, multi-omic approaches that fully explore the environmental and molecular basis of complex disease.

How Our Genomes are Regulated—Potential Areas for Environmental Influence

Perhaps one of the most well-defined epigenetic signals for gene/genome regulation is DNA methylation. DNA methylation is a common form of epigenetic genome regulation, wherein methyl groups are added to the 5' positions of cytosines in cytosine-guanine dinucleotides (CpGs), which further correlates with histone modifications and chromatin accessibility. Importantly, patterns of DNA methylation can be altered by environmental exposures (Jaenisch and Bird, 2003) and we know that they can be influenced by early-life environment (*in utero* and early post-natal), which, itself, is associated with variation in later-life disease susceptibility (Gluckman et al., 2008; Barouki et al., 2012; Lillycrop and Burdge, 2012). However, while changes in DNA methylation are often identified in response to a changing environment (Feil and Fraga, 2012), methylation by itself does not explain the full complexity and diversity of the genomic response to the environment (Freeman et al., 2016). That is because DNA methylation is just one type of epigenetic signal that can work to regulate gene expression (Jaenisch and Bird, 2003; Jirtle and Skinner, 2007; Bonev and Cavalli, 2016). Other mechanisms include non-coding RNA (ncRNA) transcription (Jaenisch and Bird, 2003; Weber et al., 2007) and modification of histone protein tails within nucleosomes, both of which directly affect 3-dimensional (3D) genome organisation and, ultimately, nuclear functions (Risca and Greenleaf, 2015; Bonev and Cavalli, 2016).

The 3D organisation of the genome emerges from the sum of the functions that are occurring within the nucleus, and is widely considered to have a role in the regulation of gene expression (Cremer and Cremer, 2001; Lieberman-Aiden et al., 2009). For example, DNA looping brings distant gene enhancers and promoters together, which promotes the recruitment of RNA polymerase and ultimately gene transcription. Chromosomes are organised into highly conserved territories (Dixon et al., 2012; Sexton et al., 2012) and at a finer scale, precise domains, termed topologically associating domains (TAD). Genes located in the same domain are often co-expressed and are insulated from genes in neighbouring domains by domain boundaries (Nora et al., 2012). Perturbations of domain boundaries can disrupt both short- and long-range genomic interactions, sometimes with pathological outcomes (Franke et al., 2016). 3D genome organisation and chromatin accessibility can be studied using techniques such as ATAC-seq [Assay for Transposase-Accessible Chromatin with high-throughput sequencing (Buenrostro et al.,

2015)] and proximity ligation experiments such as Hi-C (Bickmore and van Steensel, 2013; Stevens et al., 2017). Understanding 3D genome structure is important because chromatin remodelling is a dynamic and often adaptive response to the environment (De Nadal et al., 2011; Matilainen et al., 2017). For example, exposure to inhaled industrial chemicals (Fang et al., 2014) or heat stress (Petesch and Lis, 2008) results in alterations to chromatin structure, changing chromatin accessibility, with associated downstream effects on gene expression. 3D genome organisation has recently been implicated in the pathogenesis of obesity and diabetes (Fadason et al., 2017), highlighting the importance of integrating spatial information into interrogations of the genetic basis of complex disease.

Non-coding RNAs (ncRNAs) are transcribed from DNA but not translated into protein. Despite this, ncRNAs have broad roles in genomic regulation. For example, microRNAs (miRNAs) guide argonaute proteins to degrade mRNAs containing sequence targeted by the seed region of the miRNA, culminating in transcriptional silencing (Peters and Meister, 2007). ncRNA transcription can be altered by environmental factors (Saxena and Carninci, 2011; Tani et al., 2014) to directly influence gene expression patterns (Guttman and Rinn, 2012; Engreitz et al., 2016). Further, DNA methylation can influence ncRNA transcription to produce health-related phenotypic effects (Lujambio et al., 2010), suggesting the two processes can work together. Importantly, ncRNA interacts with chromatin and can alter the accessibility of genomic regions for transcription (Castel and Martienssen, 2013), and can remodel 3D genome structure (Cubeñas-Potts and Corces, 2015; Dekker and Misteli, 2015; Engreitz et al., 2016; Rowley and Corces, 2016).

Histone protein tails can be modified by post-translational modifications that include the addition of either acetyl groups or methyl groups (Bannister and Kouzarides, 2011), which can affect chromatin accessibility and alter transcription profiles. Patterns of histone modification can be explored by techniques such as mass spectrometry (MS) and chromatin immunoprecipitation and sequencing (ChIP-seq) (Esteller, 2007). Histone modifications can interact with DNA methylation, and this interaction has been associated with disease phenotypes [e.g., cancer (Vaissière et al., 2008)], as both are sensitive to the environment (Jirtle and Skinner, 2007; Dai and Wang, 2014). Therefore, histone modifications, and chromatin accessibly, are strong determinants of gene expression profiles.

In addition to the more traditionally recognised records of environmental impact on the genome, there are other sources of information that reflect and respond to the interactions between the host genome and environment. For example, the microbiome, proteome and metabolome each emerge from the complex web of environment, genetic, structural and epigenomic interactions. It is clear that perturbations of these systems can indicate an effect on health, which can be interrogated under an 'omics platform, and integrated into subsequent analyses:

The human microbiome is generally defined as the 'complete set of genes and genomes of the microbiota (bacteria, archaea, eukaryotes, and viruses) that reside in and on a person'. More extensive definitions also include aspects of the surrounding environmental conditions in their definition (Marchesi and Ravel, 2015). Microbiomes can be analysed at different levels, be that their metagenome (DNA) to assess community composition and functional capacity, or at the metatranscriptome (RNA) level; at this level, RNA is used to define community composition as well as characterise the activity of the organisms at the time of sampling. The microbiome, and by extension, the metagenome and metatranscriptome, is variable, depending on many environmental factors, such as anti- and probiotic use, age, diet, environment and physical activity levels. Despite few causal examples, it is widely recognized that changes in the gut microbiota are associated with the onset and progression of non-communicable diseases [reviewed in (Noce et al., 2019)], including autoimmune diseases such as multiple sclerosis (MS) and rheumatoid arthritis [RA, (Tsai et al., 2021)]. Comprehensive investigations of the microbiome are, by their nature, integrative requiring analyses of the metagenome and metatranscriptome; direct and indirect interactions between the microbiome and the host, and environment-microbe-metabolism interactions (Kurilshikov et al., 2019).

The proteome is the complete set of proteins expressed by an organism, tissue or cell at a particular point in time. Naturally, the proteome shares components of its dynamism with the transcriptome and the epigenome as a result of the process of gene expression. However, the proteome is widely recognized as not having a 1:1 relationship with the transcriptome, due to factors affecting translation and post-translational modifications [e.g., (Ghazalpour et al., 2011; Wang et al., 2019)]. Therefore, a complete picture of cellular activity cannot be determined from the transcriptome alone. Recent studies connect the proteome to immune dysregulation and obesity (Garrison et al., 2017) and traits relevant to the DOHaD hypothesis (Sarli et al., 2021). Moreover, the proteome is known to respond to environmental stimuli (Koga et al., 2011; Calamini and Morimoto, 2012) including diet (Vileigas et al., 2019), chemical exposure (Lee et al., 2018) and smoking (Sinha et al., 2021). Therefore, given the often imperfect correlation between the transcriptome and the proteome, proteomic data is one layer of omic information that adds value to integrated, multi-omic approaches; it allows for the refinement of the number of target genes deemed necessary to investigate further, since gene expression changes that are not correlated with a coordinated change in protein expression can be discounted from downstream analyses.

The metabolome comprises all current low molecular weight cellular metabolites, indicating current cellular activity levels. The metabolome essentially denotes the end product of cellular processes, allowing a functional readout of an organism (Wang et al., 2020). Altered metabolomes have been identified in many NCDs, including Type II diabetes and obesity (Fiehn et al., 2010; Zhang et al., 2014; Merino et al., 2018). Furthermore, hormones and other metabolites can be programmed *in utero* through epigenetic mechanisms (Rauschert et al., 2017), such that a child's metabolism can be influenced by its environmental

exposures (Cottrell and Ozanne, 2008). This means that the metabolome may contribute to our understanding of the developmental basis of disease by refining our ability to assign functional changes to environmental exposures. Recent studies also demonstrate the value of integrating metabolome with microbiome data to profile disease pathogenesis, for example, a recent review of autoimmune disease describes how the microbiome and associated metabolic profile are altered by 'modern' lifestyles, which is impacting on inflammatory responses (Tsai et al., 2021). Given that the metabolome itself is the end product of cellular processes, it stands to reason that, like the proteome, it can be indirectly altered by environmentally-induced genomic, epigenetic, structural and microbiotic changes. This means that integration of the microbiome in multi-omics studies will provide indications of the functional significance of observed genomic and epigenetic changes, which may highlight mechanistic pathways that are important for the aetiology of disease.

Lastly, it is important to also consider the implication of an individual's underlying genetic variation, and its interaction with environmental factors, when assessing the impact of the environment on genome regulation. For instance, we know that GWAS explain only a portion of the heritability of NCDs such as obesity and MS (Silventoinen et al., 2016; International Multiple Sclerosis Genetics Consortium, 2019) and that GWAS cannot explain all the variability of traits; many causative loci exist in intergenic regions of the genome (Manolio et al., 2009), and further, disease heritability has been observed to interact with an individual's environment [e.g., (Hüls et al., 2021; Jacobs et al., 2021; Ye et al., 2021)]. Therefore, since underlying genetic variation may influence, e.g., methylation patterns if that variation is at a modifiable cytosine residue, genetic variation cannot be discounted when attempting to predict the phenotypic effects of our environments.

How do the layers of complexity interact to influence phenotypes? Reductionist approaches that do not integrate these different levels of information may miss many of the crucial interactions that determine how our genomes orchestrate a biological response to our environments. In so doing, we will lessen our ability to investigate the effects of our environmental exposures and lose our power to predict how they might be influencing our health.

How Research is Exploring Integrated Approaches to Understanding the Impact of the Environment on Disease

Non-communicable diseases are diseases that are non-infectious in nature, but nevertheless cause severe and debilitating disease, and are a major public health burden and cause of morbidity and mortality (W.H. Organization, 2019). Perhaps due to our 'transition to modernity' (Corbett et al., 2018) such diseases are increasing in prevalence globally, making them the focus of intense research. Here, we focus on examples of the application of integrated, multi-omic approaches to several NCDs that are all themselves a product of the interaction between environmental exposures and genetic predisposition.

Obesity

Obesity is by far the most prevalent NCD for which data on integrated, multi-omic approaches exist. This is because obesity is a major public health burden, increasing in prevalence (Abarca-Gómez et al., 2017) and is a risk factor for many other metabolic diseases such as type II diabetes, cardiovascular disease, and some cancers (Johnson et al., 2015; Weihrauch-Blüher et al., 2019). Obesity is driven by a combination of an underlying genetic predisposition, and environmental factors (Albuquerque et al., 2017), including *in utero* exposures (Tounian, 2011). This means that unpacking underlying genetics, maternal and individual environmental effects to determine which environmental impacts are causative, versus those that are correlational, is difficult.

A handful of studies demonstrate an integrated approach, not necessarily on the impact of the environment, but rather, exploring multiple layers of genomic data to detect genomic changes relevant to a phenotype. For example, a recent study by van der Kolk et al. (van der Kolk et al., 2021) investigated the link between obesity and metabolic complications through the application of RNA sequencing, proteomics and metabolomics; their study cohort contained 49 BMI-discordant monozygotic twin pairs, meaning their shared genetic background enabled the researchers to build a metabolic and genomic profile of acquired (environmentally-dependent) obesity. The authors detected a downregulation of mitochondrial pathways and an upregulation of inflammatory pathways, along with alterations to the metabolome that were specific to acquired obesity. However, while this is a strong example of investigations of multiple types of genomic data, these data are not strictly integrated in their analyses. Rather, Kolk et al. present these data side-by-side as independent profiles in a manner that reinforces the biological interpretations without achieving true integration as a means of tracing cause and effect.

Integration, has been attempted in other obesity studies. For example, Chen et al. (Chen et al., 2021), citing a recent epigenome-wide association study that linked individual CpG sites with obesity traits (Sayols-Baixeras et al., 2017), explored the correlation between DNA methylation and gene expression. Their study reported associations between genes that were differentially expressed and differentially methylated. They also identified two novel genes, S100A8 and S100A9, expression of which correlated negatively with methylation and were associated with increased obesity. This study highlights the strength of integrating DNA methylation and gene expression data to deepen our understanding of the relationships between DNA methylation and gene expression in complex phenotypes.

Many genomic analyses are applied to human studies retrospectively as part of post-hoc analyses, and many are also limited in their scope, in terms of type of data available, tissue of origin, and cohort size. Unsurprisingly, then, we can gain more insight into integrated multi-omic approaches using models of human disease. For example, Joslin et al. (Joslin et al., 2021) recently attempted to functionally interpret genome-wide association study data in obesity, by capturing information on chromatin accessibility, gene expression, and long-range enhancer-promoter interactions, in human-induced pluripotent

stem cell (iPSC)-derived hypothalamic neurons. Their data indicated that the genetic architecture at GWAS loci is complex, but nevertheless they were able to detect putative enhancer-modulating variants that have regulatory properties in their cell line, at obesity-related loci. The strength of their highly integrated approach, and the increase in scope offered by using a cell line (therefore only a 'single' genome) has allowed them to develop a pipeline to prioritize GWAS target genes for functional follow up, potentially limiting the number of functional loci that further studies may need to investigate in human-based cohorts. Joslin et al.'s study highlights the potential of integrating multiple layers of genomic complexity. The expansion of their pipeline to include, e.g., DNA methylation data, along with environmental variables that drive epigenetic variation between individuals, could drive further discoveries in this area, by facilitating a clearer understanding of how epigenetic mechanisms contribute to the association between SNPs and enhancer-promoter interactions.

These examples (Chen et al., 2021; Joslin et al., 2021) demonstrate a role for integrated genomic analyses in the relationship between the environment and a phenotype. As more data is generated, and as techniques improve, finding a way to integrate environmental variables into models of integrated multi-omic approaches to obesity will be a key driver of our ability to assign causation to both environmental factors and genetic and epigenetic mechanisms in the development of obesity. This is because an individual's specific environment can tell us something meaningful about the exposures that may be driving differences at the cellular level, that may be impacting on, e.g., gene expression, the microbiome, and the proteome. Current computational capabilities and research methodologies suggest that the further integration of 3D genome structure, to aid in the linking of risk variants to target genes (Krijger and De Laat, 2016), and the unequivocal role of the microbiome in obesity, (Hartstra et al., 2015; Jayasinghe et al., 2016; Maruvada et al., 2017), is a natural focal point for future research in this disease.

Type II Diabetes

Type II diabetes (T2D) is characterised by a resistance to insulin, meaning that blood glucose levels in the body are not able to be controlled properly, often leading to hyperglycemia, obesity, hypertension and hyperlipidemia, and eventually severe complications such as blindness and kidney failure (W.H. Organization, 2016; Khawandanah, 2019). T2D is becoming a global epidemic (Kaiser et al., 2018), thus, understanding environmental drivers of T2D, and how they may interact with an individual's underlying genetics to cause disease, will be fundamental to a global approach to mitigate its rise in prevalence.

In a meta-analysis of diabetes GWAS, Schierding et al. (Schierding and O'Sullivan, 2015) integrated SNP data, 3D genome and eQTL data, to identify 'spatial hubs', or connections between loci in genes that contribute to disease. Additionally, Xue et al. (Xue et al., 2018) combined data from gene expression studies of human blood with GWAS and identified a suite of putative functional genes for T2D, linking GWAS data with a potential functional (gene expression) output. Further, Xue's research integrated of DNA methylation data with epigenome annotation data and identified three genes (*CAMK1D, TP53INP1, ATP5G1*) as having a plausible regulatory mechanism in T2D. In the context of this review, these findings are instrumental for their ability to refine large (e.g., GWAS) datasets and improve their predictive power, by associating disease-associated SNPs with downstream and integrated layers of genomic regulation (Schierding and O'Sullivan, 2015), and by the integration of SNPs with gene expression and epigenome annotation data (Xue et al., 2018). These approaches could be readily applied to other NCDs, with environmental covariables included where studies allow, for example, integrating lifestyle and family data; accounting for heritable genetic variation and lifestyle risk factors will to strengthen the ability to assign causation to a particular risk factor (environment/lifestyle or genetic).

There are multiple examples of GWAS to determine susceptibility loci for T2D. However, the large majority of the loci identified fall in non-coding regions of the genome, meaning that it can be highly challenging to determine which genes and transcripts their variation is relevant to, and which molecular pathway they may influence. Integrated multi-omic approaches are valuable to attempt to predict which disease-associated loci are functionally relevant, in the context of the phenotype of interest. This is important when considering environmental drivers of complex disease, because if individual variation at a particular locus is associated with an environmentally-influenced disease, determination of the functional impact of that locus may help us predict whether that locus may be causative for disease, or simply correlated. For instance, that locus may mark an underlying CpG site, or be located within a ncRNA sequence, which we know are sensitive to environmental influence, and therefore may influence the expression of genes that may be phenotypically relevant. To this end, efforts have been made to develop analytical pipelines that integrate genetic, genomic and biological data to produce networks that indicate connectivity between GWAS loci and candidate causal genes. For example, Nyaga et al. (Nyaga et al., 2021) used integrated genomics to ask whether there were any shared genetic features of type I diabetes (T1D) and T2D; their functional approach integrated Hi-C and eQTL data to characterise the functional impacts of disease-associated SNPs, identifying genetic regulatory regions that alter regulation of genes common to both T1D and T2D, that are associated with disease development. Additionally, Fernández-Tajes et al. (Fernández-Tajes et al., 2019) present an analytical pipeline to define the transcriptional activity of T2D-associated SNPs, integrating genomic data to reveal connectivity between candidate genes at T2D GWAS loci. These approaches, while distinct, can be applied to other diseases, using other types of genomic data, thereby providing insights into the diseases that are identifying new means of stratification, prevention and treatment, which collectively prove the importance of these types of approaches.

Mens and colleagues (Mens et al., 2020) used large-scale GWAS data to detect variants associated with T2D traits, and integration of DNA methylation and miRNA expression data

confirmed that several of these miRNAs were associated with T2D traits. The data used by Mens et al. was obtained from human peripheral blood, therefore their identified miRNAs could be considered as biomarkers for T2D. Their study highlights yet another strength of integrated analyses; the computational reduction of a huge study into practical targets by assigning a more likely function to those targets, prioritising areas for follow up.

Multiple Sclerosis

Multiple sclerosis (MS) is an autoimmune disorder that affects the central nervous system (CNS). It leads to the destruction of myelin in the CNS, blocking the transport of signals along the nerves to the brain, meaning that movement, sensation and body functions are impacted. MS is characterised by periods of recovery and relapse over a number of years, before ending in disability, with no treatment available at the progressive disease stage.

The International Multiple Sclerosis Genetics Consortium, and others, have published numerous GWAS to identify genetic factors that contribute to MS [e.g., (Consortium et al., 2007; International Multiple Sclerosis Genetics Consortium, 2019; Mo et al., 2019)], however, there are also multiple known environmental risk factors for MS development, most prominently, lack of vitamin D exposure, smoking, and exposure to Epstein-Barr virus (EBV) (O'Gorman et al., 2012). For example, a recent longitudinal study in military veterans demonstrated a strong link between EBV and MS (Bjornevik et al., 2022). Despite these known associations, and the complex interplay between genetics and the environment, most studies of MS focus only on GWAS, or are conducted at the candidate-gene level, for example, correlations between promoter methylation and gene expression levels (Hosseini et al., 2020). A small number of studies have started to integrate genomic information across multiple genome technologies and layers of regulation. For example, Gokuladhas and colleagues (Gokuladhas et al., 2020) integrated SNPs from (amongst other neuromuscular disorders) MS patients with Hi-C and eQTL data, to identify target genes to prioritise for therapy and treatment of MS. Their approach, essentially determining SNP-mediated gene regulation, highlights the potential for the integration of SNP and spatial data for more precisely identifying the molecular mechanisms of complex disease, as well as providing evidence of disease-related SNP functionality, particularly given that most SNPs are in intergenic regions of the genome. Gokuladhas et al. have since expanded the approach to include protein-protein interaction data, and applied it more widely to autoimmune diseases to reveal shared biological processes across autoimmune diseases (Gokuladhas et al., 2021).

Mo et al. (Mo et al., 2019) employed Mendelian randomisation to explore GWAS, gene expression (eQTL) and epigenome-wide association study (mQTL) data, to determine whether e- and mQTL data, in combination with GWAS, was a viable way to prioritise relevant GWAS loci for further investigation. While not integrated in the strict sense (the authors explored overlap and validation in the individual datasets) this technique was highly successful in identifying potentially causal SNPs and DNA methylation differences, demonstrating the strength of this methodology to identify genomic features that may participate in the pathogenesis of MS.

Rather than interrogating genomic loci such as SNPs, Cervantes-Gracia and others (Cervantes-Gracia and Husi, 2018) used publicly available expression datasets to identify the most common molecules relevant to MS. Their approach was to generate interaction networks to identify molecular pathways/conserved networks that are deregulated across MS. They integrated mRNA and miRNA expression profile datasets, and impressively, combined these with differentially expressed genes identified through studies of, e.g., EBV infection, allergies and other autoimmune diseases. Their research uncovered a suite of molecules (mRNAs, miRNAs) that were correlated and deregulated in their datasets, that they could use to infer novel findings about the primary cause of the molecular changes seen in MS blood samples.

Based on evidence gathered from research into other NCDs, namely, that an integrated, multi-omic approach is valuable and insightful, together with the paucity of such approaches being applied to MS, highlights how much MS research will benefit from the integration of multiple layers of genomic data, particularly in light of the strong and well-identified environmental factors [e.g., (Bjornevik et al., 2022)]; this approach will allow us to interrogate the impact of the environment and the genome on MS progression, providing novel insights into the biological basis of disease development and progression.

Alzheimer's Disease

Alzheimer's disease (AD) is the most common type of dementia, a form of brain degeneration that ablates memory and cognitive functions. AD is increasing in prevalence, most likely due to longer life expectancies globally, and it is estimated that 100 million people will have AD or dementia by the year 2050 (Palmqvist et al., 2020). There is currently no treatment to prevent the progression of dementia (Mehta et al., 2017), although some drugs can help to manage symptoms, if detected early. However, because AD is usually only diagnosed at an advanced age, early diagnosis and rapid treatment is challenging.

Alzheimer's disease has been the focus of multiple GWAS over recent years, with over 20 independent loci associated with the disease (Van Cauwenberghe et al., 2016). Further, there are known environmental risk factors that associate with a diagnosis of AD, such as obesity, hypertension and tobacco smoking (Østergaard et al., 2015). Because AD is currently incurable, understanding the environmental and genetic determinants of AD is paramount if we wish to be able to both prolong life *via* early diagnosis, and develop effective and additional therapies, and integrated, multi-omic studies are the clear pathway to achieving this. Currently, genomics, transcriptomics, proteomics and metabolomics are offering a more comprehensive view of molecular pathways underlying the development of neurodegenerative diseases. For example, they are helping to differentiate subtypes of patients based on their specific molecular signatures, to aid individual treatment

plans for patients (La Cognata et al., 2021). Additionally, genomic technologies are profiling the transcriptome of the brain with neurodegenerative diseases (Neff et al., 2021), and while studies explore the 'omics' of AD, few have done so in an integrated manner, to improve the power of their associations.

Thus, as with many complex diseases, truly integrated, multi-omic studies are scant. However, a recent comprehensive study by Nativio and colleagues (Nativio et al., 2020) used transcriptome profiling of human brain samples to inform proteomic analysis and ChIP-seq, followed by an exploration of the overlap of their identified genes, with GWAS and eQTL data. This powerful study identified upregulation of transcription- and chromatin-related genes (including the histone acetyltransferase genes for H3K27ac and H3K9ac) in AD, culminating in the new knowledge that the histone modifications H3K27ac and H3K9ac and genome reconfiguration are potentially important AD. Further, multi-omic atlases of AD from human brain tissue are currently being constructed (De Jager et al., 2018), that include genotypes, whole genome sequencing, DNA methylation, chromatin immunoprecipitation, RNA and miRNA profiles, with the focus of understanding the molecular mechanisms of AD in the target organ, rather than a cell line or animal model. Nativio's study suggests that we can use integrated data to explore genomic mechanisms associated with AD, and genome atlases will allow us to integrate data across multiple levels. This is important in an uncurable disease such as AD, where the identification of targets and the development of new therapies is imperative.

Rheumatoid Arthritis

Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disease that affects joints. It is characterised by increased inflammation in the synovial membrane, which causes swelling and damages the joint *via* bone erosion. Like other autoimmune diseases, RA is characterised by periods of flare/exacerbation and of remission, and, again like other autoimmune diseases, its aetiology is a complex mix of underlying genetic risk factors (e.g., particular HLA class II genotypes (Raychaudhuri et al., 2012) and environmental risk factors [age, *in utero* exposure to tobacco, personal tobacco use, obesity, and a high sodium diet (Deane et al., 2017)]. There is no cure for RA, but joint destruction can be delayed by prompt and aggressive anti-inflammatory treatment.

Thus, because of the known genetic determinants of RA that lead to increased susceptibility, which is further enhanced by the multiple environmental risk factors, that are themselves all able to modify the epigenome [e.g., (Joubert et al., 2012; Besingi and Johansson, 2014; Florath et al., 2014; Jaffe and Irizarry, 2014; Ivorra et al., 2015; Sayols-Baixeras et al., 2017; Noble et al., 2021)], researchers are applying multi-omics approaches to identify networks that drive disease progression, and to prioritise candidate genes for study, all of which may aid in the identification of targets for novel therapeutics. For example, Whitaker et al. (Whitaker et al., 2015) used an unbiased approach to integration (i.e., they did not assume a relationship between DNA methylation and gene expression) to prioritise candidate genes; they devised a strategy to identify 'multi-evidence genes' (MEGs) to identify triple-evidence genes that overlap between epigenome, transcriptome and sequence data, to collate sets of genes that were implicated in RA. Their approach identified seven triple-evidence genes, validating some as candidates for new RA therapies. Further, an assessment of RA pathogenesis was undertaken *via* an integrated DNA methylation and gene expression approach by Li Yim and colleagues (Li Yim et al., 2021); they identified a suite of 59 genes with coordinated changes at the gene transcript and DNA methylation level, which were associated with immune response pathways. Their research provided more evidence for molecular changes associated with RA pathogenesis, and their approach, like that of Whitaker et al., will be useful in aiding in the prioritisation of targets for new therapeutics, *via* the identification of potential new drug targets. Another benefit of a multi-omic approach is that it provides the power to interrogate disease-relevant tissue in a dynamic way, allowing a fuller understanding of the variants that shape disease. A relevant example of this is that of Ha et al. (Ha et al., 2021), who explore GWAS, gene expression and DNA methylation in CD4+ T cells in patients with RA; CD4+ T cells are the most disease-relevant tissue in RA. Their research identified a larger number (2575) of RA-specific differentially expressed genes that correlated with RA-specific differentially methylated regions of the genome, and that were enriched in T cell differentiation and activation pathways. They were also able to show, through their multidimensional approach, that many of the differentially expressed genes were explained by eQTLS (771, for transcripts) and mQTLs (83, for differentially methylated regions). This comprehensive study clearly demonstrates that integrating SNP, gene expression and DNA methylation data can aid in the dissection of genome regulation in a complex disease state, and Han's methodology has the potential to be applied readily to other complex diseases.

Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) includes Ulcerative colitis (UC) and Crohn's disease (CD), both of which are characterised by chronic inflammation in the gastrointestinal tract, are debilitating, and can lead to severe and life-threatening complications. Like MS and RA, IBD patients experience remission and relapse of symptoms. IBD is thought to arise from activation of the intestinal mucosa immune system, and the disease has been subject to extensive genetic and epigenetic examination in patients. This is largely driven *via* the collaborative effort from groups in the International IBD Genetics Consortium (Ventham et al., 2013) and currently, over 200 susceptibility loci have been identified as playing a role in IBD (Liu et al., 2015), with methylation sites in genes linked to inflammation detected in whole blood of IBD patients (Adams et al., 2014; Somineni et al., 2019). In addition to genetic risk loci, several environmental factors have been associated with the development of IBD, for example, geographic location, cigarette smoking, diet and gastrointestinal infection (Baumgart and Carding, 2007). Thus, given the complexity of IBD, an approach that integrates multi-omic data, including that of the microbiome, metabolome and proteome, will enable the identification of genomic loci that are more likely to mediate

disease risk, and those which may be modified or influenced by the environment. Work in this area has begun, with a study by Ventham et al. (Ventham et al., 2016) which related genomic and gene expression data to cell methylation profile; Ventham's study identified five differentially methylation regions and 439 differentially methylated positions that were IBD-specific, and further, by using paired genetic and epigenetic data, showed how site-specific DNA methylation changes correlate with underlying genotype differences. Therefore, since their methodology enables the relation of site-specific DNA methylation changes to underlying genotype, it provides a platform for future studies in this area, *via* the ability of this pipeline to identify biomarkers that can be used for early diagnosis and treatment.

As well as biomarkers for early diagnosis, multi-omic approaches have been used to identify patients at risk from IBD relapse. IBD is an inflammatory disease which is located in the gut, therefore many studies of the disease focus on non-genetic omics, as exemplified in a study by Borren et al. (Borren et al., 2020), who employed proteomic and metabolomic profiling of patient blood samples, with the addition of fecal metagenome sampling, to determine whether they could identify biomarkers that suggest relapse. Their data was used to generate a combined risk score of relapse from the proteomic and metabolomic profile, which was correlated with fecal microbiome composition, to indicate a correlation between particular gut microbes and risk scores, which was predictive of future risk of relapse in patients. The strength of this method is in the integration of 'omic data from a non-invasive source (blood samples), leading to the development of a risk profile that will go some way to improving the ability to predict relapse, helping patients with the uncertainty of the future path of their condition.

The gut microbiome experiences alterations during periods of active IBD, termed functional dysbiosis, and work by Lloyd-Price and colleagues (Lloyd-Price et al., 2019) set out to understand these changes comprehensively, at the system level. As part of the Integrative Human Microbiome Project, they characterised the gut microbial ecosystem gathering multiple microbial profiles (metagenome, metatranscriptome, proteome, metabolome, and virome). Their research demonstrated characteristic changes in microbe composition, and changes to microbe transcription and metabolite pools, that were disease-specific, indicating that their integrated approach had identified relationships between multi-omic features that were potentially central to periods of IBD activity.

Can Integrated Genomics Shed Light on the Aetiology of Multimorbidities?

Two or more NCDs often co-exist in the same individual (multimorbidity), which can further complicate the dissecting of the gene-environment interactions that are important for disease progression. Coupled with the fact that many disease-associated SNPs are in non-coding regions of the genome (Hindorff et al., 2009), and that gene regulatory elements can strongly impact distant genes strongly, limiting our capacity to make assumptions of regulatory function based on gene proximity alone (Schierding et al., 2016), a non-integrated approach that does not take into account the spatial landscape of the genome, and the interactions therein, will limit our ability to understand the aetiology of multimorbid traits. Research is currently underway to address this question, challenging our understanding of the functional impact of SNPs; specifically, Fadason et al. (Fadason et al., 2018) integrated spatial data across multiple human traits from the GWAS catalogue, identifying eQTL-eGene pairs (phenotype-associated SNP-gene pairs with confirmatory interaction data), that were missed by proximity GWAS association, as well as inter-chromosomal eQTL associations. The highlight of this approach was the ability of the research methodology to identify phenotype-associated genetic components relative to multimorbidity and individual disorders, demonstrating the strength of this approach in understanding the aetiology of complex disease, and providing a platform for the future integration of other multi-omic and environmental data.

Current Computational and Bioinformatic Challenges/Limitations

Exploration of the highly interactive and dynamic layers of genome regulation in an integrated and informed way will enable us to begin to understand the biological basis of human diseases that are not 'Mendelian' (one gene, one disease) in nature. By doing so, we can improve our power to understand more about the aetiology of diseases that have a large environmental component. However, there are series of specific limitations that need to be addressed before progress can be made in this area. For example, the inclusion of additional omic data types in a study will increase the cost of that study, meaning that if budgets remain the same, sample sizes will be smaller and our power to detect associations will be reduced. Moreover, epistatic interactions are frequently ignored in analyses of genetics, and the sample sizes required to properly assess epistatic interactions are larger than those required to test single measurements against some outcome variable. This means that the addition of multiple omics further increases the complexity.

Complexity is compounded by the need for statistical power—single data sets are becoming larger as more simultaneous measurements are performed (e.g., arrays, transcriptomes, single-cell omics) and correction for multiple testing already makes discovery more difficult - combining omics increases the number of tested relationships exponentially, reducing power and increasing computational load. The difficulty is not just a matter of the availability of sufficient computational resources; development of new statistical techniques is required, and datasets must be of a sufficient size to discover underlying effects.

Thus, producing sufficiently sized datasets for multi-omic analysis will require a combination of multiple analysis runs, and either combining cohorts from multiple centres, or meta-analysis of previously produced datasets. This necessity means that dealing effectively with batch effects is highly important (Leek et al., 2010; Lazar et al., 2013; Price and Robinson, 2018). Another complication is capturing each multi-omic variable

which is on its own trajectory. For instance, the timing of the process for addition of a methyl group to a cytosine will differ from the timeframe over which a difference in mRNA or protein levels are detected, or when proteins are activated by post-translational modification and localised to a particular cellular subcompartment, or the timeframe over which biological effect is observed. Analysing these events can lead to very different results purely based upon when samples are collected.

Another paradox is that DNA methylation, predominantly seen in promoter regions, is known to negatively correlate with gene expression *via* silencing of genes (Herman and Baylin, 2003). However, methylation present in the gene body is less well characterised for its involvement in gene expression (Jones, 1999). Thus, understanding this process is dependent on where events are taking place within the genome.

A major limitation currently in our investigations of the etiology of complex traits is the well-characterised bias in genomic data in public data repositories. A large proportion of genomic data is derived from populations of European ancestry, which is a major limitation given the known differences in genomic architecture between populations. This means that calculated effect sizes and risk scores based on underlying genetic variation cannot be assumed to be relevant to a global population (Martin et al., 2017; Duncan et al., 2019), with recent research clearly demonstrating the value of including diverse populations in the discovery and replication phases of GWAS, increasing the powers of discovery (Wojcik et al., 2019). The generation of diverse, globally-representative datasets for multi-omic studies is therefore a current limitation that will need to be addressed to demonstrate the applicability of techniques and research findings.

Future Technologies

As well as improving our bioinformatics capacities to fully integrate data and environmental variables as far as possible, future strength in this area will be driven by new and emerging genome sequencing technologies, which are not yet represented in the research examples reviewed here. For example, Oxford Nanopore Technology (ONT) sequencing devices utilise nanopore channels through which DNA strands pass, each nucleotide base causing a different ionic current, which can be called. Nucleotide base calling allows for the differentiation between all five different cytosine residues (cytosine C, methylcytosine 5-mC, 5-hydroxymethlcytosine 5-hmC, formylcytosine 5-fC and 5-carboxylcytoine 5-caC) (Laszlo et al., 2013; Jain et al., 2016; Rand et al., 2017), potentially serving as a powerful tool for the integration of these data, but which is currently technically and computationally challenging on a platform that sequences all of these factors individually. A limiting factor of ONT is the currently high error rate, which has discouraged some researcher. This, however, is continuing to be addressed with the introduction of new chemistry and further optimised computational corrections (Senol Cali et al., 2019).

Rapid progress has been made in the development of single-cell omics, the profiling of single cells from a heterogenous cell population. We now have the capability to profile the genome, epigenome, transcriptome, and proteome at the single-cell level, unlike bulk sequencing, which provides comprehensive data as a single population of cells. For instance: single-cell RNA sequencing is widely applied to profile transcriptome-wide gene expression in individual cells (Tang et al., 2009); single-cell epigenomic technologies, such as chromatin immunoprecipitation sequencing (Rotem et al., 2015) and assays for transposase-accessible chromatin using sequencing (Cusanovich et al., 2015) are used to define epigenetic states of individual cells, and; several advanced approaches, including cellular indexing of transcriptomes and epitopes by sequencing (Stoeckius et al., 2017) and RNA expression and protein sequencing assay (Peterson et al., 2017) allow the simultaneous investigation of both gene and protein expression. Thus, single-cell omics is a powerful tool for elucidating cellular and microenvironmental heterogeneity in order to characterise rare cell types and explore genomic, epigenetic, transcriptomic, and proteomic regulatory mechanisms at the cellular level.

The fundamental features of single-cell omics technologies are isolating, barcoding, and sequencing individual cells to determine their DNA, mRNA, or proteins which can all be carried out in parallel. These integrative analyses allow molecules to be deciphered for genotypic and phenotypic characteristics of individual cells, and their underlying regulatory mechanisms (Chappell et al., 2018). The first and most critical step in performing single-cell omics is isolating viable single cells from a population of interest (Wang and Navin, 2015), followed by the challenge of identifying sequences from the same cells (Klein et al., 2015). Despite the technical challenges, it is now possible to define a landscape of intercellular heterogeneity and functions associated with pathophysiological processes (Lee et al., 2020). The advancement of single-cell omics and integrative analysis of the genome, epigenome, transcriptome, and proteome at the single-cell level will undoubtedly enable unprecedented levels of precision and resolution in our understanding of complex cellular systems, while also providing an unprecedented opportunity to uncover novel biological processes. Furthermore, single-cell technologies have been adapted to studies of the spatial and higher-order chromatin structure of the genome, for example, ligation- and non-ligation-based sequencing technologies [reviewed in (Zhou et al., 2021)], as well as large-scale single-cell proteomic studies [e.g., (Specht et al., 2021; Slavov, 2020)], both of which indicate the depth to which the impact of the environment on genome regulation can now be probed in a cell-specific manner. Of importance to our ability to undertake advanced multi-omic studies in single cells will be in the analysis of single-cell data, particularly as such analyses will require the generation and analysis of complex networks. Progress in this area will likely require Deep Learning and Machine Learning approaches, which are currently being developed (Ji et al., 2021) and are highlighting the use of such approaches for identifying novel biological features that has not been possible up to now. Single-cell omics will therefore be a fundamental tool in studies of the impact of the environment on genome regulation; genomic features that are modified by the environment do so in a cell- and tissue-dependent manner, meaning our ability to determine the mode of action of particular environmental variables, with respect to disease, will be greatly enhanced.

Lastly, a major consideration for our future ability to assign phenotypic impacts to environmental factors will be our power to integrate an individual's underlying genetic propensity with the environmental risks associated with disease traits. Recent advances in the development of polygenic risk scores (PRS) for multifactorial diseases with a large degree of heritability and genetic determinants, such as most NCDs, are enabling new developments that have not been possible *via* traditional GWAS. For instance, a recent study by Hüls et al. (Hüls et al., 2021) demonstrated associations between high PRS for obesity, and sociodemographic and lifestyle factors in obese children; these associations were undetectable *via* traditional GWAS (due to the lack of power associated with large cohorts and individual loci). Further, a study of cardiovascular disease and T2D combined GWAS datasets to calculate PRS, and identified an association between high PRS and an improved disease status upon adherence to a modified lifestyle (Ye et al., 2021). This clearly highlights the potential that PRS has in capturing more of the variance of polygenic traits, and the ability of PRS to be associated with the environment. Integration of PRS may drastically improve our ability to determine the phenotypic impact of environmental factors.

CONCLUSION

Integrated, multi-omic approaches, in collaboration with environmental data, will help us to robustly decipher the complex relationship between the environment, genome regulation, and associated phenotypes, to produce confirmation of the genome regulatory impacts of environmental exposures that we know are drivers of health impacts. Working in an integrated, multi-layer fashion will give us more power to predict how the environment interacts with genome regulation and influences health; however, there are challenges to develop novel statistical methods, collect cohorts of sufficient size, access sufficient computational resources to perform the analysis, and interpret the results. Future research in this area will transform our understanding of how our genomes respond to and translate an environmental exposure into a phenotype, providing new pathways for investigation into the biological basis of disease.

AUTHOR CONTRIBUTIONS

AO coordinated the review and prepared the manuscript. AN coordinated the review and prepared the manuscript. RP, AA, YL, PR, and JA researched and prepared individual sections for this manuscript. AA reviewed and provided comments on the manuscript.

ACKNOWLEDGMENTS

We are thankful to Professor Justin O'Sullivan and Professor Jack Satsangi for valuable comments on this manuscript.

REFERENCES

Abarca-Gómez, L., Abdeen, Z. A., Hamid, Z. A., Abu-Rmeileh, N. M., Acosta-Cazares, B., Acuin, C., et al. (2017). Worldwide Trends in Body-Mass index, Underweight, Overweight, and Obesity from 1975 to 2016: a Pooled Analysis of 2416 Population-Based Measurement Studies in 128·9 Million Children, Adolescents, and Adults. *Lancet* 390, 2627–2642. doi:10.1016/S0140-6736(17)32129-3

Adams, A. T., Kennedy, N. A., Hansen, R., Ventham, N. T., O'Leary, K. R., Drummond, H. E., et al. (2014). Two-stage Genome-wide Methylation Profiling in Childhood-Onset Crohn's Disease Implicates Epigenetic Alterations at the VMP1/MIR21 and HLA Loci. *Inflamm. Bowel Dis.* 20, 1784–1793. doi:10.1097/mib.0000000000000179

Albuquerque, D., Nóbrega, C., Manco, L., and Padez, C. (2017). The Contribution of Genetics and Environment to Obesity. *Br. Med. Bull.* 123, 159–173. doi:10.1093/bmb/ldx022

Bannister, A. J., and Kouzarides, T. (2011). Regulation of Chromatin by Histone Modifications. *Cell Res* 21, 381–395. doi:10.1038/cr.2011.22

Barouki, R., Gluckman, P. D., Grandjean, P., Hanson, M., and Heindel, J. J. (2012). Developmental Origins of Non-communicable Disease: Implications for Research and Public Health. *Environ. Health* 11, 42–49. doi:10.1186/1476-069X-11-42

Baumgart, D. C., and Carding, S. R. (2007). Inflammatory Bowel Disease: Cause and Immunobiology. *The Lancet* 369, 1627–1640. doi:10.1016/s0140-6736(07)60750-8

Besingi, W., and Johansson, Å. (2014). Smoke-related DNA Methylation Changes in the Etiology of Human Disease. *Hum. Mol. Genet.* 23, 2290–2297. doi:10.1093/hmg/ddt621

Bickmore, W. A., and van Steensel, B. (2013). Genome Architecture: Domain Organization of Interphase Chromosomes. *Cell* 152, 1270–1284. doi:10.1016/j.cell.2013.02.001

Bjornevik, K., Cortese, M., Healy, B. C., Kuhle, J., Mina, M. J., Leng, Y., et al. (2022). Longitudinal Analysis Reveals High Prevalence of Epstein-Barr Virus Associated with Multiple Sclerosis. *Science* 375 (6578), 296–301. doi:10.1126/science.abj8222

Bonev, B., and Cavalli, G. (2016). Organization and Function of the 3D Genome. *Nat. Rev. Genet.* 17, 661–678. doi:10.1038/nrg.2016.112

Borren, N. Z., Plichta, D., Joshi, A. D., Bonilla, G., Sadreyev, R., Vlamakis, H., et al. (2020). Multi-"-Omics" Profiling in Patients with Quiescent Inflammatory Bowel Disease Identifies Biomarkers Predicting Relapse. *Inflamm. Bowel Dis.* 26, 1524–1532. doi:10.1093/ibd/izaa183

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: a Method for Assaying Chromatin Accessibility Genome-wide. *Curr. Protoc. Mol. Biol.* 109, 21.29. 1–21.29. 9. doi:10.1002/0471142727.mb2129s109

Calamini, B., and Morimoto, R. I. (2012). Protein Homeostasis as a Therapeutic Target for Diseases of Protein Conformation. *Curr. Top. Med. Chem.* 12, 2623–2640. doi:10.2174/1568026611212220014

Castel, S. E., and Martienssen, R. A. (2013). RNA Interference in the Nucleus: Roles for Small RNAs in Transcription, Epigenetics and beyond. *Nat. Rev. Genet.* 14, 100–112. doi:10.1038/nrg3355

Cervantes-Gracia, K., and Husi, H. (2018). Integrative Analysis of Multiple Sclerosis Using a Systems Biology Approach. *Sci. Rep.* 8, 5633. doi:10.1038/s41598-018-24032-8

Chappell, L., Russell, A. J. C., and Voet, T. (2018). Single-Cell (Multi)omics Technologies. *Annu. Rev. Genom. Hum. Genet.* 19, 15–41. doi:10.1146/annurev-genom-091416-035324

Chen, N., Miao, L., Lin, W., Zou, D., Huang, L., Huang, J., et al. (2021). Integrated DNA Methylation and Gene Expression Analysis Identified S100A8 and S100A9 in the Pathogenesis of Obesity. *Front. Cardiovasc. Med.* 8, 350. doi:10.3389/fcvm.2021.631650

Consortium, I. M. S. G., Hafler, D. A., Compston, A., Sawcer, S., Lander, E. S., Daly, M. J., et al. (2007). Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study. *N. Engl. J. Med.* 357, 851–862. doi:10.1056/NEJMoa073493

Corbett, S., Courtiol, A., Lummaa, V., Moorad, J., and Stearns, S. (2018). The Transition to Modernity and Chronic Disease: Mismatch and Natural Selection. *Nat. Rev. Genet.* 19, 419–430. doi:10.1038/s41576-018-0012-3

Cottrell, E. C., and Ozanne, S. E. (2008). Early Life Programming of Obesity and Metabolic Disease. *Physiol. Behav.* 94, 17–28. doi:10.1016/j.physbeh.2007.11.017

Cremer, T., and Cremer, C. (2001). Chromosome Territories, Nuclear Architecture and Gene Regulation in Mammalian Cells. *Nat. Rev. Genet.* 2, 292–301. doi:10.1038/35066075

Cubeñas-Potts, C., and Corces, V. G. (2015). Architectural Proteins, Transcription, and the Three-dimensional Organization of the Genome. *FEBS Lett.* 589, 2923–2930.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., et al. (2015). Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing. *Science* 348, 910–914. doi:10.1126/science.aab1601

Dai, H., and Wang, Z. (2014). Histone Modification Patterns and Their Responses to Environment. *Curr. Envir Health Rpt* 1, 11–21. doi:10.1007/s40572-013-0008-2

De Jager, P. L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B. N., Felsky, D., et al. (2018). A Multi-Omic Atlas of the Human Frontal Cortex for Aging and Alzheimer's Disease Research. *Sci. Data* 5, 180142. doi:10.1038/sdata.2018.142

De Nadal, E., Ammerer, G., and Posas, F. (2011). Controlling Gene Expression in Response to Stress. *Nat. Rev. Genet.* 12, 833–845. doi:10.1038/nrg3055

Deane, K. D., Demoruelle, M. K., Kelmenson, L. B., Kuhn, K. A., Norris, J. M., and Holers, V. M. (2017). Genetic and Environmental Risk Factors for Rheumatoid Arthritis. *Best Pract. Res. Clin. Rheumatol.* 31, 3–18. doi:10.1016/j.berh.2017.08.003

Dekker, J., and Misteli, T. (2015). Long-range Chromatin Interactions. *Cold Spring Harb Perspect. Biol.* 7, a019356. doi:10.1101/cshperspect.a019356

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature* 485, 376–380. doi:10.1038/nature11082

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., et al. (2019). Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations. *Nat. Commun.* 10, 3328–3329. doi:10.1038/s41467-019-11112-0

Engreitz, J. M., Ollikainen, N., and Guttman, M. (2016). Long Non-coding RNAs: Spatial Amplifiers that Control Nuclear Structure and Gene Expression. *Nat. Rev. Mol. Cel Biol* 17, 756–770. doi:10.1038/nrm.2016.126

Esteller, M. (2007). Cancer Epigenomics: DNA Methylomes and Histone-Modification Maps. *Nat. Rev. Genet.* 8, 286–298. doi:10.1038/nrg2005

Fadason, T., Ekblad, C., Ingram, J. R., Schierding, W. S., and O'Sullivan, J. M. (2017). Physical Interactions and Expression Quantitative Traits Loci Identify Regulatory Connections for Obesity and Type 2 Diabetes Associated SNPs. *Front. Genet.* 8, 150. doi:10.3389/fgene.2017.00150

Fadason, T., Schierding, W., Lumley, T., and O'Sullivan, J. M. (2018). Chromatin Interactions and Expression Quantitative Trait Loci Reveal Genetic Drivers of Multimorbidities. *Nat. Commun.* 9, 5198. doi:10.1038/s41467-018-07692-y

Fang, L., Wuptra, K., Chen, D., Li, H., Huang, S. K., Jin, C., et al. (2014). Environmental-stress-induced Chromatin Regulation and its Heritability. *J. Carcinog Mutagen* 5. doi:10.4172/2157-2518.1000156

Feil, R., and Fraga, M. F. (2012). Epigenetics and the Environment: Emerging Patterns and Implications. *Nat. Rev. Genet.* 13, 97–109. doi:10.1038/nrg3142

Fernández-Tajes, J., Gaulton, K. J., van de Bunt, M., Torres, J., Thurner, M., Mahajan, A., et al. (2019). Developing a Network View of Type 2 Diabetes Risk Pathways through Integration of Genetic, Genomic and Functional Data. *Genome Med.* 11, 1–14.

Fiehn, O., Garvey, W. T., Newman, J. W., Lok, K. H., Hoppel, C. L., and Adams, S. H. (2010). Plasma Metabolomic Profiles Reflective of Glucose Homeostasis in Non-diabetic and Type 2 Diabetic Obese African-American Women. *PloS one* 5, e15234. doi:10.1371/journal.pone.0015234

Florath, I., Butterbach, K., Muller, H., Bewerunge-Hudler, M., and Brenner, H. (2014). Cross-sectional and Longitudinal Changes in DNA Methylation with Age: an Epigenome-wide Analysis Revealing over 60 Novel Age-Associated CpG Sites. *Hum. Mol. Genet.* 23, 1186–1201. doi:10.1093/hmg/ddt531

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., et al. (2016). Formation of New Chromatin Domains Determines Pathogenicity of Genomic Duplications. *Nature* 538, 265–269. doi:10.1038/nature19800

Freeman, J. R., Chu, S., Hsu, T., and Huang, Y.-T. (2016). Epigenome-wide Association Study of Smoking and DNA Methylation in Non-small Cell Lung Neoplasms. *Oncotarget* 7, 69579–69591. doi:10.18632/oncotarget.11831

Garrison, C. B., Lastwika, K. J., Zhang, Y., Li, C. I., and Lampe, P. D. (2017). Proteomic Analysis, Immune Dysregulation, and Pathway Interconnections with Obesity. *J. Proteome Res.* 16, 274–287. doi:10.1021/acs.jproteome.6b00611

Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., et al. (2011). Comparative Analysis of Proteome and Transcriptome Variation in Mouse. *Plos Genet.* 7, e1001393. doi:10.1371/journal.pgen.1001393

Gluckman, P. D., Hanson, M. A., Cooper, C., and Thornburg, K. L. (2008). Effect of In Utero and Early-Life Conditions on Adult Health and Disease. *N. Engl. J. Med.* 359, 61–73. doi:10.1056/nejmra0708473

Gokuladhas, S., Schierding, W., Cameron-Smith, D., Wake, M., Scotter, E. L., and O'Sullivan, J. (2020). Shared Regulatory Pathways Reveal Novel Genetic Correlations between Grip Strength and Neuromuscular Disorders. *Front. Genet.* 11, 393. doi:10.3389/fgene.2020.00393

Gokuladhas, S., Schierding, W., Golovina, E., Fadason, T., and O'Sullivan, J. M. (2021). Unravelling the Shared Genetic Mechanisms Underlying 18 Autoimmune Diseases Using a Systems Approach. *Front. Immunol.* doi:10.3389/fimmu.2021.693142

Guttman, M., and Rinn, J. L. (2012). Modular Regulatory Principles of Large Non-coding RNAs. *Nature* 482, 339–346. doi:10.1038/nature10887

Ha, E., Bang, S.-Y., Lim, J., Yun, J. H., Kim, J.-M., Bae, J.-B., et al. (2021). Genetic Variants Shape Rheumatoid Arthritis-specific Transcriptomic Features in CD4+ T Cells through Differential DNA Methylation, Explaining a Substantial Proportion of Heritability. *Ann. Rheum. Dis.* 80, 876–883. doi:10.1136/annrheumdis-2020-219152

Hartstra, A. V., Bouter, K. E. C., Bäckhed, F., and Nieuwdorp, M. (2015). Insights into the Role of the Microbiome in Obesity and Type 2 Diabetes. *Diabetes care* 38, 159–165. doi:10.2337/dc14-0769

Herman, J. G., and Baylin, S. B. (2003). Gene Silencing in Cancer in Association with Promoter Hypermethylation. *N. Engl. J. Med.* 349, 2042–2054. doi:10.1056/nejmra023075

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential Etiologic and Functional Implications of Genome-wide Association Loci for Human Diseases and Traits. *Proc. Natl. Acad. Sci.* 106, 9362–9367. doi:10.1073/pnas.0903103106

Hosseini, A., Babaloo, Z., Gharibi, T., Shomali, N., Marofi, F., Hashemi, V., et al. (2020). Epigenetic Mechanisms Shape the Underlining Expression Regulatory Mechanisms of the STAT3 in Multiple Sclerosis Disease. *BMC Res. Notes* 13, 568. doi:10.1186/s13104-020-05427-1

Hüls, A., Wright, M. N., Bogl, L. H., Kaprio, J., Lissner, L., Molnár, D., et al. (2021). Polygenic Risk for Obesity and its Interaction with Lifestyle and Sociodemographic Factors in European Children and Adolescents. *Int. J. Obes.* 45, 1321–1330.

International Multiple Sclerosis Genetics Consortium (2019). Multiple Sclerosis Genomic Map Implicates Peripheral Immune Cells and Microglia in Susceptibility. *Science* 365, eaav7188. doi:10.1126/science.aav7188

Ivorra, C., Fraga, M. F., Bayón, G. F., Fernández, A. F., Garcia-Vicent, C., Chaves, F. J., et al. (2015). DNA Methylation Patterns in Newborns Exposed to Tobacco In Utero. *J. Transl Med.* 13, 25–29. doi:10.1186/s12967-015-0384-5

Jacobs, B. M., Noyce, A. J., Bestwick, J., Belete, D., Giovannoni, G., and Dobson, R. (2021). Gene-Environment Interactions in Multiple Sclerosis: A UK Biobank Study. *Neurology-Neuroimmunology Neuroinflammation* 8. doi:10.1212/nxi.0000000000001007

Jaenisch, R., and Bird, A. (2003). Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals. *Nat. Genet.* 33, 245–254. doi:10.1038/ng1089

Jaffe, A. E., and Irizarry, R. A. (2014). Accounting for Cellular Heterogeneity Is Critical in Epigenome-wide Association Studies. *Genome Biol.* 15, R31–R39. doi:10.1186/gb-2014-15-2-r31

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). Erratum to: The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community. *Genome Biol.* 17, 239. doi:10.1186/s13059-016-1103-0

Jayasinghe, T. N., Chiavaroli, V., Holland, D. J., Cutfield, W. S., and O'Sullivan, J. M. (2016). The new era of Treatment for Obesity and Metabolic Disorders:

Evidence and Expectations for Gut Microbiome Transplantation. *Front. Cel. Infect. Microbiol.* 6, 15. doi:10.3389/fcimb.2016.00015

Ji, Y., Lotfollahi, M., Wolf, F. A., and Theis, F. J. (2021). Machine Learning for Perturbational Single-Cell Omics. *Cel Syst.* 12, 522–537. doi:10.1016/j.cels.2021.05.016

Jirtle, R. L., and Skinner, M. K. (2007). Environmental Epigenomics and Disease Susceptibility. *Nat. Rev. Genet.* 8, 253–262. doi:10.1038/nrg2045

Johnson, W., Li, L., Kuh, D., and Hardy, R. (2015). How Has the Age-Related Process of Overweight or Obesity Development Changed over Time? Co-ordinated Analyses of Individual Participant Data from Five United Kingdom Birth Cohorts. *Plos Med.* 12, e1001828. discussion e1001828. doi:10.1371/journal.pmed.1001828

Jones, P. A. (1999). The DNA Methylation Paradox. *Trends Genetics* 15, 34–37. doi:10.1016/s0168-9525(98)01636-9

Joslin, A. C., Sobreira, D. R., Hansen, G. T., Sakabe, N. J., Aneas, I., Montefiori, L. E., et al. (2021). A Functional Genomics Pipeline Identifies Pleiotropy and Cross-Tissue Effects within Obesity-Associated GWAS Loci. *Nat. Commun.* 12, 1–15. doi:10.1038/s41467-021-25614-3

Joubert, B. R., Håberg, S. E., Nilsen, R. M., Wang, X., Vollset, S. E., Murphy, S. K., et al. (2012). 450K Epigenome-wide Scan Identifies Differential DNA Methylation in Newborns Related to Maternal Smoking during Pregnancy. *Environ. Health Perspect.* 120, 1425–1431. doi:10.1289/ehp.1205412

Kaiser, A. B., Zhang, N., and Van Der Pluijm, W. (2018). Global Prevalence of Type 2 Diabetes over the Next Ten Years (2018-2028). *Diabetes* 67 (Supplement_1), 202-LB. doi:10.2337/db18-202-LB

Khawandanah, J. (2019). Double or Hybrid Diabetes: A Systematic Review on Disease Prevalence, Characteristics and Risk Factors. *Nutr. Diabetes* 9, 33–39. doi:10.1038/s41387-019-0101-1

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* 161, 1187–1201. doi:10.1016/j.cell.2015.04.044

Koga, H., Kaushik, S., and Cuervo, A. M. (2011). Protein Homeostasis and Aging: The Importance of Exquisite Quality Control. *Ageing Res. Rev.* 10, 205–215. doi:10.1016/j.arr.2010.02.001

Krijger, P. H. L., and De Laat, W. (2016). Regulation of Disease-Associated Gene Expression in the 3D Genome. *Nat. Rev. Mol. Cel Biol* 17, 771–782. doi:10.1038/nrm.2016.138

Kurilshikov, A., van den Munckhof, I. C. L., Chen, L., Bonder, M. J., Schraa, K., Rutten, J. H. W., et al. (2019). Gut Microbial Associations to Plasma Metabolites Linked to Cardiovascular Phenotypes and Risk. *Circ. Res.* 124, 1808–1820. doi:10.1161/circresaha.118.314642

La Cognata, V., Morello, G., and Cavallaro, S. (2021). Omics Data and Their Integrative Analysis to Support Stratified Medicine in Neurodegenerative Diseases. *Ijms* 22, 4820. doi:10.3390/ijms22094820

Laszlo, A. H., Derrington, I. M., Brinkerhoff, H., Langford, K. W., Nova, I. C., Samson, J. M., et al. (2013). Detection and Mapping of 5-methylcytosine and 5-hydroxymethylcytosine with Nanopore MspA. *Proc. Natl. Acad. Sci.* 110, 18904–18909. doi:10.1073/pnas.1310240110

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., et al. (2013). Batch Effect Removal Methods for Microarray Gene Expression Data Integration: a Survey. *Brief. Bioinformatics* 14, 469–490. doi:10.1093/bib/bbs037

Lee, Y. M., Li, C., Lam, S. H., Gong, Z., and Lin, Q. (2018). Proteomic Analysis of Zebrafish (*Danio rerio*) after Chemical Exposure. *Springer* 1797, 443–459. doi:10.1007/978-1-4939-7883-0_24

Lee, J., Hyeon, D. Y., and Hwang, D. (2020). Single-cell Multiomics: Technologies and Data Analysis Methods. *Exp. Mol. Med.* 52, 1428–1442. doi:10.1038/s12276-020-0420-2

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data. *Nat. Rev. Genet.* 11, 733–739. doi:10.1038/nrg2825

Li Yim, A. Y., Ferrero, E., Maratou, K., Lewis, H. D., Royal, G., Tough, D. F., et al. (2021). Novel Insights into Rheumatoid Arthritis through Characterization of Concordant Changes in DNA Methylation and Gene Expression in Synovial Biopsies of Patients with Differing Numbers of Swollen Joints. *Front. Immunol.* 12, 1215. doi:10.3389/fimmu.2021.651475

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *science* 326, 289–293. doi:10.1126/science.1181369

Lillycrop, K. A., and Burdge, G. C. (2012). Epigenetic Mechanisms Linking Early Nutrition to Long Term Health. *Best Pract. Res. Clin. Endocrinol. Metab.* 26, 667–676. doi:10.1016/j.beem.2012.03.009

Liu, J. Z., Van Sommeren, S., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., et al. (2015). Association Analyses Identify 38 Susceptibility Loci for Inflammatory Bowel Disease and Highlight Shared Genetic Risk across Populations. *Nat. Genet.* 47, 979–986. doi:10.1038/ng.3359

Lloyd-Price, J., Arze, C., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., et al. (2019). Multi-omics of the Gut Microbial Ecosystem in Inflammatory Bowel Diseases. *Nature* 569, 655–662. doi:10.1038/s41586-019-1237-9

Lujambio, A., Portela, A., Liz, J., Melo, S. A., Rossi, S., Spizzo, R., et al. (2010). CpG Island Hypermethylation-Associated Silencing of Non-coding RNAs Transcribed from Ultraconserved Regions in Human Cancer. *Oncogene* 29, 6390–6401. doi:10.1038/onc.2010.361

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the Missing Heritability of Complex Diseases. *Nature* 461, 747–753. doi:10.1038/nature08494

Marchesi, J. R., and Ravel, J. (2015). *The Vocabulary of Microbiome Research: A Proposal*. Springer.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., et al. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* 100, 635–649. doi:10.1016/j.ajhg.2017.03.004

Maruvada, P., Leone, V., Kaplan, L. M., and Chang, E. B. (2017). The Human Microbiome and Obesity: Moving beyond Associations. *Cell Host & Microbe* 22, 589–599. doi:10.1016/j.chom.2017.10.005

Matilainen, O., Sleiman, M. S. B., Quiros, P. M., Garcia, S. M. D. A., and Auwerx, J. (2017). The Chromatin Remodeling Factor ISW-1 Integrates Organismal Responses against Nuclear and Mitochondrial Stress. *Nat. Commun.* 8, 1818. doi:10.1038/s41467-017-01903-8

Mehta, D., Jackson, R., Paul, G., Shi, J., and Sabbagh, M. (2017). Why Do Trials for Alzheimer's Disease Drugs Keep Failing? A Discontinued Drug Perspective for 2010-2015. *Expert Opin. Investig. Drugs* 26, 735–739. doi:10.1080/13543784.2017.1323868

Mens, M. M. J., Maas, S. C. E., Klap, J., Weverling, G. J., Klatser, P., Brakenhoff, J. P. J., et al. (2020). Multi-omics Analysis Reveals microRNAs Associated with Cardiometabolic Traits. *Front. Genet.* 11, 110. doi:10.3389/fgene.2020.00110

Merino, J., Leong, A., Liu, C.-T., Porneala, B., Walford, G. A., von Grotthuss, M., et al. (2018). Metabolomics Insights into Early Type 2 Diabetes Pathogenesis and Detection in Individuals with normal Fasting Glucose. *Diabetologia* 61, 1315–1324. doi:10.1007/s00125-018-4599-x

Mo, X.-B., Lei, S.-F., Qian, Q.-Y., Guo, Y.-F., Zhang, Y.-H., and Zhang, H. (2019). Integrative Analysis Revealed Potential Causal Genetic and Epigenetic Factors for Multiple Sclerosis. *J. Neurol.* 266, 2699–2709. doi:10.1007/s00415-019-09476-w

Nativio, R., Lan, Y., Donahue, G., Sidoli, S., Berson, A., Srinivasan, A. R., et al. (2020). An Integrated Multi-Omics Approach Identifies Epigenetic Alterations Associated with Alzheimer's Disease. *Nat. Genet.* 52, 1024–1035. doi:10.1038/s41588-020-0696-0

Neff, R. A., Wang, M., Vatansever, S., Guo, L., Ming, C., Wang, Q., et al. (2021). Molecular Subtyping of Alzheimer's Disease Using RNA Sequencing Data Reveals Novel Mechanisms and Targets. *Sci. Adv.* 7, eabb5398. doi:10.1126/sciadv.abb5398

Noble, A. J., Pearson, J. F., Noble, A. D., Boden, J. M., Horwood, L. J., Kennedy, M. A., et al. (2021). *DNA Methylation Analysis Using Bisulfite-Based Amplicon Sequencing of Individuals Exposed to Maternal Tobacco Use during Pregnancy, and Offspring Conduct Problems in Childhood and Adolescence. Reproduction*. Fertility and Development In Press.

Noce, A., Marrone, G., Di Daniele, F., Ottaviani, E., Wilson Jones, G., Bernini, R., et al. (2019). Impact of Gut Microbiota Composition on Onset and Progression of Chronic Non-communicable Diseases. *Nutrients* 11, 1073. doi:10.3390/nu11051073

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., et al. (2012). Spatial Partitioning of the Regulatory Landscape of the X-Inactivation centre. *Nature* 485, 381–385. doi:10.1038/nature11049

Nyaga, D. M., Vickers, M. H., Jefferies, C., Fadason, T., and O'Sullivan, J. M. (2021). Untangling the Genetic Link between Type 1 and Type 2 Diabetes Using Functional Genomics. *Scientific Rep.* 11, 1–15. doi:10.1038/s41598-021-93346-x

O'Gorman, C., Lucas, R., and Taylor, B. (2012). Environmental Risk Factors for Multiple Sclerosis: a Review with a Focus on Molecular Mechanisms. *Int. J. Mol. Sci.* 13, 11718–11752.

Østergaard, S. D., Mukherjee, S., Sharp, S. J., Proitsi, P., Lotta, L. A., Day, F., et al. (2015). Associations between Potentially Modifiable Risk Factors and Alzheimer Disease: a Mendelian Randomization Study. *PLoS Med.* 12, e1001841.

Palmqvist, S., Janelidze, S., Quiroz, Y. T., Zetterberg, H., Lopera, F., Stomrud, E., et al. (2020). Discriminative Accuracy of Plasma Phospho-Tau217 for Alzheimer Disease vs Other Neurodegenerative Disorders. *Jama* 324, 772–781. doi:10.1001/jama.2020.12134

Peters, L., and Meister, G. (2007). Argonaute Proteins: Mediators of RNA Silencing. *Mol. Cel.* 26, 611–623. doi:10.1016/j.molcel.2007.05.001

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., et al. (2017). Multiplexed Quantification of Proteins and Transcripts in Single Cells. *Nat. Biotechnol.* 35, 936–939. doi:10.1038/nbt.3973

Petesch, S. J., and Lis, J. T. (2008). Rapid, Transcription-independent Loss of Nucleosomes over a Large Chromatin Domain at Hsp70 Loci. *Cell* 134, 74–84. doi:10.1016/j.cell.2008.05.029

Pingault, J.-B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F., and Dudbridge, F. (2018). Using Genetic Data to Strengthen Causal Inference in Observational Research. *Nat. Rev. Genet.* 19, 566–580. doi:10.1038/s41576-018-0020-3

Price, E. M., and Robinson, W. P. (2018). Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned. *Front. Genet.* 9, 83. doi:10.3389/fgene.2018.00083

Rand, A. C., Jain, M., Eizenga, J. M., Musselman-Brown, A., Olsen, H. E., Akeson, M., et al. (2017). Mapping DNA Methylation with High-Throughput Nanopore Sequencing. *Nat. Methods* 14, 411–413. doi:10.1038/nmeth.4189

Rauschert, S., Kirchberg, F. F., Marchioro, L., Koletzko, B., Hellmuth, C., and Uhl, O. (2017). Early Programming of Obesity throughout the Life Course: a Metabolomics Perspective. *Ann. Nutr. Metab.* 70, 201–209. doi:10.1159/000459635

Raychaudhuri, S., Sandor, C., Stahl, E. A., Freudenberg, J., Lee, H.-S., Jia, X., et al. (2012). Five Amino Acids in Three HLA Proteins Explain Most of the Association between MHC and Seropositive Rheumatoid Arthritis. *Nat. Genet.* 44, 291–296. doi:10.1038/ng.1076

Risca, V. I., and Greenleaf, W. J. (2015). Unraveling the 3D Genome: Genomics Tools for Multiscale Exploration. *Trends Genet.* 31, 357–372. doi:10.1016/j.tig.2015.03.010

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., et al. (2015). Single-cell ChIP-Seq Reveals Cell Subpopulations Defined by Chromatin State. *Nat. Biotechnol.* 33, 1165–1172. doi:10.1038/nbt.3383

Rowley, M. J., and Corces, V. G. (2016). The Three-Dimensional Genome: Principles and Roles of Long-Distance Interactions. *Curr. Opin. Cel Biol.* 40, 8–14. doi:10.1016/j.ceb.2016.01.009

Sarli, P.-M., Manousopoulou, A., Efthymiou, E., Zouridis, A., Potiris, A., Pervanidou, P., et al. (2021). Liver Proteome Profile of Growth Restricted and Appropriately Grown Newborn Wistar Rats Associated with Maternal Undernutrition. *Front. Endocrinol.* 12. doi:10.3389/fendo.2021.684220

Saxena, A., and Carninci, P. (2011). Long Non-coding RNA Modifies Chromatin. *Bioessays* 33, 830–839. doi:10.1002/bies.201100084

Sayols-Baixeras, S., Subirana, I., Fernández-Sanlés, A., Sentí, M., Lluís-Ganella, C., Marrugat, J., et al. (2017). DNA Methylation and Obesity Traits: An Epigenome-wide Association Study. The REGICOR Study. *Epigenetics* 12, 909–916. doi:10.1080/15592294.2017.1363951

Schierding, W., and O'Sullivan, J. M. (2015). Connecting SNPs in Diabetes: a Spatial Analysis of Meta-GWAS Loci. *Front. Endocrinol.* 6, 102. doi:10.3389/fendo.2015.00102

Schierding, W., Antony, J., Cutfield, W. S., Horsfield, J. A., and O'Sullivan, J. M. (2016). Intergenic GWAS SNPs Are Key Components of the Spatial and Regulatory Network for Human Growth. *Hum. Mol. Genet.* 25, 3372–3382. doi:10.1093/hmg/ddw165

Senol Cali, D., Kim, J. S., Ghose, S., Alkan, C., and Mutlu, O. (2019). Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks and Future Directions. *Brief. Bioinformatics* 20, 1542–1559. doi:10.1093/bib/bby017

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., et al. (2012). Three-dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* 148, 458–472. doi:10.1016/j.cell.2012.01.010

Silventoinen, K., Jelenkovic, A., Sund, R., Hur, Y.-M., Yokoyama, Y., Honda, C., et al. (2016). Genetic and Environmental Effects on Body Mass index from Infancy to the Onset of Adulthood: an Individual-Based Pooled Analysis of 45 Twin Cohorts Participating in the COllaborative Project of Development of Anthropometrical Measures in Twins (CODATwins) Study. *Am. J. Clin. Nutr.* 104, 371–379. doi:10.3945/ajcn.116.130252

Sinha, I., Modesto, J., Krebs, N. M., Stanley, A. E., Walter, V. A., Richie, J. P., Jr, et al. (2021). Changes in Salivary Proteome before and after Cigarette Smoking in Smokers Compared to Sham Smoking in Nonsmokers: A Pilot Study. *Tob. Induced Dis.* 19. doi:10.18332/tid/138336

Slavov, N. (2020). Unpicking the Proteome in Single Cells. *Science* 367, 512–513. doi:10.1126/science.aaz6695

Somineni, H. K., Venkateswaran, S., Kilaru, V., Marigorta, U. M., Mo, A., Okou, D. T., et al. (2019). Blood-derived DNA Methylation Signatures of Crohn's Disease and Severity of Intestinal Inflammation. *Gastroenterology* 156, 2254–2265. e3. doi:10.1053/j.gastro.2019.01.270

Specht, H., Emmott, E., Petelski, A. A., Huffman, R. G., Perlman, D. H., Serra, M., et al. (2021). Single-cell Proteomic and Transcriptomic Analysis of Macrophage Heterogeneity Using SCoPE2. *Genome Biol.* 22, 50–27. doi:10.1186/s13059-021-02267-5

Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., et al. (2017). 3D Structures of Individual Mammalian Genomes Studied by Single-Cell Hi-C. *Nature* 544, 59–64. doi:10.1038/nature21429

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., et al. (2017). Simultaneous Epitope and Transcriptome Measurement in Single Cells. *Nat. Methods* 14, 865–868. doi:10.1038/nmeth.4380

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq Whole-Transcriptome Analysis of a Single Cell. *Nat. Methods* 6, 377–382. doi:10.1038/nmeth.1315

Tani, H., Onuma, Y., Ito, Y., and Torimura, M. (2014). Long Non-coding RNAs as Surrogate Indicators for Chemical Stress Responses in Human-Induced Pluripotent Stem Cells. *PloS one* 9, e106282. doi:10.1371/journal.pone.0106282

Timpson, N. J., Lawlor, D. A., Harbord, R. M., Gaunt, T. R., Day, I. N., Palmer, L. J., et al. (2005). C-reactive Protein and its Role in Metabolic Syndrome: Mendelian Randomisation Study. *The Lancet* 366, 1954–1959. doi:10.1016/s0140-6736(05)67786-0

Tounian, P. (2011). Programming towards Childhood Obesity. *Ann. Nutr. Metab.* 58, 30–41. doi:10.1159/000328038

Tsai, Y.-W., Dong, J.-L., Jian, Y.-J., Fu, S.-H., Chien, M.-W., Liu, Y.-W., et al. (2021). Gut Microbiota-Modulated Metabolomic Profiling Shapes the Etiology and Pathogenesis of Autoimmune Diseases. *Microorganisms* 9, 1930. doi:10.3390/microorganisms9091930

Vaissière, T., Sawan, C., and Herceg, Z. (2008). Epigenetic Interplay between Histone Modifications and DNA Methylation in Gene Silencing. *Mutat. Research/Reviews Mutat. Res.* 659, 40–48.

Van Cauwenberghe, C., Van Broeckhoven, C., and Sleegers, K. (2016). The Genetic Landscape of Alzheimer Disease: Clinical Implications and Perspectives. *Genet. Med.* 18, 421–430. doi:10.1038/gim.2015.117

van der Kolk, B. W., Saari, S., Lovric, A., Arif, M., Alvarez, M., Ko, A., et al. (2021). Molecular Pathways behind Acquired Obesity: Adipose Tissue and Skeletal Muscle Multiomics in Monozygotic Twin Pairs Discordant for BMI. *Cel Rep. Med.* 2, 100226. doi:10.1016/j.xcrm.2021.100226

Ventham, N. T., Kennedy, N. A., Nimmo, E. R., and Satsangi, J. (2013). Beyond Gene Discovery in Inflammatory Bowel Disease: the Emerging Role of Epigenetics. *Gastroenterology* 145, 293–308. doi:10.1053/j.gastro.2013.05.050

Ventham, N. T., Kennedy, N. A., Adams, A. T., Kalla, R., Heath, S., O'leary, K. R., et al. (2016). Integrative Epigenome-wide Analysis Demonstrates that DNA Methylation May Mediate Genetic Risk in Inflammatory Bowel Disease. *Nat. Commun.* 7, 13507–13514. doi:10.1038/ncomms13507

Vileigas, D. F., Harman, V. M., Freire, P. P., Marciano, C. L. C., Sant'Ana, P. G., de Souza, S. L. B., et al. (2019). Landscape of Heart Proteome Changes in a Diet-Induced Obesity Model. *Sci. Rep.* 9, 18050. doi:10.1038/s41598-019-54522-2

W.H. Organization (2016). *Global Report on Diabetes*. Geneva, Switzerland.

W.H. Organization (2019). *World Health Statistics 2019: Monitoring Health for the SDGs, Sustainable Development Goals*. Geneva: World Health Organization.

Wang, Y., and Navin, N. E. (2015). Advances and Applications of Single-Cell Sequencing Technologies. *Mol. Cel.* 58, 598–609. doi:10.1016/j.molcel.2015.05.005

Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., et al. (2019). A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. *Mol. Syst. Biol.* 15, e8503. doi:10.15252/msb.20188503

Wang, R., Li, B., Lam, S. M., and Shui, G. (2020). Integration of Lipidomics and Metabolomics for In-Depth Understanding of Cellular Mechanism and Disease Progression. *J. Genet. Genomics* 47, 69–83. doi:10.1016/j.jgg.2019.11.009

Wang, C. A., Attia, J. R., Lye, S. J., Oddy, W. H., Beilin, L., Mori, T. A., et al. (2021). The Interactions between Genetics and Early Childhood Nutrition Influence Adult Cardiometabolic Risk Factors. *Scientific Rep.* 11, 1–13. doi:10.1038/s41598-021-94206-4

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., et al. (2007). Distribution, Silencing Potential and Evolutionary Impact of Promoter DNA Methylation in the Human Genome. *Nat. Genet.* 39, 457–466. doi:10.1038/ng1990

Weihrauch-Blüher, S., Schwarz, P., and Klusmann, J.-H. (2019). Childhood Obesity: Increased Risk for Cardiometabolic Disease and Cancer in Adulthood. *Metabolism* 92, 147–152.

Whitaker, J. W., Boyle, D. L., Bartok, B., Ball, S. T., Gay, S., Wang, W., et al. (2015). Integrative Omics Analysis of Rheumatoid Arthritis Identifies Non-obvious Therapeutic Targets. *PloS one* 10, e0124254. doi:10.1371/journal.pone.0124254

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., et al. (2019). Genetic Analyses of Diverse Populations Improves Discovery for Complex Traits. *Nature* 570, 514–518. doi:10.1038/s41586-019-1310-4

Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K. E., Zheng, Z., et al. (2018). Genome-wide Association Analyses Identify 143 Risk Variants and Putative Regulatory Mechanisms for Type 2 Diabetes. *Nat. Commun.* 9, 2941. doi:10.1038/s41467-018-04951-w

Ye, Y., Chen, X., Han, J., Jiang, W., Natarajan, P., and Zhao, H. (2021). Interactions between Enhanced Polygenic Risk Scores and Lifestyle for Cardiovascular Disease, Diabetes, and Lipid Levels. *Circ. Genomic Precision Med.* 14, e003128. doi:10.1161/circgen.120.003128

Zhang, A.-h., Qiu, S., Xu, H.-y., Sun, H., and Wang, X.-j. (2014). Metabolomics in Diabetes. *Clinica Chim. Acta* 429, 106–110. doi:10.1016/j.cca.2013.11.037

Zhou, T., Zhang, R., and Ma, J. (2021). The 3D Genome Structure of Single Cells. *Annu. Rev. Biomed. Data Sci.* 4. doi:10.1146/annurev-biodatasci-020121-084709

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Noble, Purcell, Adams, Lam, Ring, Anderson and Osborne. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 February 2022
doi: 10.3389/fgene.2022.762514

Immune ULBP1 is Elevated in Colon Adenocarcinoma and Predicts Prognosis

Guo-Tian Ruan[1,2†], Hai-Lun Xie[1,2†], Li-Chen Zhu[3†], Yi-Zhong Ge[1,2], Lin Yan[4], Cun Liao[5], Yi-Zhen Gong[6] and Han-Ping Shi[1,2*]*

[1]Department of Gastrointestinal Surgery/Department of Clinical Nutrition, Beijing Shijitan Hospital, Capital Medical University, Beijing, China, [2]Beijing International Science and Technology Cooperation Base for Cancer Metabolism and Nutrition, Beijing, China, [3]Department of Immunology, School of Preclinical Medicine, Guangxi Medical University, Nanning, China, [4]Department of Thoracic Surgery, Affiliated Hospital of Guilin Medical College, Guilin, China, [5]Department of Colorectal and Anal Surgery, The First Affiliated Hospital of Guangxi Medical University, Nanning, China, [6]Division of Colorectal and Anal Surgery, Department of Gastrointestinal Surgery, Guangxi Medical University Cancer Hospital, Nanning, China

Background: Colon adenocarcinoma (COAD) is still the main cause of cancer deaths worldwide. Although immunotherapy has made progress in recent years, there is still a need to improve diagnosis, prognosis, and treatment tools. UL-16 binding protein 1 (ULBP1) is a ligand that activates the receptor natural killer cell group 2 receptor D (NKG2D) and plays an important immunomodulatory role. We aimed to investigate the clinical significance of *ULBP1* in COAD.

Methods: We obtained the relevant data from The Cancer Genome Atlas (TCGA). A total of 438 patients with COAD were included in this study, with a mean age of 67.1 ± 13.03 years old, of which 234 (53.42%) were male. The diagnostic value of COAD tumor tissues and adjacent tissues was analyzed by ROC curve. Univariate and multivariate survival analysis investigated the prognostic value of *ULBP1* gene, and Gene Set Enrichment Analysis (GSEA) curve was performed to analyze the biological process and enriched enrichment pathway of *ULBP1* in COAD. Combination survival analysis investigated the combined prognostic effect of prognostic genes.

Results: *ULBP1* gene had a high diagnostic value in COAD [AUC (TCGA) = 0.959; AUC (Guangxi) = 0.898]. Up-regulated *ULBP1* gene of patients with COAD predicted a worse prognosis compared to those patients with down-regulated *ULBP1* gene (Adjusted HR = 1.544, 95% CI = 1.020–2.337, $p = 0.040$). The GSEA showed that *ULBP1* was involved in the apoptotic pathway and biological process of T cell mediated cytotoxicity, regulation of natural killer cell activation, and T cell mediated immunity of COAD. The combination survival analysis showed that the combination of high expression of *ULBP1*, *AARS1*, and *DDIT3* would increase the 2.2-fold death risk of COAD when compared with those of low expression genes.

Conclusion: The immune-related *ULBP1* gene had diagnostic and prognostic value in COAD. The combination of *ULBP1*, *AARS1*, and *DDIT3* genes could improve the prognostic prediction performance in COAD.

Keywords: ULBP1, COAD, NKG2D, GSEA, biomarker (BM)

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Emanuele Micaglio,
IRCCS San Donato Polyclinic, Italy
Andrew Dellinger,
Elon University, United States

***Correspondence:**
Yi-Zhen Gong
gongyizhen@gxmu.edu.cn
Han-Ping Shi
shihp@ccmu.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: *22 August 2021*
Accepted: *10 January 2022*
Published: *08 February 2022*

Citation:
Ruan G-T, Xie H-L, Zhu L-C, Ge Y-Z, Yan L, Liao C, Gong Y-Z and Shi H-P (2022) Immune ULBP1 is Elevated in Colon Adenocarcinoma and Predicts Prognosis. Front. Genet. 13:762514. doi: 10.3389/fgene.2022.762514

INTRODUCTION

According to Global Sung et al. (2021), it is estimated that there will be more than 1.9 million new cases of colorectal cancer (CRC) and 935,000 deaths, accounting for about one-tenth of cancer cases and deaths. Overall, the incidence of CRC ranks third, but the mortality rate ranks second (Sung et al., 2021). Colon adenocarcinoma (COAD) is a common malignant tumor of the digestive system, and it is the most frequently diagnosed histological subtype of CRC (Siegel et al., 2020). The patient's clinical symptoms usually manifest as diarrhea, abdominal pain, and bloody stools, which develop in the middle and late stages of the disease. The quality of life of patients is usually very low, and the prognosis of most patients is poor. The occurrence and development of COAD are the results of a variety of mixed factors *in vivo* and *in vitro*, which involve a series of molecules and signal pathways (Wang et al., 2021). Although substantive diagnosis and treatment strategies such as surgery, neoadjuvant therapy, and targeted therapy are constantly being developed, the recurrence rate of postoperative COAD is still high, and the 5-year survival rate of patients with advanced COAD is still very low (The 5-year survival rate after distant metastasis is less than 15%) (Patel et al., 2021). Thus, there is an urgent need to explore new biomarkers and therapeutic targets in clinical practice to improve the survival rate of patients with COAD.

Natural killer cell group 2 receptor D (NKG2D) is an alkaline-activated receptor belonging to the c-type lectin-like family. It is expressed in NK cells, most NKT cells, some $\gamma\delta$ T+, and CD8 T+ cells. It is different from other NKG2 receptors, which is not associated with CD94 (Mondelli, 2012) and has nothing to do with CD94 (Mondelli, 2012). The seemingly unchanged activation receptor NKG2D is mixed with a variety of ligands, such as the major histocompatibility complex class I-related chain A and B (MICA and MICB) and the unique long 16 (UL16)-binding protein family (ULBPs, ULBP1-6) which are poor (Champsaur and Lanier, 2010). NKG2D ligand expression is usually lacking in healthy tissue but can induce expression under stress, infection, and DNA damage. NKG2D ligand is also widely expressed in a variety of cancer cell lines and primary solid tumors (McGilvray et al., 2009; Champsaur and Lanier, 2010). The upregulation of these ligands may break NK cells from inhibiting the balance of activation (induced self-identification), with significant biological significance. The interaction of NKG2D is variable between different types of cancers. In the mouse model, the tumor cell line of transfection of RaE1 is rejected by NKG2D-mediated immunization (Diefenbach et al., 2001). The most recent NKG2D knockout mice provide the most convincing evidence for NKG2D to participate in anti-tumor immune responses (Guerra et al., 2008). Many mechanisms have been proposed, cancer can evade NKG2D-mediated immune response. In some systems, the persistent expression of NKG2D ligands can cause NKG2D expression to be lowered (Oppenheim et al., 2005). These results indicate that NKG2D's participation in the anti-cancer immune response is significantly different between different types of cancer. It is also proposed that tumors may release soluble NKG2D ligands, or secrete immunosuppressive cytokines, such as transforming growth factor-beta to reduce NKG2D expression (Groh et al., 2002; Castriconi et al., 2003; Lee et al., 2004). Notably, NKG2D ligands can be independently expressed in human cell lines and primary tumors, the expression of NKG2D ligands among different tumors in knockout mice is also heterogeneous (McGilvray et al., 2009). The complex interaction between NKG2D and its ligands may involve the natural history and treatment response of various cancers (Mondelli, 2012).

The authors showed that *ULBP1*, one of the important ligands of NKG2D, is up-regulated in COAD cancer tissues, but is low-expressed in normal adjacent tissues. Although most previous studies reported that *ULBP1* was related to recurrence-free survival, disease-free survival, or overall survival (OS) in different cancers (McGilvray et al., 2009; McGilvray et al., 2010; Mondelli, 2012; Chen et al., 2013; Maccalli et al., 2017), the relationship between *ULBP1* and OS in COAD has not been reported yet. Therefore, our study uncovers and investigates the diagnosis, prognosis, and immune mechanism of *ULBP1* gene in COAD, which may help make this immune receptor an exceptional candidate for basic and applied cancer research in COAD.

MATERIALS AND METHODS

Public Data Collection

We downloaded the COAD-related *ULBP1* gene mRNA expression data set and the corresponding patient clinical information parameters from the public cancer database-The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/, obtained on December 10, 2020) (Giordano, 2014; Hutter and Zenklusen, 2018). Based on the TCGA-COAD project data, the differential expression level of *ULBP1* in tumor tissues and adjacent normal tissues in pan-cancers was obtained from the TIMER website (https://cistrome.shinyapps.io/timer/, obtained on June 10, 2021) (Liu et al., 2021). The expression level of *ULBP1* gene in COAD tumor tissues and adjacent normal tissues was also obtained from the GEPIA website, which integrated the expression levels of normal tissues in the TCGA database and the Genotype-Tissue Expression (GTEx) database. Additionally, the expression level of *ULBP1* in different COAD tumor stages was obtained from Gene Expression Profiling Interactive Analysis (GEPIA; http://gepia.cancer-pku.cn/index.html; obtained on June 10, 2021) (Tang et al., 2017). We obtained the methylation level and mutation status of *ULBP1* gene from UALCAN database (http://ualcan.path.uab.edu/index.html, obtained on June 10, 2021) (Li et al., 2021) and cBio Cancer Genomics Portal (cBioPortal, https://www.cbioportal.org/, obtained on June 12, 2021), respectively (Harbig et al., 2021). The ULBP1-expressed protein in COAD cancer tissues and adjacent normal tissues was obtained from THE HUMAN PROTEIN ATLAS (HPA, https://www.proteinatlas.org/, obtained on June 12, 2021) (Ullah et al., 2021). Finally, we obtained the information on immune infiltration associated with *ULBP1* in COAD and the correlation between *ULBP1* gene expression level and Genomics of Drug Sensitivity in

Cancer (GDSC) drug sensitivity test or The Cancer Therapeutics Response Portal (CTRP) drug sensitivity test in pan-cancer from the Gene Set Cancer Analysis (GSCA, https://www.proteinatlas.org/, obtained on June 13, 2021) (Ji et al., 2016).

Validation of the Differential Expression and Diagnostic Value of UL-16 Binding Protein 1

COAD tumor tissues and adjacent normal tissues were obtained from the Department of Colorectal Surgery, First Affiliated Hospital of Guangxi Medical University. Inclusion criteria included: 1. The age is not less than 18 years old; 2. The postoperative pathological diagnosis is COAD; 3. Sign the surgical consent form and informed consent form; 4. The length of the hospital stay is more than 48 h. Exclusion criteria included: 1. Suffer from multiple tumors at the same time; 2. Has received preoperative neoadjuvant radiotherapy and chemotherapy; 3. Refuse to sign the informed consent form; 4. The age is less than 18 years old; 5. The length of hospitalization is less than 48 h. After the tissue was excised, it was cut into RNA protection solution and quickly stored in the refrigerator at −80°C. The total RNA extracted from the tissue was reverse transcribed into cDNA and then the qPCR reaction program was performed. The PCR reaction program was performed according to the following conditions: 95°C for 10 min, 1 cycle; 95°C for 15 s, 60°C for 1 min, and 95°C for 30 s, 40 cycles; 95°C for 15 s, 60°C 1 min, 95°C for 30 s, 60°C for 15 s, 1 cycle. Use *GAPDH* as a reference gene: upstream, 5′-GTCAGCCGCATCTTCTTT-3'; downstream, 5′-CGCCCAATACGACCAAAT-3'. The target *ULBP1* gene sequence was: upstream, 5′-CACACACTGTCT TTGCTATGAC-3'; downstream, 5′- CCAGGTTTTTGTGAC ATTGACT-3'. The relative expression level of *ULBP1* gene was performed according to previous descript method of 2-ΔΔ Cq (Ruan et al., 2020a).

Comprehensive Analysis of the Clinical Value of UL-16 Binding Protein 1 Gene Based on the The Cancer Genome Atlas Cohort

In the TCGA database, patients were divided into two high- and low-expression groups based on the median cut-off value of *ULBP1* gene expression. Univariate and multivariate survival analysis was performed to assess the potential prognostic value of *ULBP1* gene expression in patients with COAD.

According to the expression level of *ULBP1* gene, the COAD expression genome-wide data in the TCGA database was divided into high expression group and low expression group. When the gene expression satisfied $|\log_2 \text{foldchange}| \geq 1$ and $p < 0.05$, it was considered to be a differential expression gene in this study.

We investigated the co-expression analysis of *ULBP1* and COAD-related genes in the TCGA cohort. When the Pearson correlation coefficient ≥0.3 or ≤ −0.3, and the *p*-value < 0.05, it was considered to be a co-expressed gene with the *ULBP1* in COAD. The top 20 co-expressed genes were selected to analyze the prognostic value of genes in COAD. Significant prognostic genes were selected to construct a risk score model based on the prognostic contribution coefficients (β) of different genes. The risk score was generated based on the calculation formula: gene expression of $1^*\beta 1+$ gene expression of $2^*\beta 2+\ldots+$gene expression of $n^*\beta n$ (Ruan et al., 2020b).

Statistical Analysis

In this study, the unpaired Student's *t*-test or paired *t*-test was used to compare the expression levels between two groups. The gene expression level was expressed by using the mean ± standard deviation (SD). If the data did not conform to the normal distribution, the rank sum test was used. Univariate and multivariate cox regression analyses were performed to investigate the prognostic value of genes. The selection of adjustment variables adopted single-factor meaningful clinical parameters, and the TNM stage was used as an adjusted factor for prognostic adjustment to reduce the clinical deviation. All two-tailed $p < 0.05$ were considered statistically significant. The statistical analyses in this study were performed using SPSS 22.0 version and R platform, version 4.0.1.

RESULTS

Baseline Characteristics

In this study, a total of 456 patients with *ULBP1* mRNA expression data set in COAD were obtained from the TCGA database, including 480 tumor tissue samples and 41 adjacent normal tissue samples. After removing the duplicate information and the information with a survival time of 0, we obtained a total of 438 tumor sample information and 42 adjacent normal tissue sample information. The mean age of the 438 patients was 67.1 ± 13.03 years old, including 234 males (53.42%) and 204 females (46.58%). Clinical parameter information included age, sex, and TNM stage. The univariate survival analysis of clinical parameters showed that only the TNM stage had a significant prognostic value in COAD ($p < 0.001$, **Table 1**).

Investigating the Association Between UL-16 Binding Protein 1 Gene Expression and Immune Infiltration and Drug Sensitivity

Based on the GSCA website, the association between *ULBP1* gene expression and immune infiltration suggested that the *ULBP1* expression was significantly positively related to the cells of nTreg, iTreg, Neutrophil, Monocyte, Gamma_delta, Exhausted, and CD8_navie. However, the inverse relationship was observed in the cells of NK, NKT, Tfh, Th17, Th2, Tr1, MAIT, Cytotoxic, CD8_T, CD4_T, and B cell. Additionally, a correlational relationship was observed in *ULBP1* gene expression and the majority of drug sensitivity (**Figure 1**).

Differential Expression Analysis and Diagnostic ROC Curve Analysis

We obtained the *ULBP1* gene in COAD tumor tissues and normal tissues from the GEPIA database that matched the information from the GTEx database and found that *ULBP1* expression was

TABLE 1 | Baseline patient characteristics in TCGA cohort

Variables	Patients	OS			
	(*n* = 438)	No. of events	MST (days)	HR(95%CI)	Log-rank P
Age (years)					
≤65	175	29	NA	1	0.062
>65	261	68	2,134	0.064 (0.429–1.024)	—
Missing*	2	—	—	—	—
Sex					
Male	234	54	2,475	1	0.545
Female	204	44	NA	0.884 (0.593–1.318)	—
TNM stage					
I	73	4	NA	1	<0.001
II	167	27	2,821	2.240 (0.781–6.421)	—
III	126	31	NA	4.068 (1.434–11.538)	—
IV	61	31	858	11.291 (3.980–32.026)	—
Missing#	11	—	—	—	—

Notes: Missing, information of age was unknown in 2 patients; Missing#, information of TNM stage was not reported in 10 patients; TCGA, The Cancer Genome Atlas; OS, overall survival; MST, median survival time; 95 % CI, 95 % confidence interval; HR, hazards ratio; NA, not available.*

up-regulated in tumor tissues and also found that the expression level of *ULBP1* gene increased with the progression of tumor stages (**Figures 2A,B**). It was also found that the expression level of *ULBP1* in most tumor tissues was higher than that in normal tissues adjacent to cancer (**Figure 3A**). There was no significant difference in the methylation level of *ULBP1* gene in tumor tissues and adjacent normal tissues in COAD, and the mutation rate of *ULBP1* gene in COAD was 0% (**Figures 2C,D**). *ULBP1* expressed protein was mainly expressed in the cytoplasm (**Figures 2E–G**).

Based on the TCGA cohort, we analyzed the expression level and diagnostic value of *ULBP1* gene in COAD. The results showed that the expression level of *ULBP1* gene in COAD tumor tissues was significantly higher than that in adjacent normal tissues ($p < 0.001$). Simultaneously, it had a higher diagnostic value in COAD (AUC = 0.898, 95%CI = 0.784–1.000, $p < 0.0001$) (**Figure 3B**).

The validation result based on the Guangxi cohort found that the expression level of *ULBP1* gene in COAD tumor tissue was significantly higher than that in adjacent normal tissues ($p = 0.0028$). The diagnostic ROC curve results showed that *ULBP1* has a higher diagnostic value in COAD (AUC = 0.959, 95%CI = 0.942–0.976, $p < 0.001$) (**Figure 3C**).

Survival Analysis of UL-16 Binding Protein 1 Gene in Colon Adenocarcinoma

The results of univariate survival analysis included the TNM stage as an adjusted prognostic factor. After adjustment, the high expression of *ULBP1* gene in COAD predicted a worse OS compared to patients with low expression of *ULBP1* (Adjusted HR = 1.544, 95%CI = 1.020–2.337, $p = 0.040$) (**Figure 4A**; **Table 2**).

We divided 438 COAD patients into high- and low-expression groups based on the median cut-off value of *ULBP1* expression. At the same time, the COAD genome-wide data was also divided into two groups, and the differential analysis and enrichment pathway analysis of these two groups were carried out. The *PHGDH* gene in the high expression group was significantly up-regulated, while the down-regulated genes included *ITLN1*, *JCHAIN*, *DUOXA2*, *CLCA1*, *PRAC1*, *ADH1B*, *GCG*, *IGLL5*, *NXPE4*, *DUOX2*, *CHP2*, and *SI*. The enrichment pathway analysis showed that these differential genes might involve in the process of extracellular exosome and immunoglobulin receptor binding (**Figure 5**).

Based on the *ULBP1* expression levels, we performed the Gene Set Enrichment Analysis (GSEA) to investigate the potential prognosis molecular mechanism of *ULBP1* in COAD. The GSEA was performed by the tool of GSEA 4.1.0 version. The internal reference genes of GSEA were obtained from the Molecular Characterization Database (MSIGDB): KEGG pathway: c2.cp.kegg.v7.4.symbols.gmt; GO term: c5.go.v7.4.symbols.gm. In this study, nominal $p < 0.05$ and false discovery rate (FDR) < 0.25 were considered statistically significant. The results showed that *ULBP1* gene might involve in the development of COAD by participating in the apoptosis pathway and the biological process of T cell mediated cytotoxicity, regulation of natural killer cell activation, and T cell mediated immunity. (**Figure 6**; **Supplementary Figure S1**).

UL-16 Binding Protein 1 Related Co-expression Analysis and Prognostic Analysis in Colon Adenocarcinoma

Based on all gene expression sequences of the TCGA database, the potential clinical value of *ULBP1* gene and *ULBP1* related co-expressed genes was investigated. A total of 87 co-expressed genes related to *ULBP1* in COAD were mined. Pathway analysis of 87 co-expressed genes showed that co-expressed genes were involved in metabolic pathways of COAD (**Figure 7**).

The prognostic value of the top 20 *ULBP1*-related co-expressed genes in COAD has also been investigated. The multivariate survival analysis showed that patients with COAD with high expression of alanyl-tRNA synthetase 1 (*AARS1*) (Adjusted HR = 1.583, 95%CI = 1.043–2.401, $p = 0.031$) or DNA damage inducible transcript 3 (*DDIT3*) (Adjusted HR =

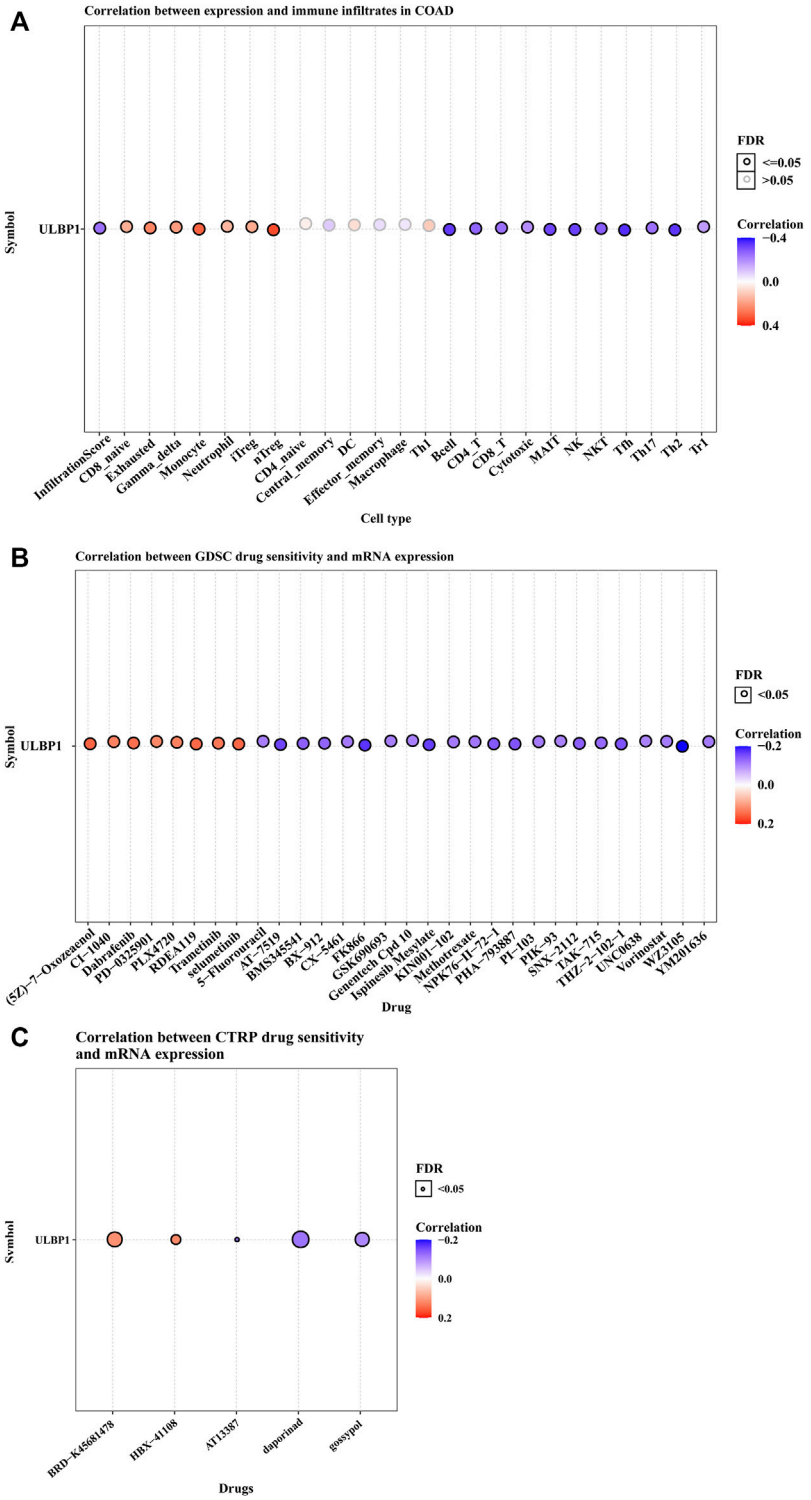


FIGURE 1 | ULBP1's immune infiltration in COAD and the relationship between ULBP1 and pan-cancer drug sensitivity tests based on GSCA. **(A)** immune infiltration; **(B,C)** GDSC and CTRP drug sensitivity test. Notes: COAD, colon adenocarcinoma; ULBP1, unique long 16 (UL16)-binding protein 1; GDSC, Genomics of Drug Sensitivity in Cancer; CTRP, The Cancer Therapeutics Response Portal; GSCA, Gene Set Cancer Analysis.

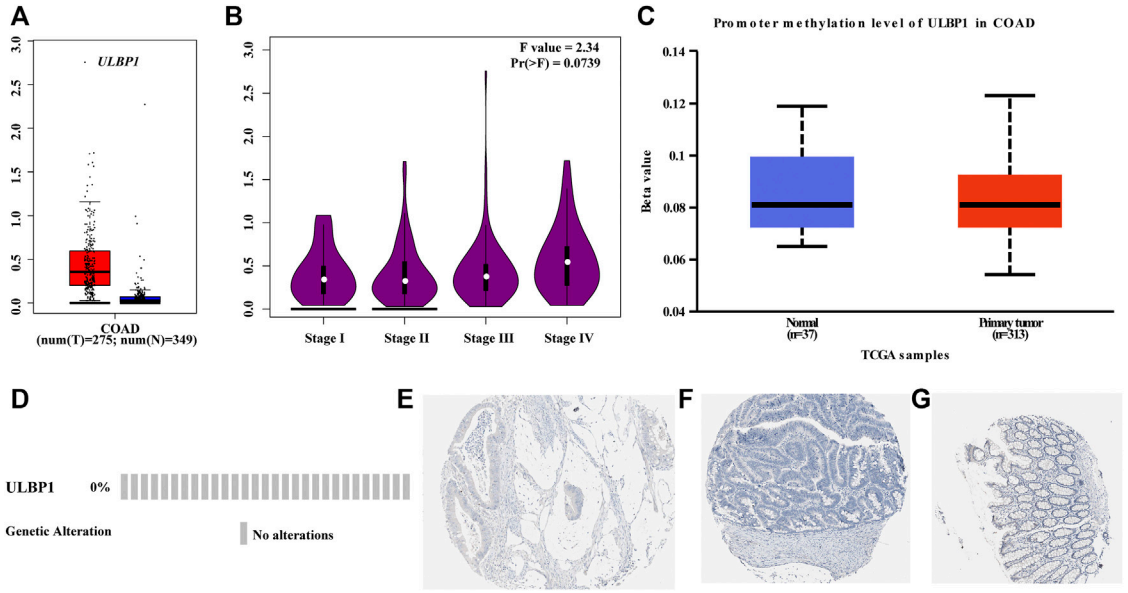


FIGURE 2 | *ULBP1* and *ULBP1* methylation levels in tumor tissues and adjacent normal tissues, ULBP1 mutations, and protein expression in COAD. **(A,B)** GEPIA data: *ULBP1* expression level in COAD and different COAD tumor stages; **(C)** *ULBP1* methylation levels in COAD; **(D)** *ULBP1* mutation; **(E–G)** immunohistochemistry. Notes: COAD, colon adenocarcinoma; *ULBP1*, unique long 16 (UL16)-binding protein 1; GEPIA, Gene Expression Profiling Interactive Analysis.

1.556, 95%CI = 1.013–2.390, p = 0.044) had worse OS when compared with patients with low expression of *DDIT3* or *AARS*, respectively (**Figures 4B,C**; **Table 2**). The combined analysis of *ULBP1*, *DDIT3*, and *AARS* genes showed that the risk of death in COAD patients with High *ULBP1* & High *DDIT3* & High *AARS1* was 2.210-fold higher than that of COAD patients with Low *ULBP1* & Low *DDIT3*& Low *AARS1* (Adjusted HR = 1.180–4.140, p = 0.013) (**Figure 4D**; **Table 3**).

The model of risk score we constructed by the formula: Risk score = *ULBP1**0.434 + *DDIT3**0.442 + *AARS* *0.459. The higher the gene expression, the higher the risk score, and the higher the patient's risk of death. The time-dependent ROC curve results showed that the 1-, 3-, and 5-year AUCs were 59.2, 56.8, and 57.5, respectively (**Figure 8**).

DISCUSSION

It is now widely accepted that tumors develop methods to evade anti-cancer immunity through a process called immunoediting (Dunn et al., 2004). Various evidence from *in vivo* models indicates that the immune system attacks early-stage tumors. Surviving cancer cells must adapt to avoid the immune system. This process is described as immunoediting, immune sculpting, or cancer immune evasion (Dunn et al., 2004). In recent years, studies on tumor models *in vivo* strongly indicate that the activated immune receptor NKG2D participates in the anti-cancer immune response, and it has also attracted much attention as a ligand of the NKG2D receptor (Smyth et al., 2005; Guerra et al., 2008; McGilvray et al., 2009). In humans, primary tumors and tumor cell lines express NKG2D ligands at a high frequency (McGilvray et al., 2009). As an important member of NKG2D ligand, *ULBP1* also plays an important role in immune regulation. The expression of *ULBP1* was associated with majority of immune cells, including NK cells, most NKT cells, $\gamma\delta$ T^+ and CD8 T+ cells. etc. Interestingly, we can take targeted chemotherapy based on the results of drug susceptibility to the immune-related *ULBP1* gene.

After mining through the database, it was found that *ULBP1* was highly expressed in the majority of tumors (including COAD), compared with adjacent tumor tissues. However, some tumors expressed the opposite trend, such as lung adenocarcinoma (LAUD) (**Figure 2A**). As previously described, it also showed that the expression of *ULBP1* in different cancers was different, but the general expression was frequently expressed in cancer tissues. When we analyzed the diagnostic value of *ULBP1* gene expression differences in COAD, whether it was the TCGA cohort or the Guangxi validated cohort, we found that *ULBP1* had a higher diagnostic value (TCGA cohort:0.959; Guangxi cohort:0.898) in COAD. In other words, we can take advantage of this high expression characteristic in cancer tissues, and the immune-related *ULBP1* can better distinguish cancer tissues from normal tissues. Additionally, we also found that with the progress of TMN staging, the expression of *ULBP1* showed an upward trend. The expression level of ULBP1 gene was related to the tumor grade and prognosis, and the differential expression level of ULBP1 gene was different in tumor tissues and adjacent normal tissues of patients with different tumor stages. Univariate and multivariate survival analysis results showed that low expression of *ULBP1* in patients with COAD had a worse prognosis when compared with those patients with high expression of *ULBP1*. The differential

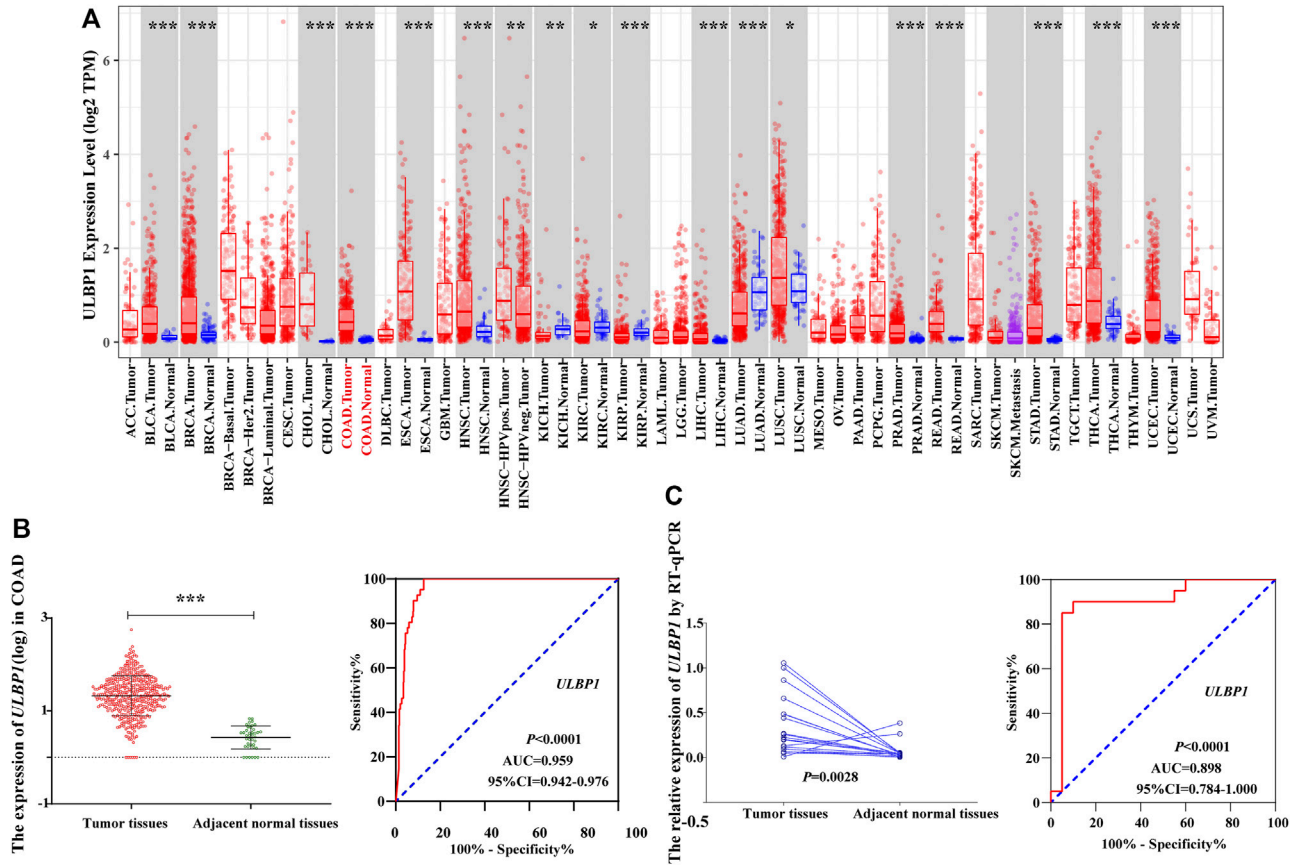


FIGURE 3 | The expression level of *ULBP1* gene in pan-carcinoma and the diagnostic ROC curve. **(A)** The expression level of *ULBP1* gene in pan-cancers based on TIMER; **(B,C)** ROC curve of TCGA cohort and Guangxi validation cohort. Notes: COAD, colon adenocarcinoma; ULBP1, unique long 16 (UL16)-binding protein 1; ROC, receiver operating characteristic curve. *$p < 0.05$; ***$p < 0.001$; ****$p < 0.0001$.

expression results between the high and low groups of *ULBP1* expression indicated that it was related to the binding of immunoglobulins. In addition, the GSEA of *ULBP1* gene in COAD suggested that *ULBP1* was involved in the occurrence and development of COAD through enrichment of apoptosis pathways, and was related to the immunoregulation of T cells and NK cells. We suspected that upregulated-*ULBP1* might participate in the apoptosis process of COAD through its unique immune regulation mechanism.

A study by CADOUX et al. also found that *ULBP1* was expressed at a higher level in hepatocellular carcinoma (HCC) tumors with lower differentiation and higher grades, but the difference is not significant (Cadoux et al., 2021). Interestingly, a study of 462 primary colorectal tumors by McGilvray et al. investigated the ULBP1-expressed protein in different TNM stages, the result showed the opposite trend was that high expression level of ULBP1 was common in TNM stage I tumors, but gradually decreased in stage II, III, and IV tumors (McGilvray et al., 2009). To understand the difference, we also investigated the expression level of *ULBP1* in rectal adenocarcinoma from the data platform (**Supplementary Figure S2**), the trend was consistent with the description of McGilvray et al. (McGilvray et al., 2009), indicating that the expression of *ULBP1* in the colon and rectum was also heterogeneous. Changes in the expression level of *ULBP1* are inseparable from tumor differentiation and grade. In other words, *ULBP1* is closely related to tumor prognosis. However, previous reports described the potential mechanism and prognosis of *ULBP1* expression changed. A study of genome-wide screen to identify novel drivers of *ULBP1* expression by Gowen et al. showed that in the multiple stages of *ULBP1* biogenesis, independent pathways gradually play a role. The transcription factor ATF4 drives the expression of *ULBP1* gene in cancer cells, while the RNA binding protein RBM4 supports the expression of *ULBP1* by inhibiting a new alternative splicing subtype of *ULBP1* mRNA, and explains its mechanism of activating the body's immune system (Gowen et al., 2015). The study by Chava et al. indicated that DOT1L inhibition could regulate apoptotic and metabolic pathways as well as upregulate the expression of *ULBP1* that increased in NK cell-mediated ovarian cancer eradication (Chava et al., 2021). Maccalli et al. showed that patients with melanoma with the negative expression of sULBP-1 were associated with a better prognosis than those patients with positive expression of sULBP-1 (OS: 25.3 months *vs.* 12.1 months) (Maccalli et al., 2017). On the contrary, a study by CADOUX et al. showed that the high expression of *ULBP1* was

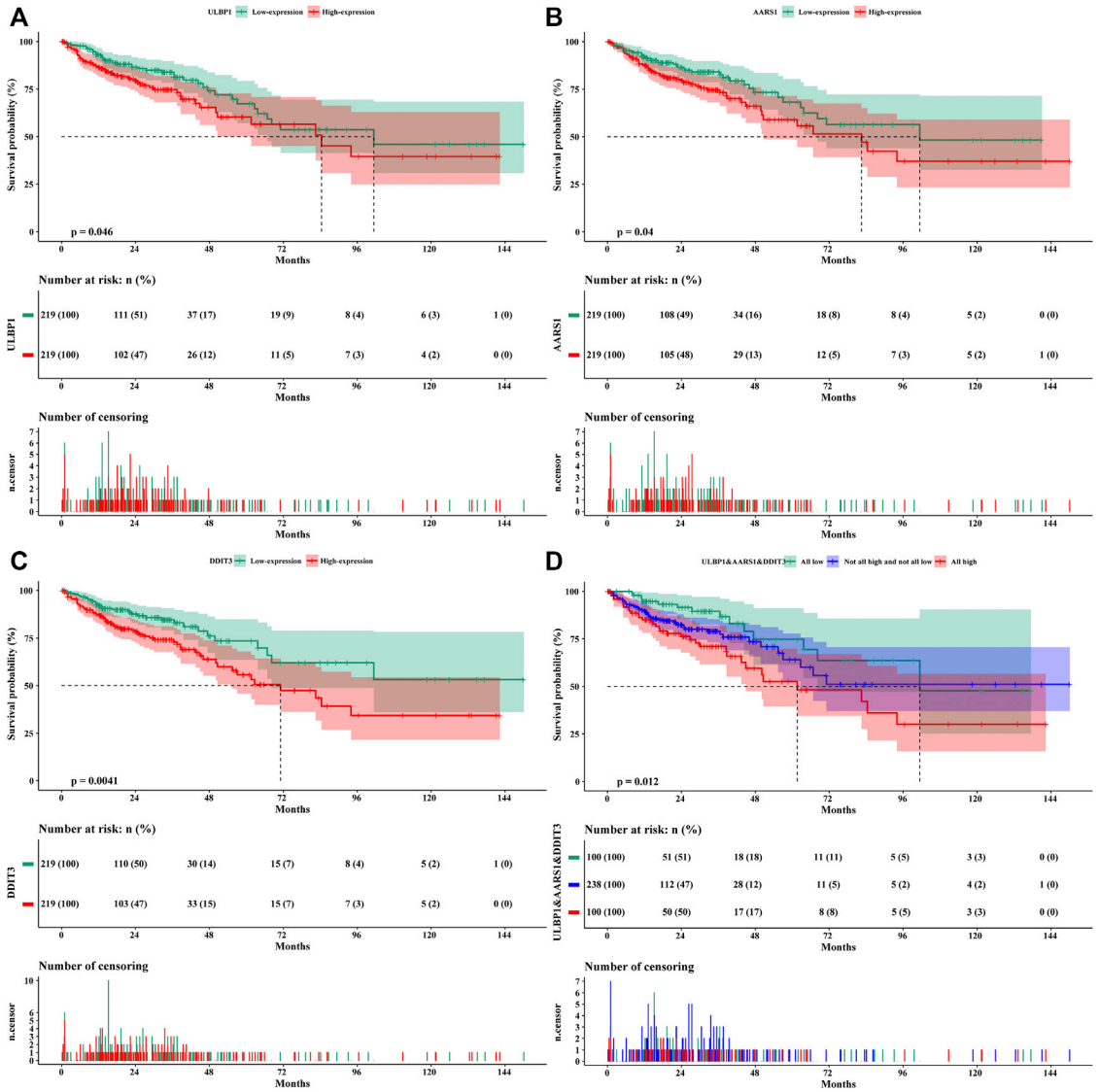


FIGURE 4 | Kaplan–Meier survival curves of genes in COAD. **(A)** *ULBP1*; **(B)** *AARS1*; **(C)** *DDIT3*; **(D)** *ULBP1*& *AARS1* &DDIT3. Notes: COAD, colon adenocarcinoma; ULBP1, unique long 16 (UL16)-binding protein 1; AARS1, alanyl-tRNA synthetase 1; DDIT3, DNA damage inducible transcript 3.

related to the aggressiveness of hepatocellular carcinoma, and the expression of *ULBP1* could be down-regulated through the β-catenin signaling pathway (Cadoux et al., 2021). The study by McGilvray et al. also indicated that patients with ovarian cancer with high expression of *ULBP1* had a worse survival than those patients with no expression of *ULBP1* (disease-specific survival: 14 months *vs.* 30 months) (McGilvray et al., 2010).

The interaction between NKG2D and its ligands may play a central role in anti-tumor surveillance. The level of NKG2D ligands may determine the strength of the anti-tumor immune response (Wu et al., 2012). As described above, tumors can lead to tumor re-editing through immune evasion or ligand shedding. Different cancers are heterogeneous, and we should treat different cancers differently in their anticancer immune responses. To directly avoid NKG2D recognition, tumors may secrete TGF-β and/or release soluble NKG2D ligands, thereby down-regulating the expression of NKG2D (Lee et al., 2004). This observation was also observed in NKG2D knockout mice. For example, the incidence of MCA-induced fibrosarcoma was not affected when knocked out, but the incidence of large prostate tumors, when knocked out, was much higher than that of wild-type (Guerra et al., 2008). Butler et al. confirmed that p53 family members play an important role in the upregulation of *ULBP1* in head and neck squamous cell carcinoma induced by proteasome inhibitor drugs (Butler et al., 2009). It is well known that the activation of immune response by NKG2D depends on the tissue microenvironment and synergizes/antagonizes the signals induced by other cell receptors and cytokines (Eagle et al., 2009). A similar description was suggested by CADOUX et al. that the activated NKG2D system led to a strong inflammatory

TABLE 2 | Prognostic values of *ULBP1* and top 20 *ULBP1*-coexpression genes in COAD

Gene	Patients	OS			
	(*n* = 438)	No. of events	MST (days)	HR (95%CI)	Adjusted $P^{\&}$
ULBP1					
Low	219	43	3,042	1	0.040
High	219	55	2,532	1.544 (1.020–2.337)	—
ZNF534					
Low	219	47	3,042	1	0.645
High	219	51	2,475	0.908 (0.602–1.369)	—
ZNF578					
Low	219	47	2,821	1	0.144
High	219	51	2,532	1.357 (0.901–2.043)	—
ZNF761					
Low	219	42	3,042	1	0.382
High	219	56	2,532	1.373 (0.908–2.075)	—
CLGN					
Low	219	45	3,042	1	0.440
High	219	53	2,134	0.850 (0.563–1.284)	—
ASNS					
Low	219	46	2,134	1	0.405
High	219	52	2,821	1.191(0.789–1.798)	—
TUBE1					
Low	219	50	2,532	1	0.924
High	219	48	2,475	0.980(0.648–1.482)	—
DMGDH					
Low	219	47	3,042	1	0.810
High	219	51	2047	0.950 (0.623–1.448)	—
UPK1A					
Low	219	40	3,042	1	0.185
High	219	58	2,134	1.332 (0.872–2.037)	—
DCDC1					
Low	219	48	2,821	1	0.876
High	219	50	2047	1.033 (0.687–1.553)	—
PSAT1					
Low	219	48	2047	1	0.217
High	219	50	2,821	1.297 (0.858–1.961)	—
AGBL3					
Low	219	44	2,532	1	0.439
High	219	54	2,134	1.177 (0.779–1.777)	—
SLC4A5					
Low	219	47	3,042	1	0.848
High	219	51	2,532	0.960 (0.633–1.456)	—
YARS					
Low	219	47	2,475	1	0.545
High	219	51	3,042	1.135 (0.754–1.707)	—
DDIT3					
Low	219	35	NA	1	0.044
High	219	63	2,134	1.556 (1.013–2.390)	—
AARS1					
Low	219	41	3,042	1	0.031
High	219	57	2,475	1.583 (1.043–2.401)	—
AGXT2					
Low	219	45	2,475	1	0.372
High	219	53	2,821	1.208 (0.798–1.830)	—
GARS					
Low	219	49	2047	1	0.848
High	219	49	NA	0.961 (0.638–1.447)	—
XPOT					
Low	219	47	2,532	1	0.984
High	219	51	2,821	1.004 (0.666–1.513)	—
NOL4					
Low	219	46	NA	1	0.750
High	219	52	2,475	1.069 (0.710–1.609)	—
PHGDH					
Low	219	43	2,134	1	0.155
High	219	55	2,821	1.353 (0.892–2.054)	—

Notes: Adjusted $P^{\&}$, adjustment for TNM stage; COAD, colon adenocarcinoma.

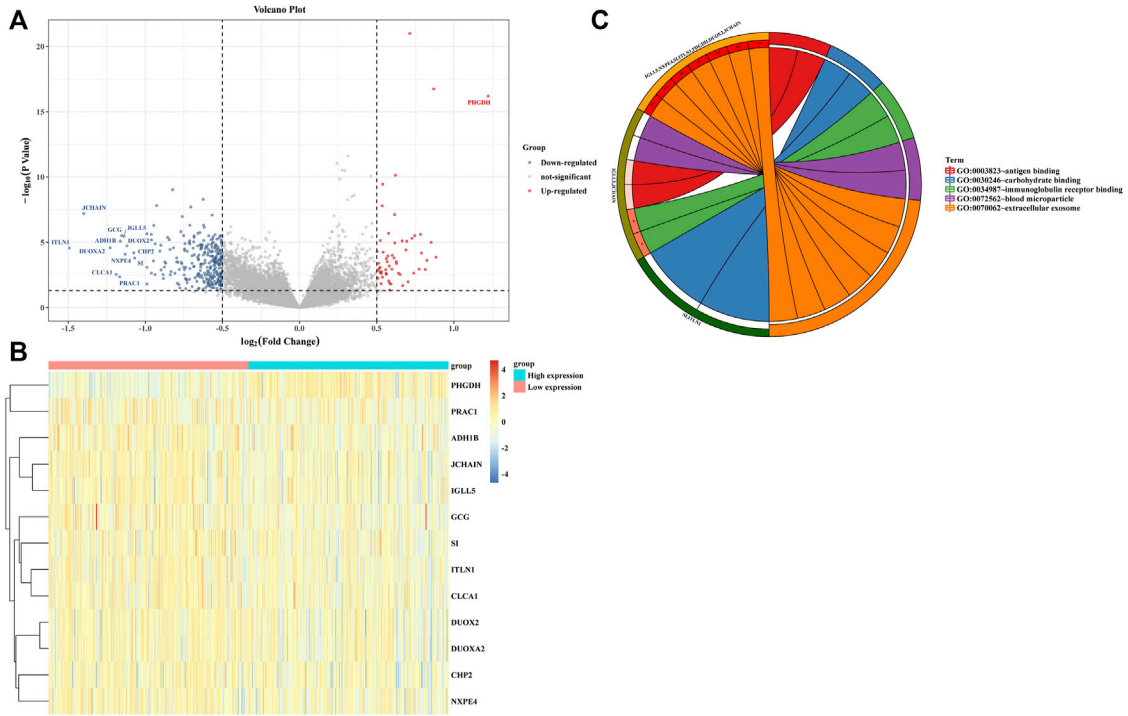


FIGURE 5 | Difference and enrichment analysis of high- and low- expressed *ULBP1* groups in COAD. **(A,B)** differential expression analysis; **(C)** enrichment analysis. Notes: COAD, colon adenocarcinoma; *ULBP1*, unique long 16 (UL16)-binding protein 1.

response, leading to a strong aggressiveness and poor prognosis (Cadoux et al., 2021). These factors vary for different types of cancer. It is also clear that NKG2D ligand can be independently expressed on cancer cells and can be expressed in response to different cancer-related pathways. Such as ULBP1-2, but not ULBP3, is induced by the expression of the BCR/ABL oncogene (McGilvray et al., 2010).

The enriched pathways of *ULBP1* gene and its co-expressed genes showed that co-expressed genes might participate in the metabolic pathway and Aminoacyl-tRNA biosynthesis of COAD. When we selected the top 20 co-expressed genes and performed prognostic analysis, we found that both ARRS1 and DDIT3 genes have prognostic value in COAD. Notably, the combination of High expression of ULBP1, AARS1, and DDIT3 would increase the 2.2-fold death risk of COAD, when compared with those of low expression genes. AARS1 is a family member of the aminoacyl-tRNA synthetases (AARSs), which is a housekeeping protein widely present in all organisms, it can catalyze the combination of amino acids and tRNA and convert nucleic acid coding information into amino acids, playing an important role in protein synthesis (Zhang et al., 2020). In addition to these translation functions, AARSs are also involved in many other important physiological activities, such as translation and transcription regulation, signal transduction, cell migration, angiogenesis, inflammation, and tumorigenesis (Kim et al., 2011; Datt and Sharma, 2014; Kim et al., 2014). Cancer is a disease of cell disorders, which can be affected by using translation in unexpected ways, using the catalytic function of AARSs in an untranslated environment, or manipulating its regulatory function independent of enzyme activity (Wang and Yang, 2020). If the expression of tRNA exceeds a certain level, it may cause abnormal cell and tissue growth. On the other hand, with the strong demand for protein synthesis by cancer, the classic enzyme action of AARSs is needed to maintain tumor growth (Grewal, 2015). DDIT3 gene, also called CHOP, is an endoplasmic reticulum (ER). This gene encodes a member of CCAAT/enhanced binding protein (C/EBP) family transcription factors (Ron and Habener, 1992). DDIT3, activated by p38 mitogen-related protein kinase, is a major pro-apoptotic transcription factor induced by ER stress (Woo et al., 2007). It has been reported that DDIT3 overexpression can lead to cell cycle arrest and/or apoptosis (Woo et al., 2007). Studies have also shown that DDIT3 can trigger key early events leading to cell apoptosis, which is considered an important target for the development of anti-cancer drugs (Oyadomari and Mori, 2004). Additionally, DDIT3 can participate in cell apoptosis transition and induce Bcl2 down-regulation and DR5 (death receptor 5) activated protein (Farooqi et al., 2015). RASK et al. also indicated that increased DDIT3 was associated with the tumor invasion of CRC (Rask et al., 2000). However, Sun et al. activated the PERK-ATF4-CHOP signaling pathway through TIIA, and then increased the expression of ULBP1 and DR5 through

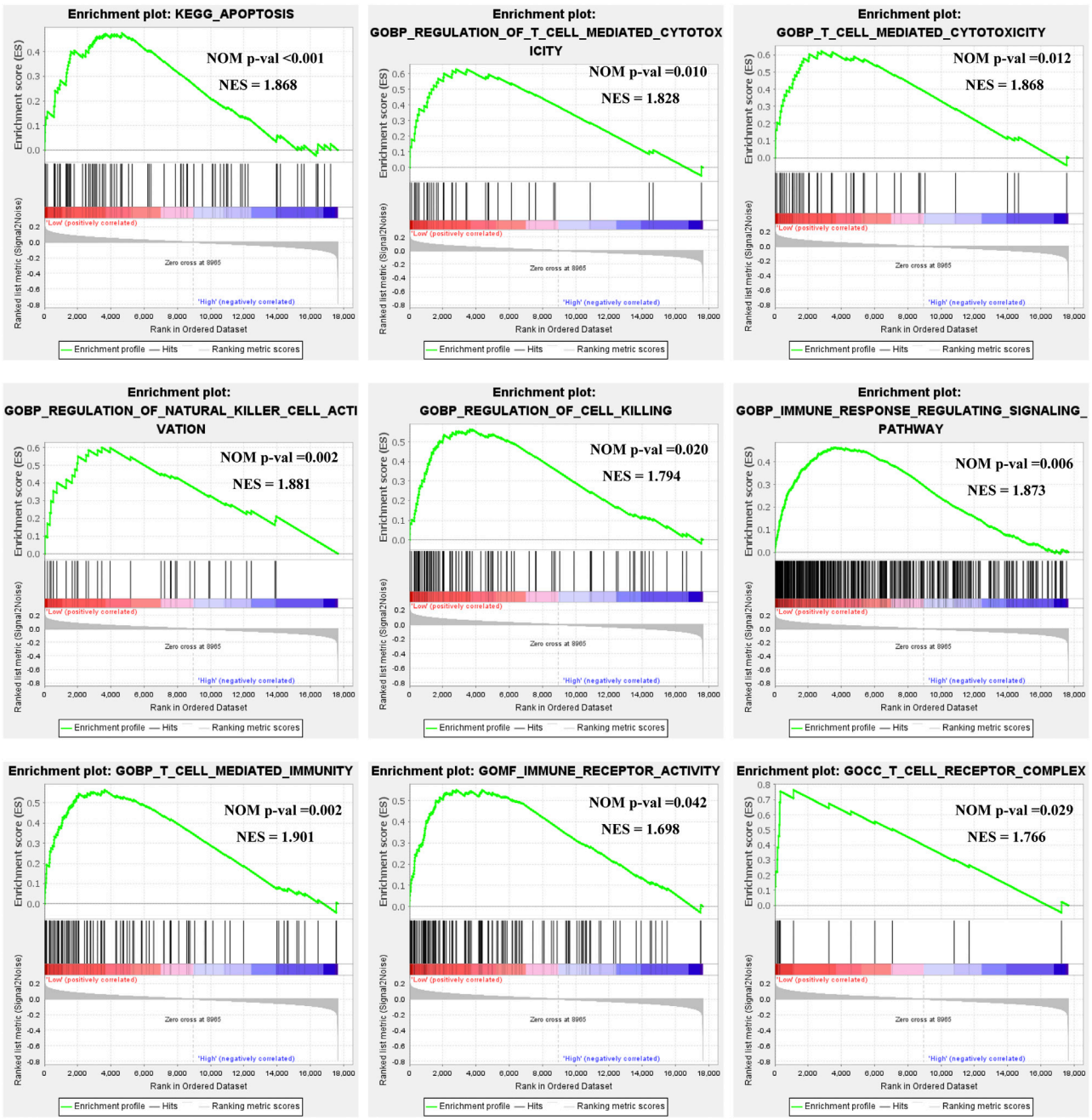


FIGURE 6 | GSEA of *ULBP1* expression in COAD. Notes: COAD, colon adenocarcinoma; ULBP1, unique long 16 (UL16)-binding protein 1; GSEA, gene set enrichment analysis.

ATF4 and CHOP, leading to enhanced NK cell-mediated killing of NSCLC cells, which seemed to indicate a connection between ULBP1 and CHOP(Sun et al., 2021). In general, the combination of these 3 genes that reflect different levels can improve the prognosis of COAD patients.

However, our research still has some unavoidable limitations. Firstly, the study obtained fewer clinical parameters from the TCGA database, and more clinical parameters need to be included to reduce clinical bias. Secondly, we only analyze from the perspective of genes, but due to the limitations of the current experimental conditions, there is no protein-level validation. In the future, more experiments including *in vivo* and *in vitro* are needed to explore. Finally, the Guangxi cohort in this study is only a single-center cohort, and multiple centers and larger samples might be needed for further validation.

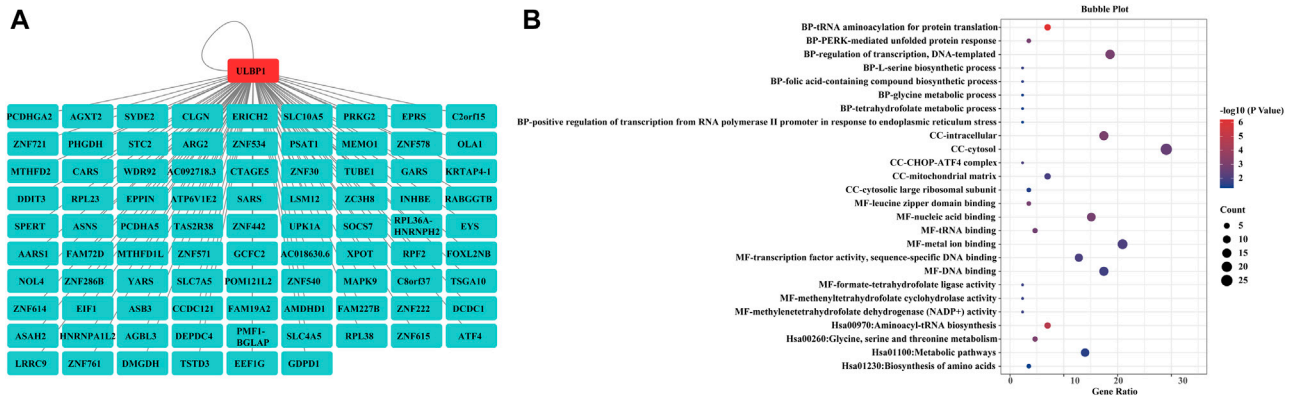


FIGURE 7 | *ULBP1* related co-expressed genes and enrichment analysis in COAD. **(A)** *ULBP1* related co-expressed genes; **(B)** enrichmentanalysis. Notes: COAD, colon adenocarcinoma; *ULBP1*, unique long 16 (UL16)-binding protein 1.

TABLE 3 | Combined effect survival analysis

Gene	Patients	OS			
	(*n* = 438)	HR (95%CI)	Crude *P**	HR (95%CI)	Adjusted *P*[&]
Low *ULBP1*& Low *DDIT3*& Low *AARS1*	100	1	—	1	—
Not all high or low	238	1.681 (0.940–3.004)	0.080	1.424 (0.778–2.606)	0.252
High *ULBP1*& High *DDIT3*& High *AARS1*	100	2.434 (1.325–4.470)	0.004	2.210 (1.180–4.140)	0.013

Notes: Adjusted P[&]*, adjustment for TNM stage; COAD, colon adenocarcinoma.*

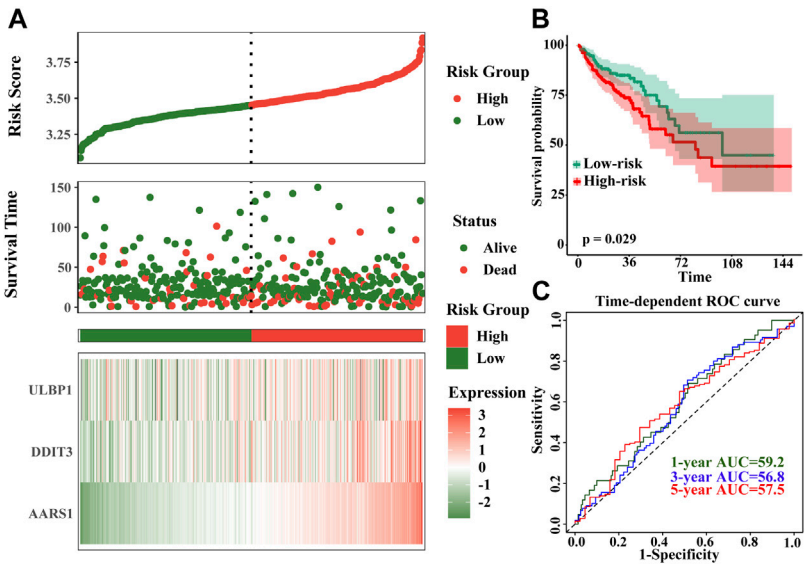


FIGURE 8 | Prognostic risk score model and time-dependent ROC curve of *ULBP1* gene in COAD. **(A)** risk score developed by *ULBP1*, *DDITS*, and *AARS*; **(B)** kaplan–Meier survival curve of risk score; **(C)** time-dependent ROC curve. Notes: COAD, colon adenocarcinoma; ULBP1, unique long 16 (UL16)-bindingprotein 1; *AARS1*, alanyl-tRNA synthetase 1; *DDIT3*, DNA damage inducible transcript 3. ROC, receiver operating characteristic curve.

CONCLUSION

Our study was the first to investigate the diagnostic and prognostic value of the immune-related *ULBP1* gene in COAD. *ULBP1* gene had a high diagnostic value in COAD. Up-regulated *ULBP1* gene of patients with COAD predicted a worse prognosis compared to those patients with down-regulated *ULBP1* gene. GSEA results showed that *ULBP1* was involved in the apoptotic pathway and biological process of T cell mediated cytotoxicity, regulation of natural killer cell activation, and T cell mediated immunity of COAD. The combination survival analysis showed that the combination of high expression of *ULBP1*, *AARS1*, and *DDIT3* would increase the 2.2-fold death risk of COAD when compared with those of low expression genes. However, these findings need to be further validated.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the First Affiliated Hospital of Guangxi Medical University. [Ethics No. 2020 (K-Y-E-050)]. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

G-TR wrote the manuscript. G-TR, H-LX, and L-CZ analyzed and interpreted the patient data, G-TR, H-LX, L-CZ, Y-ZG, and H-PS made substantial contributions to the conception, design, and intellectual content of the studies. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Key Research and Development Program (2017YFC1309200).

ACKNOWLEDGMENTS

We thank the contributor of TCGA (https://cancergenome.nih.gov/) for sharing the COAD dataset on open access. We are grateful to all the participants who have enabled this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.762514/full#supplementary-material

Supplementary Figure S1 | Apoptosis pathway.

Supplementary Figure S2 | Expression of *ULBP1* in READ based on individual cancer stages.

REFERENCES

Butler, J. E., Moore, M. B., Presnell, S. R., Chan, H.-W., Chalupny, N. J., and Lutz, C. T. (2009). Proteasome Regulation of ULBP1 Transcription. *J. Immunol.* 182 (10), 6600–6609. doi:10.4049/jimmunol.0801214

Cadoux, M., Caruso, S., Pham, S., Gougelet, A., Pophillat, C., Riou, R., et al. (2021). Expression of NKG2D Ligands Is Downregulated by β-catenin Signalling and Associates with HCC Aggressiveness. *J. Hepatol.* 74 (6), 1386–1397. doi:10.1016/j.jhep.2021.01.017

Castriconi, R., Cantoni, C., Della Chiesa, M., Vitale, M., Marcenaro, E., Conte, R., et al. (2003). Transforming Growth Factor 1 Inhibits Expression of NKp30 and NKG2D Receptors: Consequences for the NK-Mediated Killing of Dendritic Cells. *Proc. Natl. Acad. Sci.* 100 (7), 4120–4125. doi:10.1073/pnas.0730640100

Champsaur, M., and Lanier, L. L. (2010). Effect of NKG2D Ligand Expression on Host Immune Responses. *Immunol. Rev.* 235 (1), 267–285. doi:10.1111/j.0105-2896.2010.00893.x

Chava, S., Bugide, S., Edwards, Y. J. K., and Gupta, R. (2021). Disruptor of Telomeric Silencing 1-like Promotes Ovarian Cancer Tumor Growth by Stimulating Pro-tumorigenic Metabolic Pathways and Blocking Apoptosis. *Oncogenesis* 10 (7), 48. doi:10.1038/s41389-021-00339-6

Chen, Y., Lin, G., Guo, Z.-q., Zhou, Z.-f., He, Z.-y., and Ye, Y.-b. (2013). Effects of MICA Expression on the Prognosis of Advanced Non-small Cell Lung Cancer and the Efficacy of CIK Therapy. *PLoS One* 8 (7), e69044. doi:10.1371/journal.pone.0069044

Datt, M., and Sharma, A. (2014). Evolutionary and Structural Annotation of Disease-Associated Mutations in Human Aminoacyl-tRNA Synthetases. *BMC Genomics* 15, 1063. doi:10.1186/1471-2164-15-1063

Diefenbach, A., Jensen, E. R., Jamieson, A. M., and Raulet, D. H. (2001). Rae1 and H60 Ligands of the NKG2D Receptor Stimulate Tumour Immunity. *Nature* 413 (6852), 165–171. doi:10.1038/35093109

Dunn, G. P., Old, L. J., and Schreiber, R. D. (2004). The Three Es of Cancer Immunoediting. *Annu. Rev. Immunol.* 22, 329–360. doi:10.1146/annurev.immunol.22.012703.104803

Eagle, R., Jafferji, I., and Barrow, A. (2009). Beyond Stressed Self: Evidence for NKG2D Ligand Expression on Healthy Cells. *Cir* 5 (1), 22–34. doi:10.2174/157339509787314369

Farooqi, A. A., Li, K.-T., Fayyaz, S., Chang, Y.-T., Ismail, M., Liaw, C.-C., et al. (2015). Anticancer Drugs for the Modulation of Endoplasmic Reticulum Stress and Oxidative Stress. *Tumor Biol.* 36 (8), 5743–5752. doi:10.1007/s13277-015-3797-0

Giordano, T. J. (2014). The Cancer Genome Atlas Research Network: a Sight to Behold. *Endocr. Pathol.* 25 (4), 362–365. doi:10.1007/s12022-014-9345-4

Gowen, B. G., Chim, B., Marceau, C. D., Greene, T. T., Burr, P., Gonzalez, J. R., et al. (2015). A Forward Genetic Screen Reveals Novel Independent Regulators of ULBP1, an Activating Ligand for Natural Killer Cells. *Elife* 4, e08474. doi:10.7554/eLife.08474

Grewal, S. S. (2015). Why Should Cancer Biologists Care about tRNAs? tRNA Synthesis, mRNA Translation and the Control of Growth. *Biochim. Biophys. Acta (Bba) - Gene Regul. Mech.* 1849 (7), 898–907. doi:10.1016/j.bbagrm.2014.12.005

Groh, V., Wu, J., Yee, C., and Spies, T. (2002). Tumour-derived Soluble MIC Ligands Impair Expression of NKG2D and T-Cell Activation. *Nature* 419 (6908), 734–738. doi:10.1038/nature01112

Guerra, N., Tan, Y. X., Joncker, N. T., Choy, A., Gallardo, F., Xiong, N., et al. (2008). NKG2D-deficient Mice Are Defective in Tumor Surveillance in Models of Spontaneous Malignancy. *Immunity* 28 (4), 571–580. doi:10.1016/j.immuni.2008.02.016

Harbig, T. A., Nusrat, S., Mazor, T., Wang, Q., Thomson, A., Bitter, H., et al. (2021). OncoThreads: Visualization of Large-Scale Longitudinal Cancer Molecular Data. *Bioinformatics* 37 (Suppl. l_1), i59–i66. doi:10.1093/bioinformatics/btab289

Hutter, C., and Zenklusen, J. C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell* 173 (2), 283–285. doi:10.1016/j.cell.2018.03.042

Ji, Z., Vokes, S. A., Dang, C. V., and Ji, H. (2016). Turning Publicly Available Gene Expression Data into Discoveries Using Gene Set Context Analysis. *Nucleic Acids Res.* 44 (1), e8. doi:10.1093/nar/gkv873

Kim, D., Kwon, N. H., and Kim, S. (2013). Association of Aminoacyl-tRNA Synthetases with Cancer. *Top. Curr. Chem.* 344, 207–245. doi:10.1007/128_2013_455

Kim, S., You, S., and Hwang, D. (2011). Aminoacyl-tRNA Synthetases and Tumorigenesis: More Than Housekeeping. *Nat. Rev. Cancer* 11 (10), 708–718. doi:10.1038/nrc3124

Lee, J.-C., Lee, K.-M., Kim, D.-W., and Heo, D. S. (2004). Elevated TGF-β1 Secretion and Down-Modulation of NKG2D Underlies Impaired NK Cytotoxicity in Cancer Patients. *J. Immunol.* 172 (12), 7335–7340. doi:10.4049/jimmunol.172.12.7335

Li, C.-J., Chiu, Y.-H., Chang, C., Chang, Y.-C. I., Sheu, J. J.-C., and Chiang, A.-J. (2021). Acetyl Coenzyme A Synthase 2 Acts as a Prognostic Biomarker Associated with Immune Infiltration in Cervical Squamous Cell Carcinoma. *Cancers* 13 (13), 3125. doi:10.3390/cancers13133125

Liu, X.-S., Zhou, L.-M., Yuan, L.-L., Gao, Y., Kui, X.-Y., Liu, X.-Y., et al. (2021). NPM1 Is a Prognostic Biomarker Involved in Immune Infiltration of Lung Adenocarcinoma and Associated With m6A Modification and Glycolysis. *Front. Immunol.* 12, 724741. doi:10.3389/fimmu.2021.724741

Maccalli, C., Giannarelli, D., Chiarucci, C., Cutaia, O., Giacobini, G., Hendrickx, W., et al. (2017). Soluble NKG2D Ligands Are Biomarkers Associated with the Clinical Outcome to Immune Checkpoint Blockade Therapy of Metastatic Melanoma Patients. *Oncoimmunology* 6 (7), e1323618. doi:10.1080/2162402X.2017.1323618

McGilvray, R. W., Eagle, R. A., Rolland, P., Jafferji, I., Trowsdale, J., and Durrant, L. G. (2010). ULBP2 and RAET1E NKG2D Ligands Are Independent Predictors of Poor Prognosis in Ovarian Cancer Patients. *Int. J. Cancer* 127 (6), 1412–1420. doi:10.1002/ijc.25156

McGilvray, R. W., Eagle, R. A., Watson, N. F. S., Al-Attar, A., Ball, G., Jafferji, I., et al. (2009). NKG2D Ligand Expression in Human Colorectal Cancer Reveals Associations with Prognosis and Evidence for Immunoediting. *Clin. Cancer Res.* 15 (22), 6993–7002. doi:10.1158/1078-0432.CCR-09-0991

Mondelli, M. U. (2012). NKG2D and its Ligands: Key to Immunotherapy of Liver Cancer? *J. Hepatol.* 56 (2), 308–310. doi:10.1016/j.jhep.2011.07.008

Oppenheim, D. E., Roberts, S. J., Clarke, S. L., Filler, R., Lewis, J. M., Tigelaar, R. E., et al. (2005). Sustained Localized Expression of Ligand for the Activating NKG2D Receptor Impairs Natural Cytotoxicity *In Vivo* and Reduces Tumor Immunosurveillance. *Nat. Immunol.* 6 (9), 928–937. doi:10.1038/ni1239

Oyadomari, S., and Mori, M. (2004). Roles of CHOP/GADD153 in Endoplasmic Reticulum Stress. *Cell Death Differ* 11 (4), 381–389. doi:10.1038/sj.cdd.4401373

Patel, S., Issaka, R. B., Chen, E., and Somsouk, M. (2021). Colorectal Cancer Screening and COVID-19. *Am. J. Gastroenterol.* 116 (2), 433–434. doi:10.14309/ajg.0000000000000970

Rask, K., Thorn, M., Ponten, F., Kraaz, W., Sundfeldt, K., Hedin, L., et al. (2000). Increased Expression of the Transcription Factors CCAAT-Enhancer Binding Protein-? (C/EB?) and C/EBP? (CHOP) Correlate with Invasiveness of Human Colorectal Cancer. *Int. J. Cancer* 86 (3), 337–343. doi:10.1002/(sici)1097-0215(20000501)86:3<337:aid-ijc6>3.0.co;2-3

Ron, D., and Habener, J. F. (1992). CHOP, a Novel Developmentally Regulated Nuclear Protein that Dimerizes with Transcription Factors C/EBP and LAP and Functions as a Dominant-Negative Inhibitor of Gene Transcription. *Genes Dev.* 6 (3), 439–453. doi:10.1101/gad.6.3.439

Ruan, G.-T., Gong, Y.-Z., Zhu, L.-C., Gao, F., Liao, X.-W., Wang, X.-K., et al. (2020a). The Perspective of Diagnostic and Prognostic Values of Lipoxygenases mRNA Expression in Colon Adenocarcinoma. *Ott* Vol. 13, 9389–9405. doi:10.2147/OTT.S251965

Ruan, G. T., Zhu, L. C., Gong, Y. Z., Liao, X. W., Wang, X. K., Liao, C., et al. (2020b). The Diagnosis and Prognosis Values of WNT mRNA Expression in colon Adenocarcinoma. *J. Cel Biochem* 121 (5-6), 3145–3161. doi:10.1002/jcb.29582

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., et al. (2020). Colorectal Cancer Statistics, 2020. *CA A. Cancer J. Clin.* 70 (3), 145–164. doi:10.3322/caac.21601

Smyth, M. J., Swann, J., Cretney, E., Zerafa, N., Yokoyama, W. M., and Hayakawa, Y. (2005). NKG2D Function Protects the Host from Tumor Initiation. *J. Exp. Med.* 202 (5), 583–588. doi:10.1084/jem.20050994

Sun, Y., Gong, C., Ni, Z., Hu, D., Ng, W., Zhu, X., et al. (2021). Tanshinone IIA Enhances Susceptibility of Non-small Cell Lung Cancer Cells to NK Cell-mediated Lysis by Up-regulating ULBP1 and DR5. *J. Leukoc. Biol.* 110 (2), 315–325. doi:10.1002/JLB.5MA1120-776RR

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a Web Server for Cancer and normal Gene Expression Profiling and Interactive Analyses. *Nucleic Acids Res.* 45 (W1), W98–W102. doi:10.1093/nar/gkx247

Ullah, M., Han, K., Hadi, F., Xu, J., Song, J., and Yu, D.-J. (2021). PScL-HDeep: Image-Based Prediction of Protein Subcellular Location in Human Tissue Using Ensemble Learning of Handcrafted and Deep Learned Features with Two-Layer Feature Selection. *Brief Bioinform* 22. doi:10.1093/bib/bbab278

Wang, J., and Yang, X.-L. (2020). Novel Functions of Cytoplasmic Aminoacyl-tRNA Synthetases Shaping the Hallmarks of Cancer. *Enzymes* 48, 397–423. doi:10.1016/bs.enz.2020.06.005

Wang, Y., Nie, H., Liao, Z., He, X., Xu, Z., Zhou, J., et al. (2021). Expression and Clinical Significance of Lactate Dehydrogenase A in Colon Adenocarcinoma. *Front. Oncol.* 11, 700795. doi:10.3389/fonc.2021.700795

Woo, K. J., Lee, T. J., Lee, S. H., Lee, J.-M., Seo, J.-H., Jeong, Y.-J., et al. (2007). Elevated Gadd153/chop Expression during Resveratrol-Induced Apoptosis in Human colon Cancer Cells. *Biochem. Pharmacol.* 73 (1), 68–76. doi:10.1016/j.bcp.2006.09.015

Wu, X., Tao, Y., Hou, J., Meng, X., and Shi, J. (2012). Valproic Acid Upregulates NKG2D Ligand Expression through an ERK-dependent Mechanism and Potentially Enhances NK Cell-Mediated Lysis of Myeloma. *Neoplasia* 14 (12), 1178–1189. doi:10.1593/neo.121236

Zhang, C., Lin, X., Zhao, Q., Wang, Y., Jiang, F., Ji, C., et al. (2020). YARS as an Oncogenic Protein that Promotes Gastric Cancer Progression through Activating PI3K-Akt Signaling. *J. Cancer Res. Clin. Oncol.* 146 (2), 329–342. doi:10.1007/s00432-019-03115-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Ruan, Xie, Zhu, Ge, Yan, Liao, Gong and Shi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

MINI REVIEW
published: 10 February 2022
doi: 10.3389/fgene.2022.760690

Regulation of DNA Methylation During Plant Endosperm Development

Dongdong Lu[1,2,3,4], *Jixian Zhai*[2,3,4]* *and Mengli Xi*[1]*

[1]Key Laboratory of Forest Genetics and Biotechnology of Ministry of Education, Nanjing Forestry University, Nanjing, China, [2]Department of Biology, Southern University of Science and Technology, Shenzhen, China, [3]Institute of Plant and Food Science, Southern University of Science and Technology, Shenzhen, China, [4]Key Laboratory of Molecular Design for Plant Cell Factory of Guangdong Higher Education Institutes, Southern University of Science and Technology, Shenzhen, China

The endosperm is a vital storage tissue in plant seeds. It provides nutrients to the embryos or the seedlings during seed development and germination. Although the genetic information in the endosperm cannot be passed directly to the next generation, its inherited epigenetic marks affect gene expression and its development and, consequently, embryo and seed growth. DNA methylation is a major form of epigenetic modification that can be investigated to understand the epigenome changes during reproductive development. Therefore, it is of great significance to explore the effects of endosperm DNA methylation on crop yield and traits. In this review, we discuss the changes in DNA methylation and the resulting imprinted gene expression levels during plant endosperm development, as well as their effects on seed development.

Keywords: DNA methylation, demethylation, imprinted genes, endosperm, symmetrical methylation, non-symmetrical methylation

OPEN ACCESS

Edited by:
Joao Batista Teixeira da Rocha, Federal University of Santa Maria, Brazil

Reviewed by:
Igor Kovalchuk, University of Lethbridge, Canada
Jordi Moreno-Romero, Universitat Autònoma de Barcelona, Spain
Zongliang Chen, Rutgers, The State University of New Jersey, United States

***Correspondence:**
Jixian Zhai
zhaijx@sustech.edu.cn
Mengli Xi
ximenglinjfu@126.com

Specialty section:
This article was submitted to Epigenomics and Epigenetics, a section of the journal Frontiers in Genetics

Received: *18 August 2021*
Accepted: *17 January 2022*
Published: *10 February 2022*

Citation:
Lu D, Zhai J and Xi M (2022) Regulation of DNA Methylation During Plant Endosperm Development. Front. Genet. 13:760690. doi: 10.3389/fgene.2022.760690

INTRODUCTION

In angiosperms, the development of seeds requires double fertilization. The egg and central cells independently combine with sperm cells and develop into the embryo and endosperm, respectively, (Bleckmann et al., 2014). The functions of endosperm are mainly to act as the nutrient supplier, to be the mechanical barrier for the embryo, and to be the growth regulator of the embryo during seed development and germination. The endosperm is also a fundamental factor for the seed prosperity of angiosperms (Yan et al., 2014). Additionally, the endosperm is a critical factor in controlling seed viability and dormancy (De Giorgi et al., 2015). The vast majorities of the edible parts of rice, wheat, and corn, which account for approximately 70% of human food consumption, are endosperm tissues, which are rich in starch, protein, vitamins, dietary fiber, and other nutrients needed in the human diet (Kang et al., 2008). Therefore, improving endosperm contents and quality is a significant issue. Regulation of endosperm development involves gene imprinting and epigenetic modifications. DNA methylation is a major epigenetic modification that participates in gene expression, TE silencing, and genome stability during plant development. It is related to epigenetic transgenerational inheritance. Owing to the importance of the endosperm, its DNA methylation and genome imprinting are reviewed here.

This review introduces the methylation and demethylation of plant endosperm and the imprinted gene expression resulting from DNA methylation. We discuss the effects of endosperm DNA methylation on plant development.

DNA METHYLATION AND DEMETHYLATION OF THE PLANT ENDOSPERM

DNA methylation, the addition of a methyl (CH_3) group at carbon 5 of cytosine by DNA methyltransferases, is a critical epigenetic marker in mammals and plants (Jin et al., 2011). It represents a heritable change in gene expression not encoded by the DNA sequence. DNA methylation is essential for genomic imprinting, transposable element (TE) silencing, gene regulation, genetic evolution, and genomic stability (Zhang et al., 2018). The loss of DNA methyltransferase function can lead to abnormal plant development (Rajkumar et al., 2020). DNA methylation occurs in three sequence contexts in plants: symmetrical CG and CHG sites and asymmetrical CHH (H = C/T/A) sites (Kawakatsu et al., 2017). Different methyltransferases accomplish different DNA methylation patterns through *de novo* methylation and maintenance of methylation (Law and Jacobsen, 2010). There are three types of DNA methyltransferase in plants: DNA Methyltransferase (MET), Domains Rearranged Methyltransferase (DRM), and plant-specific Chromomethylase (CMT). These methyltransferases perform their duties in *de novo* and maintenance methylation, and jointly complete the DNA methylation modification in plants (Ashapkin et al., 2016).

De novo methylation refers to the generation of new methylation at sites that have not undergone methylation. The plant-specific RNA-directed DNA methylation (RdDM) pathway catalyzes the *de novo* methylation of three sequence contexts (Matzke and Mosher, 2014). Asymmetric CHH site methylation can only be maintained through *de novo* methylation (Zhang et al., 2018). Some repetitive DNA sequences are transcribed by RNA Polymerase IV (Pol IV) to generate single-strand RNAs (ssRNAs). These ssRNAs produce double-strand RNA (dsRNA) by RNA-Dependent RNA Polymerase 2 (RDR2), which is then cut into 24 nt siRNA by DICER-LIKE 3 (DCL3) (Haag and Pikaard, 2011; Matzke and Mosher, 2014; Zhai et al., 2015). In addition, some inverted repeated DNA sequences can also produce dsRNA under the action Pol II and RDR6, which is further cleaved by DCL3 to produce 21 nt siRNA. These 24 nt and 21 nt siRNAs combine with Argonaute 4 (AGO4) to form a siRNA-AGO4 complex, which recruits DRM1 and DRM2 to *de novo* methylation in the three sites (CG, CHG, CHH) (Matzke and Mosher, 2014). In addition to the RdDM pathway, CMT2 and CMT3 can also catalyze *de novo* methylation (Law and Jacobsen, 2010). The nucleosome remodeling factor Decrease in DNA Methylation 1 (DDM1) changes nucleosome conformation, binds CMT2 to histone H3 lysine 9 dimethylation (H3K9me2) and mediates the *de novo* methylation of the adjacent CHG and CHH sites (Kuo et al., 2017).

Maintenance methylation refers to maintaining the methylation form of the original site in the process of DNA replication. The maintenance of CG methylation in plants is completed by the methylation regulator VIMs (Variation in Methylation, VIM1, VIM2, VIM3) protein and DNA methyltransferase MET1. VIMs recognize and bind to the hemimethylated CG site, recruit MET1 to complete CG methylation of the newly synthesized strand, and finally obtain double-stranded DNA methylation of the CG site (Kawashima and Berger, 2014). The maintenance of CHG methylation is mediated by the CMT3-H3K9me2 pathway. CMT3 binds to two H3K9me2 proteins simultaneously and methylates DNA at nearby CHG sites. The methylated CHG DNA recruits Su (var) Homologue 4 (SUVH4), and the deposition of H3K9me2 markers on the nucleosomes surrounding CHG methylated DNA by SUVH4 creates a CHG–H3K9me2 positive feedback loop (Zhang et al., 2018).

In addition, there is an active DNA demethylation process in flowering plants, and this is achieved by DNA glycosylase/lyase through a base excision repair (BER) mechanism. Three types of DNA glycosylases have been found in plants: Demeter (DME), Repressor of silencing1 (ROS1), Demeter-like (DML2 and DML3). Thus, the final methylation level in the genome is determined by the activities of both DNA methyltransferases and demethylases (Gong et al., 2002).

In *Arabidopsis*, the expression levels of the major DNA methylation enzymes are available at the *Arabidopsis RNA-seq database* (http://ipf.sustech.edu.cn/pub/athrna/). In wild-type endosperm, the expression of *MET1* is low, whereas the expression levels of *MET2a* and *MET2b*, which are specifically expressed in central cells, are high. The paternal imprinting genes *VIM5* and *MET3* are also specifically expressed and highly expressed in the endosperm. Therefore, we speculate that *MET2a*, *MET2b*, *MET3*, and *VIM5* may jointly regulate CG methylation in the endosperm (**Figure 1**), which requires further experimental proof. We profile a simple model based on DME-mediated DNA demethylation in the endosperm (**Figure 2**). The *DME* gene is predominantly expressed in the central cell, and DME induces global hypomethylation (Choi et al., 2002; Hsieh et al., 2009). Before fertilization, the central cell and vegetative cell are highly demethylated resulted from the action of DME. DME preferentially targets TE regions (Hsieh et al., 2009; Ibarra et al., 2012). The demethylation of the maternal genome during gametogenesis is also reported in other species—castor bean (Park et al., 2016), rice (Zemach et al., 2010; Park et al., 2016), and maize (Lauria et al., 2004). The vegetative cell produces siRNA into the sperm cells and maintains the sperm cell hypermethylation through the RDdM pathway (Martinez et al., 2016). So the methylation level in the endosperm is much lower than in the embryo after fertilization. The siRNAs produced by the demethylation of the endosperm are transferred to the embryo to maintain the stability of the embryo genome (McCue et al., 2012). The loss of *DME* function (*dme* mutant) in endosperm restores CG methylation but unexpectedly further diminishes non-CG methylation, suggesting demethylation in a non-CG context is regulated by a yet unknown DME-independent mechanism (Hsieh et al., 2009; Jullien et al., 2012).

THE LEVEL OF DNA METHYLATION IN ENDOSPERM VARIES AMONG DIFFERENT PLANT SPECIES AND DURING THEIR DEVELOPMENT

The endosperm of some plants, such as *Arabidopsis*, only exist in the early stage of seed development and gradually disappear with

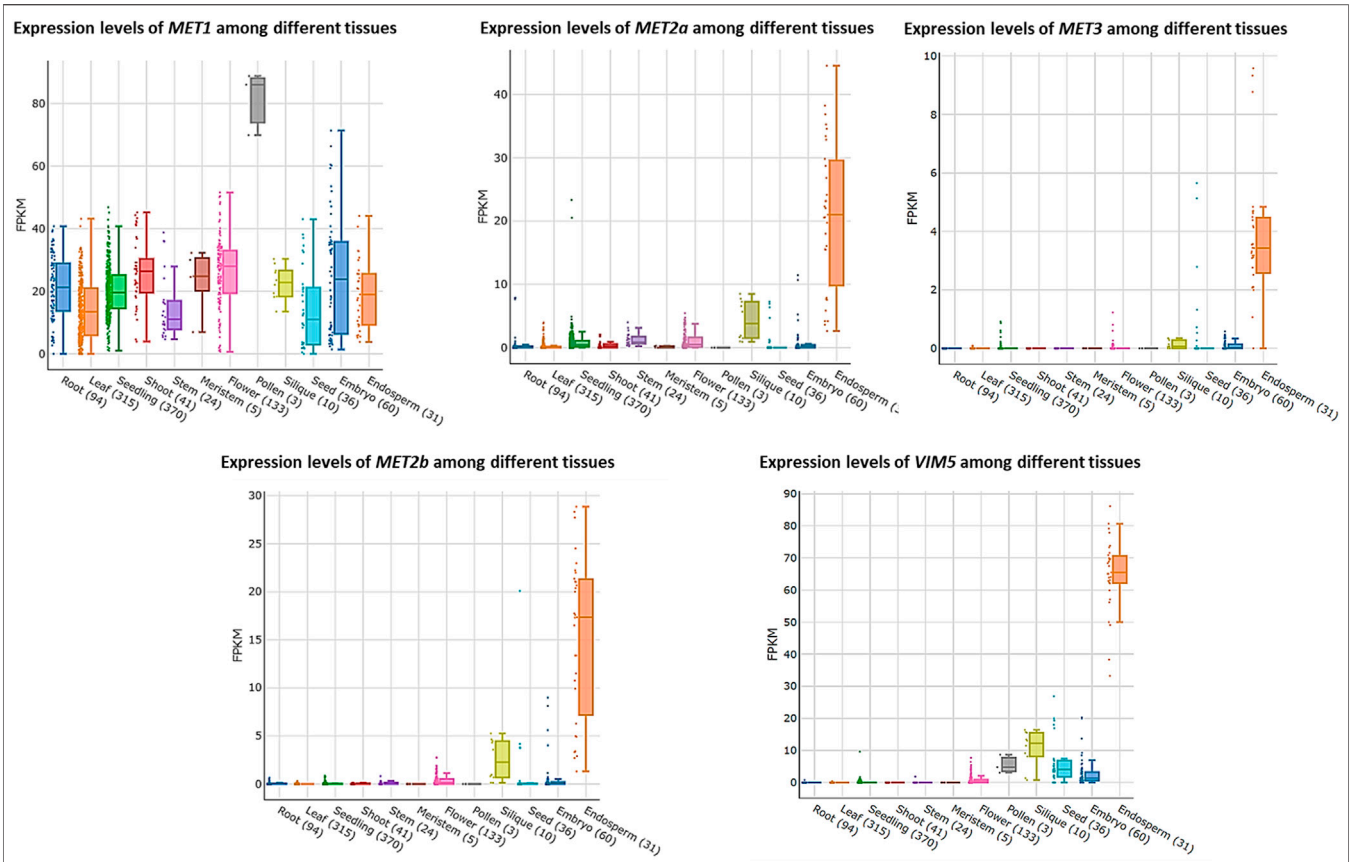


FIGURE 1 | Expression levels of *MET1*, *MET2a*, *MET2b*, *MET3* and *VIM5* among different tissues.

seed development (Brown et al., 1999). Most monocotyledons, some dicotyledons, and gymnosperms, have endosperm in their mature seeds, such as castor beans and rice (Greenwood and Bewley, 1982; Brown et al., 1996). The DNA methylation profiles in the endosperm of different plants are significantly different, suggesting that DNA methylation profiles of endosperm are not conserved. By comparing the methylation levels of the endosperm among different plants, it was found that genomic DNA hypomethylation in endosperm relative to the embryo is widespread (**Figure 3**), especially in dicotyledons (Hsieh et al., 2009; Zemach et al., 2010; Lu et al., 2015; Xu et al., 2016). CG, CHG, and CHH methylation levels were low at 4 days after pollination (DAP), but all three contexts of DNA methylation levels were elevated at 6 days after pollination by DNA methylation sequencing in *Arabidopsis* endosperm (Pignatta et al., 2014; Moreno-Romero et al., 2016). Hu et al. found that the methylation levels were higher during the early (3–5 DAP) and late stages (13–25 DAP) of endosperm development compared with the middle stage (7–11 DAP) in maize endosperm (Hu et al., 2021). Thus, DNA methylation represents a dynamic process during endosperm development. The DNA methylation changes in the endosperm affect the expression of genes and siRNAs, thereby affecting endosperm formation and seed development (Moore et al., 2013).

GENOMIC IMPRINTING BY DNA METHYLATION DURING PLANT ENDOSPERM DEVELOPMENT

Genomic imprinting is the process of inheriting the epigenetic marking for a particular segment of a chromosome from paternal or maternal alleles (Feil and Berger, 2007). The endosperm is the main organ that undergoes genomic imprinting in flowering plants (Gehring et al., 2011; Hsieh et al., 2011). The epigenetic regulation of genomic imprinting plays an indispensable role in normal endosperm development and seed fertility. The misregulation of imprinted genes affects the sizes of seeds or leads to inviable seeds (Tiwari et al., 2010; Hornslien et al., 2019). The generation of genomic imprinting is mainly caused by the different epigenetic modifications of male and female gametes before fertilization (Batista and Köhler, 2020). When the central cell and sperm cell fuses to form the primordial endosperm nucleus, the differences in epigenetic modification between the male and female genomes result in only one allele being expressed and the other being silenced. The differential loss of DNA methylation in the paternal and maternal alleles produces different chromatin marks in *Arabidopsis*. For example, the methylated paternal allele can lead to being transcriptionally silent, while the

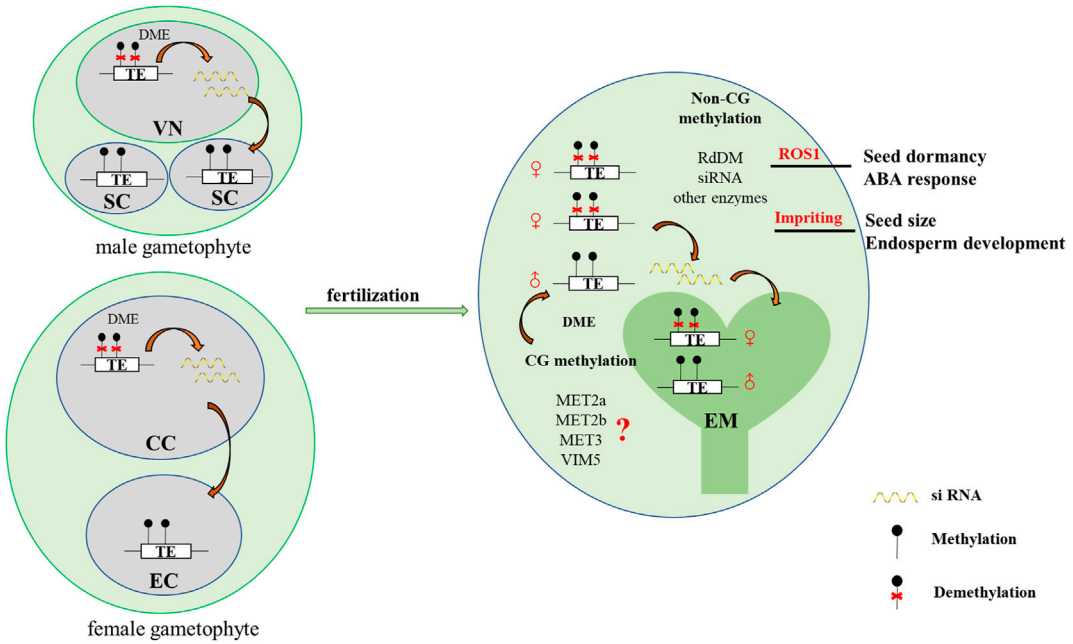


FIGURE 2 | A simplified model of DME-mediated DNA demethylation in *Arabidopsis* endosperm. Before fertilization, the central cell and vegetative cell are highly demethylated resulted from the action of DME. DME preferentially targets TE regions (Gehring et al., 2009; Hsieh et al., 2009; Ibarra et al., 2012). The vegetative cell produces siRNA into the sperm cells and maintains the sperm cell hypermethylation through the RDdM pathway (Martinez et al., 2016). So the methylation level in the endosperm is much lower than in the embryo after fertilization. The siRNAs produced by the demethylation of the endosperm are transferred to the embryo to maintain the stability of the embryo genome. MET2a, MET2b, MET3, and VIM5 may jointly regulate CG methylation in the endosperm. In addition, the differential methylation of the embryo and endosperm leads to imprinting in the endosperm, which may affect endosperm development and control seed size. Additionally, the endosperm demethylase ROS1 regulates seed dormancy. VN: vegetative cell nucleus, SC: sperm cell, CC: central cell, EC: egg cell, EM: embryo.

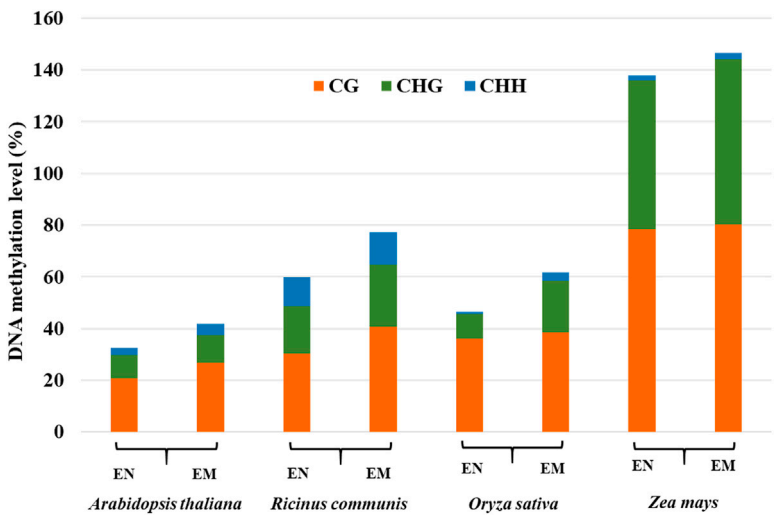


FIGURE 3 | DNA methylation levels of CG, CHG, and CHH in the endosperm (EN) and embryo (EM) among different species.

demethylation of maternal allele would become a transcriptionally active state (Kinoshita et al., 2004; Jullien et al., 2006; Tiwari et al., 2008). In *Arabidopsis*, the expression level of the DNA methyltransferase gene *MET1* is low in central cells, whereas the expression level of the demethylase gene *DME* is high (Huh et al., 2008). Therefore, the central cells maintain a lower DNA methylation level, but the sperm cells maintain a higher

DNA methylation level because *DME* is not expressed (Huh et al., 2008).

RdDM is also critical for silencing of the paternal allele at MEG (maternally expressed imprinted genes) loci; Vu et al. used Col and Cvi to distinguish the parental alleles of *Suppressor OF drm1 drm2 cmt3 (SDC)* and *MOP9.5* (also called *AtPI4Kγ3*, a type II phosphoinositide 4-kinase), and crossed wild-type ovules with pollen from mutants for RdDM (such as *nrpd2a* mutant, NRPD2A is the second largest subunit of RNA pol IV and pol V); they observed activation of *SDC* and *MOP9.5* paternal alleles from *nrpd2a* homozygous plants. Further research found that maternal-specific expression of imprinted genes *SDC* and *MOP9.5* was maintained by MET1. These results suggest that small RNAs have a significant role in setting MEG expression patterns (Vu et al., 2013). PEGs (paternally expressed imprinted genes) can also be hypomethylated at the maternal allele and hypermethylated at the paternal allele (Hsieh et al., 2009; Zhang et al., 2014). So the maternal hypomethylation is essential for the silencing of the maternal allele for many PEGs (Hsieh et al., 2011; Wolff et al., 2011). MEGs are generally more affected by DNA methylation than PEGs (Chen et al., 2018), but the latter is also regulated by histone modification, such as H3K27me3 (Wolff et al., 2011; Zhang et al., 2014). Two other repressive epigenetic marks, H3K9me2 and CHG methylation, also contribute to maternal alleles silencing of PEGs, leading to differential expression of parent-of-origin alleles in the endosperm (Inoue et al., 2017; Moreno-Romero et al., 2019). Silencing of the maternal *PHERES 1* (*PHE1*, a paternally expressed imprinted transcription factor gene) allele depends on the Polycomb Repressive Complex 2 (PRC2), and maternally inherited mutations that encode PRC2 proteins cause biallelic expression of *PHE1* (Kohler et al., 2005). The differences in the expression of PRC2 between sperm and central cells resulted in different histone methylation modifications of parental genomes in the endosperm (Luo et al., 2000; Schoft et al., 2011). DNA methylation can prevent H3K27me3 modification and interfere with PRC2 function (Weinhofer et al., 2010; Deleris et al., 2012; Jermann et al., 2014).

Genomic imprinting disruption accompanies endosperm abortion, and the expression of many imprinted genes also changes (Jullien and Berger, 2010; Kradolfer et al., 2013; Florez-Rueda et al., 2016; Tonosaki et al., 2018). Many MEGs affect seed development by regulating endosperm cytogenesis (Niu et al., 2020; Cheng et al., 2021; Tonosaki et al., 2021); whereas most PEGs knock-out mutations generally do not affect normal plant growth and development in *Arabidopsis*. But PEGs are important for endosperm development in plants, several *peg* mutants: such as *adm* (*ADMETOS*) and *peg2* (*At1g49290*) mutants—can rescue triploid seed abortion (Wolff et al., 2015). And the loss of some PEGs can also lead to serious phenotypic defects. For example, the mutants of *PEG1* (*Os01g08570*, encoding an oxygenase dependent on ketoglutarate and iron), *PEG2* (*OsFBX365*, encoding an F-box domain protein), and *PEG3* (*OsFBDUF48*, encoding a DUF295-domain protein) in rice can reduce starch content and seed fertility (Yuan et al., 2017). The PEGs may be directly involved in regulating reproductive isolation between species. In the endosperm of distant *Arabidopsis* inter-accession crosses (such as Columbia × Nossen), the expression disorder of PEGs is more significant than that of MEGs (Wolff et al., 2015). In interploidy crosses, some PEGs mutants rescue seed abortion, so they have a dramatically different phenotype than WT (Kradolfer et al., 2013; Wang et al., 2018). Hundreds of possible imprinted genes have been discovered in plants. However, there is still a lack of in-depth research on the biological functions of plant imprinted genes, even though many imprinting genes co-localize with yield-related traits (Yuan et al., 2017). For example, Chen et al. found that the rice grain weight QTL—*Grain Weight 2*—mainly expressed maternal alleles in the endosperm (Chen et al., 2016; Niu et al., 2020). These studies indicate that both MEGs and PEGs can participate in plant endosperm development.

DISCUSSION

The DNA methylation of endosperm plays a vital role in regulating seed development and storage material biosynthesis. The removal of imprinted genes can affect endosperm development and lead to seed abortion. In addition, DNA methylation can also regulate endosperm development by regulating the expression of genes and small RNAs. For example, DNA methylation affects starch synthesis in maize endosperm (Hu et al., 2021). DNA methylation also regulates seed size (Rajkumar et al., 2020) and dormancy (Zhu et al., 2018), and it directly affects crop yield and quality. At present, the research on most crops is limited to the regulation of transcription factors, and the research on DNA methylation mainly focuses on model organisms. Although the methylation sequencing of plant endosperm is gradually increasing, the regulatory pathways related to DNA methylation and demethylation in the endosperm are unclear. Therefore, it is recommended to use a combination of methylation sequencing and RNA sequencing (RNA sequencing, single-cell sequencing, small RNA sequencing) to study plant endosperm and establish a complete regulatory network profile. It is of great value to identify the cellular heterogeneity of methylation in plants, but it is still extremely challenging to sequence single-cell DNA methylation in plant endosperm. On the one hand, the presence of seed coat makes it difficult to separate pollution-free endosperm. On the other hand, it is difficult to use bisulfite-transformed DNA fragments by library construction and sequencing for highly methylated and highly repetitive genomes. The regulation of methylation in plant endosperm should be the focus of future research.

AUTHOR CONTRIBUTIONS

DL: Methodology, Investigation, Writing—original draft, Writing—review and editing. JZ: Conceptualization,

Methodology, Resources, Funding. MX: Conceptualization, Methodology, Supervision, Writing—review.

FUNDING

This work is supported by the National Key R&D Program of China Grant (2019YFA0903903), and the Key Laboratory of Molecular Design for Plant Cell Factory of Guangdong Higher Education Institutes (2019KSYS006).

ACKNOWLEDGMENTS

We thank International Science Editing (http://www.internationalscienceediting.com) for editing this manuscript.

REFERENCES

Ashapkin, V. V., Kutueva, L. I., and Vanyushin, B. F. (2016). Plant DNA Methyltransferase Genes: Multiplicity, Expression, Methylation Patterns. *Biochem. Mosc.* 81 (2), 141–151. doi:10.1134/S0006297916020085

Batista, R. A., and Köhler, C. (2020). Genomic Imprinting in Plants-Revisiting Existing Models. *Genes Dev.* 34, 24–36. doi:10.1101/gad.33292410.1101/gad.332924.119

Bleckmann, A., Alter, S., and Dresselhaus, T. (2014). The Beginning of a Seed: Regulatory Mechanisms of Double Fertilization. *Front. Plant Sci.* 5, 452. doi:10.3389/fpls.2014.00452

Brown, R. C., Lemmon, B. E., Nguyen, H., and Olsen, O.-A. (1999). Development of Endosperm in *Arabidopsis thaliana*. *Sex. Plant Reprod.* 12 (1), 32–42. doi:10.1007/s004970050169

Brown, R. C., Lemmon, B. E., and Olsen, O.-A. (1996). Development of the Endosperm in rice (*Oryza Sativa L.*): Cellularization. *J. Plant Res.* 109 (3), 301–313. doi:10.1007/BF02344477

Chen, C., Begcy, K., Liu, K., Folsom, J. J., Wang, Z., Zhang, C., et al. (2016). Heat Stress Yields a Unique MADS Box Transcription Factor in Determining Seed Size and Thermal Sensitivity. *Plant Physiol.* 171 (1), 606–622. doi:10.1104/pp.15.01992

Chen, C., Li, T., Zhu, S., Liu, Z., Shi, Z., Zheng, X., et al. (2018). Characterization of Imprinted Genes in Rice Reveals Conservation of Regulation and Imprinting with Other Plant Species. *Plant Physiol.* 177 (4), 1754–1771. doi:10.1104/pp.17.01621

Cheng, X., Pan, M., E, Z., Zhou, Y., Niu, B., and Chen, C. (2021). The Maternally Expressed Polycomb Group Gene OsEMF2a Is Essential for Endosperm Cellularization and Imprinting in rice. *Plant Commun.* 2 (1), 100092. doi:10.1016/j.xplc.2020.100092

Choi, Y., Gehring, M., Johnson, L., Hannon, M., Harada, J. J., Goldberg, R. B., et al. (2002). DEMETER, a DNA Glycosylase Domain Protein, Is Required for Endosperm Gene Imprinting and Seed Viability in Arabidopsis. *Cell.* 110 (1), 33–42. doi:10.1016/s0092-8674(02)00807-3

De Giorgi, J., Piskurewicz, U., Loubery, S., Utz-Pugin, A., Bailly, C., Mène-Saffrané, L., et al. (2015). An Endosperm-Associated Cuticle Is Required for Arabidopsis Seed Viability, Dormancy and Early Control of Germination. *Plos Genet.* 11 (12), e1005708. doi:10.1371/journal.pgen.1005708

Deleris, A., Stroud, H., Bernatavichute, Y., Johnson, E., Klein, G., Schubert, D., et al. (2012). Loss of the DNA Methyltransferase MET1 Induces H3K9 Hypermethylation at PcG Target Genes and Redistribution of H3K27 Trimethylation to Transposons in *Arabidopsis thaliana*. *Plos Genet.* 8 (11), e1003062. doi:10.1371/journal.pgen.1003062

Feil, R., and Berger, F. (2007). Convergent Evolution of Genomic Imprinting in Plants and Mammals. *Trends Genet.* 23 (4), 192–199. doi:10.1016/j.tig.2007.02.004

Florez-Rueda, A. M., Paris, M., Schmidt, A., Widmer, A., Grossniklaus, U., and Städler, T. (2016). Genomic Imprinting in the Endosperm Is Systematically Perturbed in Abortive Hybrid Tomato Seeds. *Mol. Biol. Evol.* 33 (11), 2935–2946. doi:10.1093/molbev/msw175

Gehring, M., Bubb, K. L., and Henikoff, S. (2009). Extensive Demethylation of Repetitive Elements during Seed Development Underlies Gene Imprinting. *Science.* 324 (5933), 1447–1451. doi:10.1126/science.1171609

Gehring, M., Missirian, V., and Henikoff, S. (2011). Genomic Analysis of Parent-Of-Origin Allelic Expression in *Arabidopsis thaliana* Seeds. *Plos One.* 6 (8), e23687. doi:10.1371/journal.pone.0023687

Gong, Z., Morales-Ruiz, T., Ariza, R. R., Roldán-Arjona, T., David, L., and Zhu, J.-K. (2002). ROS1, a Repressor of Transcriptional Gene Silencing in Arabidopsis, Encodes a DNA Glycosylase/Lyase. *Cell.* 111 (6), 803–814. doi:10.1016/S0092-8674(02)01133-9

Greenwood, J. S., and Bewley, J. D. (1982). Seed Development in *Ricinus communis* (castor Bean). I. Descriptive Morphology. *Can. J. Bot.* 60 (9), 1751–1760. doi:10.1139/b82-222

Haag, J. R., and Pikaard, C. S. (2011). Multisubunit RNA Polymerases IV and V: Purveyors of Non-coding RNA for Plant Gene Silencing. *Nat. Rev. Mol. Cell Biol.* 12 (8), 483–492. doi:10.1038/nrm3152

Hornslien, K. S., Miller, J. R., and Grini, P. E. (2019). Regulation of Parent-Of-Origin Allelic Expression in the Endosperm. *Plant Physiol.* 180 (3), 1498–1519. doi:10.1104/pp.19.00320

Hsieh, T.-F., Ibarra, C. A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R. L., et al. (2009). Genome-Wide Demethylation of Arabidopsis Endosperm. *Science.* 324 (5933), 1451–1454. doi:10.1126/science.1172417

Hsieh, T.-F., Shin, J., Uzawa, R., Silva, P., Cohen, S., Bauer, M. J., et al. (2011). Regulation of Imprinted Gene Expression in Arabidopsisendosperm. *Proc. Natl. Acad. Sci. USA.* 108 (5), 1755–1762. doi:10.1073/pnas.1019273108

Hu, Y., Li, Y., Weng, J., Liu, H., Yu, G., Liu, Y., et al. (2021). Coordinated Regulation of Starch Synthesis in Maize Endosperm by microRNAs and DNA Methylation. *Plant J.* 105 (1), 108–123. doi:10.1111/tpj.15043

Huh, J. H., Bauer, M. J., Hsieh, T.-F., and Fischer, R. L. (2008). Cellular Programming of Plant Gene Imprinting. *Cell.* 132 (5), 735–744. doi:10.1016/j.cell.2008.02.018

Ibarra, C. A., Feng, X., Schoft, V. K., Hsieh, T.-F., Uzawa, R., Rodrigues, J. A., et al. (2012). Active DNA Demethylation in Plant Companion Cells Reinforces Transposon Methylation in Gametes. *Science.* 337 (6100), 1360–1364. doi:10.1126/science.1224839

Inoue, A., Jiang, L., Lu, F., Suzuki, T., and Zhang, Y. (2017). Maternal H3K27me3 Controls DNA Methylation-Independent Imprinting. *Nature.* 547 (7664), 419–424. doi:10.1038/nature23262

Jermann, P., Hoerner, L., Burger, L., and Schubeler, D. (2014). Short Sequences Can Efficiently Recruit Histone H3 Lysine 27 Trimethylation in the Absence of Enhancer Activity and DNA Methylation. *Proc. Natl. Acad. Sci.* 111 (33), E3415–E3421. doi:10.1073/pnas.1400672111

Jin, B., Li, Y., and Robertson, K. D. (2011). DNA Methylation: superior or Subordinate in the Epigenetic Hierarchy? *Genes & Cancer.* 2 (6), 607–617. doi:10.1177/1947601910393957

Jullien, P. E., and Berger, F. (2010). Parental Genome Dosage Imbalance Deregulates Imprinting in Arabidopsis. *Plos Genet.* 6 (3), e1000885. doi:10.1371/journal.pgen.1000885

Jullien, P. E., Katz, A., Oliva, M., Ohad, N., and Berger, F. (2006). Polycomb Group Complexes Self-Regulate Imprinting of the Polycomb Group Gene MEDEA in Arabidopsis. *Curr. Biol.* 16 (5), 486–492. doi:10.1016/j.cub.2006.01.020

Jullien, P. E., Susaki, D., Yelagandula, R., Higashiyama, T., and Berger, F. (2012). DNA Methylation Dynamics During Sexual Reproduction in *Arabidopsis thaliana*. *Curr. Biol.* 22 (19), 1825–1830. doi:10.1016/j.cub.2012.07.061

Kang, I.-H., Steffen, J. G., Portereiko, M. F., Lloyd, A., and Drews, G. N. (2008). The AGL62 MADS Domain Protein Regulates Cellularization During Endosperm Development inArabidopsis. *Plant Cell.* 20 (3), 635–647. doi:10.1105/tpc.107.055137

Kawakatsu, T., Nery, J. R., Castanon, R., and Ecker, J. R. (2017). Dynamic DNA Methylation Reconfiguration During Seed Development and Germination. *Genome Biol.* 18 (1), 171. doi:10.1186/s13059-017-1251-x

Kawashima, T., and Berger, F. (2014). Epigenetic Reprogramming in Plant Sexual Reproduction. *Nat. Rev. Genet.* 15 (9), 613–624. doi:10.1038/nrg3685

Kinoshita, T., Miura, A., Choi, Y., Kinoshita, Y., Cao, X., Jacobsen, S. E., et al. (2004). One-Way Control of FWA Imprinting in Arabidopsis Endosperm by DNA Methylation. *Science.* 303 (5657), 521–523. doi:10.1126/science.1089835

Köhler, C., Page, D. R., Gagliardini, V., and Grossniklaus, U. (2005). The *Arabidopsis thaliana* MEDEA Polycomb Group Protein Controls Expression of PHERES1 by Parental Imprinting. *Nat. Genet.* 37 (1), 28–30. doi:10.1038/ng1495

Kradolfer, D., Wolff, P., Jiang, H., Siretskiy, A., and Köhler, C. (2013). An Imprinted Gene Underlies Postzygotic Reproductive Isolation in *Arabidopsis thaliana*. *Developmental Cell.* 26 (5), 525–535. doi:10.1016/j.devcel.2013.08.006

Kuo, H. Y., Jacobsen, E. L., Long, Y., Chen, X., and Zhai, J. (2017). Characteristics and Processing of Pol IV-dependent Transcripts in Arabidopsis. *J. Genet. Genomics.* 44 (1), 3–6. doi:10.1016/j.jgg.2016.10.009

Lauria, M., Rupe, M., Guo, M., Kranz, E., Pirona, R., Viotti, A., et al. (2004). Extensive Maternal DNA Hypomethylation in the Endosperm of Zea Mays. *Plant Cell.* 16 (2), 510–522. doi:10.1105/tpc.017780

Law, J. A., and Jacobsen, S. E. (2010). Establishing, Maintaining and Modifying DNA Methylation Patterns in Plants and Animals. *Nat. Rev. Genet.* 11 (3), 204–220. doi:10.1038/nrg2719

Lu, X., Wang, W., Ren, W., Chai, Z., Guo, W., Chen, R., et al. (2015). Genome-Wide Epigenetic Regulation of Gene Transcription in Maize Seeds. *PLoS One.* 10 (10), e0139582. doi:10.1371/journal.pone.0139582

Luo, M., Bilodeau, P., Dennis, E. S., Peacock, W. J., and Chaudhury, A. (2000). Expression and Parent-Of-Origin Effects for FIS2, MEA, and FIE in the Endosperm and Embryo of Developing Arabidopsis Seeds. *Proc. Natl. Acad. Sci.* 97 (19), 10637–10642. doi:10.1073/pnas.170292997

Martínez, G., Panda, K., Köhler, C., and Slotkin, R. K. (2016). Silencing in Sperm Cells Is Directed by RNA Movement from the Surrounding Nurse Cell. *Nat. Plants.* 2, 16030. doi:10.1038/nplants.2016.30

Matzke, M. A., and Mosher, R. A. (2014). RNA-Directed DNA Methylation: an Epigenetic Pathway of Increasing Complexity. *Nat. Rev. Genet.* 15 (6), 394–408. doi:10.1038/nrg3683

McCue, A. D., Nuthikattu, S., Reeder, S. H., and Slotkin, R. K. (2012). Gene Expression and Stress Response Mediated by the Epigenetic Regulation of a Transposable Element Small RNA. *Plos Genet.* 8 (2), e1002474. doi:10.1371/journal.pgen.1002474

Moore, L. D., Le, T., and Fan, G. (2013). DNA Methylation and its Basic Function. *Neuropsychopharmacol.* 38 (1), 23–38. doi:10.1038/npp.2012.112

Moreno-Romero, J., Jiang, H., Santos-González, J., and Köhler, C. (2016). Parental Epigenetic Asymmetry of PRC 2-Mediated Histone Modifications in the Arabidopsis Endosperm. *EMBO J.* 35 (12), 1298–1311. doi:10.15252/embj.201593534

Moreno-Romero, J., Del Toro-De León, G., Yadav, V. K., Santos-González, J., and Köhler, C. (2019). Epigenetic Signatures Associated with Imprinted Paternally Expressed Genes in the Arabidopsis Endosperm. *Genome Biol.* 20 (1), 41. doi:10.1186/s13059-019-1652-0

Niu, B., Deng, H., Li, T., Sharma, S., Yun, Q., Li, Q., et al. (2020). OsbZIP76 Interacts With OsNF-YBs and Regulates Endosperm Cellularization in rice (Oryza Sativa). *J. Integr. Plant Biol.* 62 (12), 1983–1996. doi:10.1111/jipb.12989

Park, K., Kim, M. Y., Vickers, M., Park, J.-S., Hyun, Y., Okamoto, T., et al. (2016). DNA Demethylation Is Initiated in the Central Cells of Arabidopsis and rice. *Proc. Natl. Acad. Sci. USA.* 113 (52), 15138–15143. doi:10.1073/pnas.1619047114

Pignatta, D., Erdmann, R. M., Scheer, E., Picard, C. L., Bell, G. W., and Gehring, M. (2014). Natural Epigenetic Polymorphisms lead to Intraspecific Variation in Arabidopsis Gene Imprinting. *Elife.* 3, e03198. doi:10.7554/eLife.03198

Rajkumar, M. S., Gupta, K., Khemka, N. K., Garg, R., and Jain, M. (2020). DNA Methylation Reprogramming During Seed Development and its Functional Relevance in Seed Size/Weight Determination in Chickpea. *Commun. Biol.* 3 (1), 340. doi:10.1038/s42003-020-1059-1

Schoft, V. K., Chumak, N., Choi, Y., Hannon, M., Garcia-Aguilar, M., Machlicova, A., et al. (2011). Function of the DEMETER DNA Glycosylase in the *Arabidopsis thaliana* Male Gametophyte. *Proc. Natl. Acad. Sci.* 108 (19), 8042–8047. doi:10.1073/pnas.1105117108

Tiwari, S., Schulz, R., Ikeda, Y., Dytham, L., Bravo, J., Mathers, L., et al. (2008). Maternally Expressed Pab C-Terminal, a Novel Imprinted Gene Inarabidopsis, Encodes the Conserved C-Terminal Domain of Polyadenylate Binding Proteins. *Plant Cell.* 20 (9), 2387–2398. doi:10.1105/tpc.108.061929

Tiwari, S., Spielman, M., Schulz, R., Oakey, R. J., Kelsey, G., Salazar, A., et al. (2010). Transcriptional Profiles Underlying Parent-Of-Origin Effects in Seeds of *Arabidopsis thaliana*. *BMC Plant Biol.* 10 (1), 72. doi:10.1186/1471-2229-10-72

Tonosaki, K., Ono, A., Kunisada, M., Nishino, M., Nagata, H., Sakamoto, S., et al. (2021). Mutation of the Imprinted Gene OsEMF2a Induces Autonomous Endosperm Development and Delayed Cellularization in rice. *Plant Cell.* 33 (1), 85–103. doi:10.1093/plcell/koaa006

Tonosaki, K., Sekine, D., Ohnishi, T., Ono, A., Furuumi, H., Kurata, N., et al. (2018). Overcoming the Species Hybridization Barrier by Ploidy Manipulation in the GenusOryza. *Plant J.* 93 (3), 534–544. doi:10.1111/tpj.13803

Vu, T. M., Nakamura, M., Calarco, J. P., Susaki, D., Lim, P. Q., Kinoshita, T., et al. (2013). RNA-directed DNA Methylation Regulates Parental Genomic Imprinting at Several Loci in Arabidopsis. *Development.* 140 (14), 2953–2960. doi:10.1242/dev.092981

Wang, G., Jiang, H., Del Toro de León, G., Martinez, G., and Köhler, C. (2018). Sequestration of a Transposon-Derived siRNA by a Target Mimic Imprinted Gene Induces Postzygotic Reproductive Isolation in Arabidopsis. *Developmental Cell.* 46 (6), 696–705. doi:10.1016/j.devcel.2018.07.014

Weinhofer, I., Hehenberger, E., Roszak, P., Hennig, L., and Köhler, C. (2010). H3K27me3 Profiling of the Endosperm Implies Exclusion of Polycomb Group Protein Targeting by DNA Methylation. *Plos Genet.* 6 (10), e1001152. doi:10.1371/journal.pgen.1001152

Wolff, P., Jiang, H., Wang, G., Santos-González, J., and Köhler, C. (2015). Paternally Expressed Imprinted Genes Establish Postzygotic Hybridization Barriers in *Arabidopsis thaliana*. *Elife.* 4, e10074. doi:10.7554/eLife.10074

Wolff, P., Weinhofer, I., Seguin, J., Roszak, P., Beisel, C., Donoghue, M. T. A., et al. (2011). High-resolution Analysis of Parent-Of-Origin Allelic Expression in the Arabidopsis Endosperm. *Plos Genet.* 7 (6), e1002126. doi:10.1371/journal.pgen.1002126

Xu, W., Yang, T., Dong, X., Li, D.-Z., and Liu, A. (2016). Genomic DNA Methylation Analyses Reveal the Distinct Profiles in Castor Bean Seeds With Persistent Endosperms. *Plant Physiol.* 171 (2), 1242–1258. doi:10.1104/pp.16.00056

Yan, D., Duermeyer, L., Leoveanu, C., and Nambara, E. (2014). The Functions of the Endosperm During Seed Germination. *Plant Cell Physiol.* 55 (9), 1521–1533. doi:10.1093/pcp/pcu089

Yuan, J., Chen, S., Jiao, W., Wang, L., Wang, L., Ye, W., et al. (2017). Both Maternally and Paternally Imprinted Genes Regulate Seed Development in rice. *New Phytol.* 216 (2), 373–387. doi:10.1111/nph.14510

Zemach, A., Kim, M. Y., Silva, P., Rodrigues, J. A., Dotson, B., Brooks, M. D., et al. (2010). Local DNA Hypomethylation Activates Genes in rice Endosperm. *Proc. Natl. Acad. Sci.* 107 (43), 18729–18734. doi:10.1073/pnas.1009695107

Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T.-f., Teng, C., et al. (2015). A One Precursor One siRNA Model for Pol IV-Dependent siRNA Biogenesis. *Cell.* 163 (2), 445–455. doi:10.1016/j.cell.2015.09.032

Zhang, H., Lang, Z., and Zhu, J.-K. (2018). Dynamics and Function of DNA Methylation in Plants. *Nat. Rev. Mol. Cell Biol.* 19 (8), 489–506. doi:10.1038/s41580-018-0016-z

Zhang, M., Xie, S., Dong, X., Zhao, X., Zeng, B., Chen, J., et al. (2014). Genome-wide High Resolution Parental-specific DNA and Histone Methylation Maps Uncover Patterns of Imprinting Regulation in maize. *Genome Res.* 24 (1), 167–176. doi:10.1101/gr.155879.113

Zhu, H., Xie, W., Xu, D., Miki, D., Tang, K., Huang, C.-F., et al. (2018). DNA Demethylase ROS1 Negatively Regulates the Imprinting of DOGL4 and Seed Dormancy in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA.* 115 (42), E9962–E9970. doi:10.1073/pnas.1812847115

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Lu, Zhai and Xi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 March 2022
doi: 10.3389/fgene.2022.833545

Integrated Analysis of Competitive Endogenous RNA Networks in Acute Ischemic Stroke

*Zongkai Wu[1,2], Wanyi Wei[2], Hongzhen Fan[2], Yongsheng Gu[2], Litao Li[2] and Hebo Wang[1,2]**

[1]Department of Neurology, Hebei Medical University, Shijiazhuang, China, [2]Department of Neurology, Hebei General Hospital, Shijiazhuang, China

Background: Acute ischemic stroke (AIS) is a severe neurological disease with complex pathophysiology, resulting in the disability and death. The goal of this study is to explore the underlying molecular mechanisms of AIS and search for new potential biomarkers and therapeutic targets.

Methods: Integrative analysis of mRNA and miRNA profiles downloaded from Gene Expression Omnibus (GEO) was performed. We explored differentially expressed genes (DEGs) and differentially expressed miRNAs (DEMirs) after AIS. Target mRNAs of DEMirs and target miRNAs of DEGs were predicted with target prediction tools, and the intersections between DEGs and target genes were determined. Subsequently, Gene Ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG) pathway enrichment analyses, Gene set enrichment analysis (GSEA), Gene set variation analysis (GSVA), competitive endogenous RNA (ceRNA) (lncRNA-miRNA-mRNA) network, protein–protein interaction (PPI) network, and gene transcription factors (TFs) network analyses were performed to identify hub genes and associated pathways. Furthermore, we obtained AIS samples with evaluation of immune cell infiltration and used CIBERSORT to determine the relationship between the expression of hub genes and infiltrating immune cells. Finally, we used the Genomics of Drug Sensitivity in Cancer (GDSC) database to predict the effect of the identified targets on drug sensitivity.

Result: We identified 293 DEGs and 26 DEMirs associated with AIS. DEGs were found to be mainly enriched in inflammation and immune-related signaling pathways through enrichment analysis. The ceRNA network included nine lncRNAs, 13 miRNAs, and 21 mRNAs. We used the criterion AUC >0.8, to screen a 3-gene signature (FBL, RPS3, and RPS15) and the aberrantly expressed miRNAs (hsa-miR-125a-5p, hsa-miR-125b-5p, hsa-miR-148b-3p, and hsa-miR-143-3p) in AIS, which were verified by a method of quantitative PCR (qPCR) in HT22 cells. T cells CD8, B cells naïve, and activated NK cells had statistical increased in number compared with the acute cerebral infarction group. By predicting the IC50 of the patient to the drug, AZD0530, Z.LLNle.CHO and NSC-87877 with significant differences between the groups were screened out. AIS demonstrated heterogeneity in immune infiltrates that correlated with the occurrence and development of diseases.

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Eman Toraih,
Tulane University, United States
Vincenzo Bonnici,
University of Parma, Italy

****Correspondence:***
Hebo Wang
wanghbhope@hebmu.edu.cn
https://orcid.org/0000-0002-0598-5772

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *11 December 2021*
Accepted: *25 February 2022*
Published: *25 March 2022*

Citation:
Wu Z, Wei W, Fan H, Gu Y, Li L and Wang H (2022) Integrated Analysis of Competitive Endogenous RNA Networks in Acute Ischemic Stroke. Front. Genet. 13:833545. doi: 10.3389/fgene.2022.833545

Conclusion: These findings may contribute to a better understanding of the molecular mechanisms of AIS and provide the basis for the development of novel treatment targets in AIS.

Keywords: acute ischemic stroke, differentially expressed genes, differentially expressed miRNAs, functional enrichment analyses, competitive endogenous RNA, protein-protein interaction, qPCR analysis

INTRODUCTION

Stroke is one of the leading causes of permanent disability worldwide and among the leading causes of mortality. Approximately nine million people worldwide suffer from stroke for the first time each year, and approximately 6.5 million people have long-term disabilities (Dennis et al., 2019). Acute ischemic stroke (AIS) is triggered by obstruction of blood flow to the local brain due to a clot or embolus blocking a cerebral artery (Adams et al., 1993; Fatahzadeh and Glick, 2006; Benjamin et al., 2018). In the ischemic stroke, a complex pathophysiological cascade is strongly correlated, both spatially and temporally, with the reduction of cerebral blood flow (Dirnagl et al., 1999). The pathogenesis of AIS includes ischemic brain injury caused by a chain of events (ischemic cascade) that are triggered by secondary injuries occurring hours or days after the initial event (Iadecola et al., 2020). However, the molecular mechanisms of ischemic stroke remain unclear. Therefore, to improve diagnosis and treatment, the molecular mechanism of ischemic stroke needs to be investigated. Early diagnosis and successful treatment beneficial in minimizing the damage to the brain, thus reducing mortality and improving prognosis. The current diagnosis of stroke has been severely hampered by the lack of rapid, valid, and analytically sensitive diagnostic biomarkers (Saenger and Christenson, 2010). Neuroimaging remains the most reliable tool for the diagnosis of ischemic stroke. Thrombolytic treatment through the tissue plasminogen activator (tPA) agent and surgical removal of clots represent the current therapeutic approaches for the treatment of AIS. While these therapies can restore cerebral blood flow and are efficient treatments for AIS, therapy designed to simultaneously target multiple mechanisms of cell injury is needed. Thus, to achieve improved clinical efficacy, there is an urgent need for novel biomarkers with high sensitivity and specificity for early diagnosis and treatment of ischemic stroke.

LncRNAs are non-coding RNAs with more than 200 nucleotides in length and lacking of the protein coding potential (Karagkouni et al., 2020). LncRNA molecule serves as a "sponge" and is capable to compete for miRNA binding (Karagkouni et al., 2020). Current evidence has shown that the expression and function of miRNA can exert either pro-inflammatory or anti-inflammatory effect after ischemic stroke, and that miRNAs are negatively regulated by lncRNAs. LncRNA SNHG14, which acts as a competitive sponge for miR-136–5p, miR-145–5p, and miR-199b, regulates the activation of microglia and exhibits pro-inflammatory ability (Qi et al., 2017; Zhong et al., 2019; Zhang G. et al., 2021). In contrast to the pro-inflammatory properties of the M1 microglia, the M2 microglia is responsible for the removal of debris and facilitating tissue repair through anti-inflammatory factors primarily at the recovery stage (Hu et al., 2015). Knockdown of lncRNA H19 can negatively regulate the expression of miR-29b and miR-138–5p, and therefore, can promote functional recovery after cerebral ischemia and the polarization of microglia (Li et al., 2020; Xu J. et al., 2021). A recent study shows that lncRNA MALAT1, sponging miR-30a, promotes neuronal cell death and suppresses autophagy in ischemic stroke (Guo et al., 2017). Studies have shown that miR-145 plays an essential role in inflammation after ischemic stroke. A previous study also indicated that the production of inflammatory cytokines is regulated by lncRNA TUG1 at an early stage after ischemic injury by targeting miR-145a-5p (Wang et al., 2019). Enhancing our understanding of the interactions between RNAs through competitive endogenous RNA (ceRNA) network can elucidate new AIS-related molecular mechanisms and identifying novel biomarkers for AIS. Thus, elucidation of the mechanistic details of AIS occurrence and progression, and exploration of potential biomarkers and therapeutic targets are critical to developing new treatments and diagnostics.

Neuroinflammation is driving cause of the pathophysiological processes leading to ischemic stroke (Doll et al., 2014; Zhao et al., 2016). Several pathophysiological processes could negatively affect homeostasis of physiological functions, including excitotoxicity, excessive formation of reactive oxygen species (ROS), loss of glucose, and oxygen mitochondrial dysfunction, neuronal apoptosis, and blood-brain barrier permeability (Forrester et al., 2018; Chen W. et al., 2020; Xu Q. et al., 2021). Local and peripheral immune system plays important roles in the pathophysiology of stroke, and includes both the innate and the adaptive immune responses (Benakis et al., 2016; Venkat et al., 2018). Neurotoxic factors including reactive oxygen and nitrogen species as well as exopeptidases can be released immediately after peripheral immune system contribute to secondary neurodegeneration (Benakis et al., 2016). Immune cells, including microglia, monocyte/macrophages, neutrophils, and lymphocytes infiltrate into the brain after stroke and induce inflammatory or anti-inflammatory responses via distinct pathways (Zhang S. R. et al., 2021).

Two studies (Zhang et al., 2020; Sun et al., 2021) aimed to explore possible molecular mechanisms of ischemic stroke by constructing ceRNA networks. Our study used different datasets from that of Sun et al. The study of Sun et al. included the miRNA expression profile of GSE55937, the mRNA and lncRNA expression profile of GSE122709, and the mRNA expression profile of GSE146882. However, our study included the mRNA expression profiles of GSE16561, and the miRNA dataset GSE110993. These two studies (Zhang et al., 2020; Sun et al., 2021) only performed ceRNA network analysis, while our

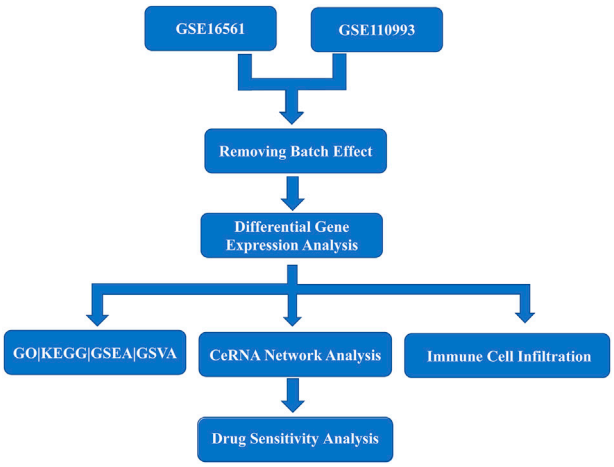


FIGURE 1 | Flow chart of overall analysis.

study used the hypergeometric distribution model to evaluate the ceRNA interaction network and performed ROC analysis for key genes. Relative to these two studies, we added GSEA analysis, GSVA analysis, TFs network analysis, immune infiltration analysis and drug network analysis. The ceRNA network we found includes 9 lncRNAs such as AL360004.1, LINC00173, LINC01089, LINC00115, 13 miRNAs such as hsa-miR-125a-5p, hsa-miR-125b-5p, hsa-miR-148b-3p and hsa-miR-143-3p and 21 mRNAs such as FBL, RPS3, RPS15. The difference was that the ceRNA network constructed by Sun et al.'s study contains 7 mRNAs, 14 lncRNAs, such as SND1-IT1, NAPA-AS1, LINC01001, LUCAT1, ASAP1-IT2, 8 miRNAs, such as miR-93 -3p, miR-24-3p. Zhang et al. found that MCM3AP-AS1, LINC01089, ITPK1-AS1 and HCG27 may be new biomarkers and potential targets for AIS therapy.

Finally, we determined the DEGs and DEMirs in AIS through a comprehensive analysis. Gene expression profiles in cerebral infarction samples were obtained through the GEO database. Subsequently, GO, KEGG, GSEA, and GSVA were used to study the molecular mechanisms for DEGs in AIS. Then, a ceRNA network and TFs network were established. Receiver operating characteristic (ROC) curve analysis was implemented to explore the diagnostic validity of the identified DEMirs and DEGs. We identified three potential mRNAs and four potential miRNAs as important predictors of AIS. Moreover, we further conducted immune infiltration analysis and drug sensitivities of the cell lines expressed, as the half maximal inhibitory concentration (IC50). Our research will clarify the molecular mechanisms of AIS and provide the basis for new applications in both diagnosis and treatment.

MATERIALS AND METHODS

Data Download and Data Pre-Processing

Acute cerebral infarction expression profile datasets GSE16561 (Barr et al., 2010) and GSE110993 (Tiedt et al., 2017) with reliable sample sources were downloaded from the GEO (https://www.ncbi.nlm.nih.gov/geo/) database by using the GEO query package (Davis and Meltzer, 2007)of R software (version 3.6.6). The samples in the datasets are all from *Homo sapiens*. The data in the two datasets were generated using different platforms: GPL6883 Illumina HumanRef-8 v3.0 expression beadchip and GPL15456 Illumina HiScanSQ respectively. GSE16561 dataset includes whole blood samples from 39 patients with acute cerebral infarction and 24 healthy controls. The GSE110993 dataset includes whole blood samples from 20 patients with acute cerebral infarction and 20 healthy patients for inclusion. The raw data were converted into an expression matrix and corrected for background and normalized by the limma package (Ritchie et al., 2015). Afterwards, the batch effect was removed in sva package (Leek et al., 2012). A flow diagram for the present analysis is shown in **Figure 1**.

Data Analysis: Differential Expression and Pathway Analyses

First, Principal Component Analysis (PCA) was conducted using FactoMineR package (Lê et al., 2008). Then, DEGs were identified using the limma package (Ritchie et al., 2015), and volcano plots of DEGs were generated using the ggplot2 package (Walter et al., 2015). Finally, adjusted $p < 0.05$ and $|\log2FC| > 0.5$ were used as the cutoff criteria to identify DEGs. We performed GO and KEGG enrichment analysis using the clusterProfiler R package (Yu et al., 2012), and differences were considered statistically different at adjusted p value <0.05.

GSEA Analysis, GSVA Analysis

GSEA was performed on the gene expression matrix using clusterProfiler R package (Yu et al., 2012), and enrichment was considered significant for false discovery rate (FDR) < 0.25 and $p < 0.05$. The gene sets were analyzed using the gene matrix transposed (gmt) file downloaded from MSigDB. Each gene set was constructed into a GSVA score matrix. Then, according to the GSVA scores, the gene sets were separated into low- and high-score groups. The limma R package was applied to define the significant differences between the low- and high-score groups. Finally, heat maps were drawn using the R package pheatmap.

ceRNA Interaction Network Analysis

Normalized data were analyzed using the GDCRNATools package (Li et al., 2018), and then screened for differential mRNAs, miRNAs, and lncRNAs. The miRcode database (Jeggari et al., 2012) was respectively utilized to pair lncRNA-miRNA and mRNA-miRNA. The hypergeometric distribution model was constructed to evaluate the ceRNA interaction network, and finally visualization was performed with Cytoscape (Shannon et al., 2003).

ROC Analysis of Key Molecules

A pROC package (Robin et al., 2011) was used to perform the leave-one-out jackknife approach and draw the ROC curve of key molecules, with the sensitivity as the ordinate and 1-specificity as

the abscissa. The area under the curve (AUC) served as the main evaluation performance. The higher the AUC value, the better is the predictive power.

Immune Cell Infiltration Level, Correlation Analysis of Immune Cells

CIBERSORT (Newman et al., 2015) is a tool used for deconvolution of the transcriptome expression matrix based on the principle of linear support vector regression, which can estimate the composition and abundance of infiltrating immune cells of the acute cerebral infarction sample in the mixed cells. After uploading the gene expression matrix data to CIBERSORT (Newman et al., 2015) and filtering the outputs ($p < 0.05$), we obtained the immune cell infiltration matrix and the immune infiltration distribution results of acute cerebral infarction. The resulting correlation of 21 types of infiltrated immune cells was visualized in a heat map format generated by the corrplot package (Friendly, 2002) of R.

The Upstream Transcription Factor Network That Regulate miRNAs and Drug Network

Prediction of transcription factors regulating differentially expressed genes was analyzed by FunRich software (Pathan et al., 2015). By selecting the intersection molecules of the predicted transcription factor, mRNA, and miRNA target genes, we reconstructed the predicted transcription factor-miRNA regulatory networks. In addition, Genomics of Drug Sensitivity in Cancer (GDSC database) (Yang et al., 2013), that covers the sensitivity and response of cells to drugs, was employed. For further network pharmacology analysis, we predicted the IC50 value and compared the *p*-value of the rank sum test between acute cerebral infarction and normal samples to determine drug sensitivity.

Cell Culture and Treatment

The immortalized mouse hippocampal neuronal cell line, HT22 (Zhejiang Ruyao Biotechnology Co. Ltd., Zhejiang, China), was cultured in Dulbecco's modified Eagle's medium (DMEM, Corning, NY, United States) containing 1% penicillin/streptomycin and 10% fetal bovine serum (FBS, BI, Israel) at 37°C in a humidified incubator containing 5% CO2. For cell lysis of adherent cells, cells were grown to 60–80% confluence and were rinsed with PBS before trypsinization. Thereafter, HT22 cells were randomly divided into the normal group and the model group. Cells were seeded in a 96-well plates at a density of 5 × 104 cells/well and cultured for 24 h. After 24 h, the culture medium was discarded, and the cells of model group were washed twice with PBS. Serum and glucose free media were used, and cells were placed in an anaerobic culture box inside the 37°C incubator for 4 h. After hypoxia, glucose free media was replaced with complete DMEM and the cells were reoxygenated in a normoxic incubator at 37°C. These cells of control group were cultured at 37°C in a 5% CO2 incubator under normal atmospheric oxygen conditions. Cell viability from the model and control groups was determined using the Cell Counting Kit-8 assay (CCK-8, Biosharp, China).

Quantitative Polymerase Chain Reaction (qPCR Analysis)

Total RNA was extracted using the TRIzol (Life Technologies, Carlsbad, CA, United States), following the manufacturer's instructions. Briefly, HT22 cells were lysed in TRIzol, then 60 μL chloroform was added, samples were shaken for 1 min, incubated at room temperature for 5 min and centrifuged for 15 min at 12000 g at 4°C. The aqueous phase was transferred into a new tube and RNA was precipitated in the presence of isopropanol. After centrifugation, the supernatant was discarded and the pellet was washed with 200 μL 75% ethanol, made with Diethyl pyrocarbonate (DEPC)-treated water, by centrifugation at 7,500 g, for 10 min at 4°C. RNA was eluted in 20 μL DEPC-treated water, quantified by SmartSpec Plus (Bio-Rad, Hercules, CA) and stored at −80°C. Reverse transcription reactions (37°C for 15 min, 85°C for 5 s, 4°C) were carried out using the Evo M-MLV kit (Accurate Biotechnology Co., Ltd, China). The cDNA was synthesized (37°C for 60 min, 85°C for 5 min, 4°C) using the miRNA first-strand cDNA synthesis kit (Accurate Biotechnology Co., Ltd, China). Then, qPCR was performed using the SYBR Green Pro Taq HS qPCR premix and PCR-amplified in a two-step process. The PCR amplification conditions were the following: 40 cycles of 30 s at 95°C, 5 s at 95°C and 30 s at 60°C. The relative expression levels were calculated using the $2^{-\triangle\triangle Ct}$ method with GAPDH as an internal reference gene. The melting curve was analyzed to assure specificity of the primers after each reaction. See **Supplementary Table S1** for primer sequences.

STATISTICAL ANALYSIS

All analyses were performed using R software for statistical calculations (version 4.0.2). Independent sample *t*-test was used to estimate normally distributed variables whereas Mann-Whitney *U* test was used to compare non-normally distributed variables. The statistical significance of categorical variables was compared using the chi-square test or Fisher exact test. Pearson correlation coefficients were calculated to define the correlation between different genes. ROC curves were generated with the R package pROC, and the corresponding area under curve (AUC) values calculated. All statistical tests were performed two-sided, and statistical significance was set at $p < 0.05$. The results were shown as the mean ± standard deviation (SD).

RESULTS

Data Pre-Processing

PData function was available for getting grouping information of the expression profile datasets GSE16561 (Barr et al., 2010) and GSE110993 (Tiedt et al., 2017). Each expression matrix from two raw datasets (GSE16561 and GSE110993) was obtained and then

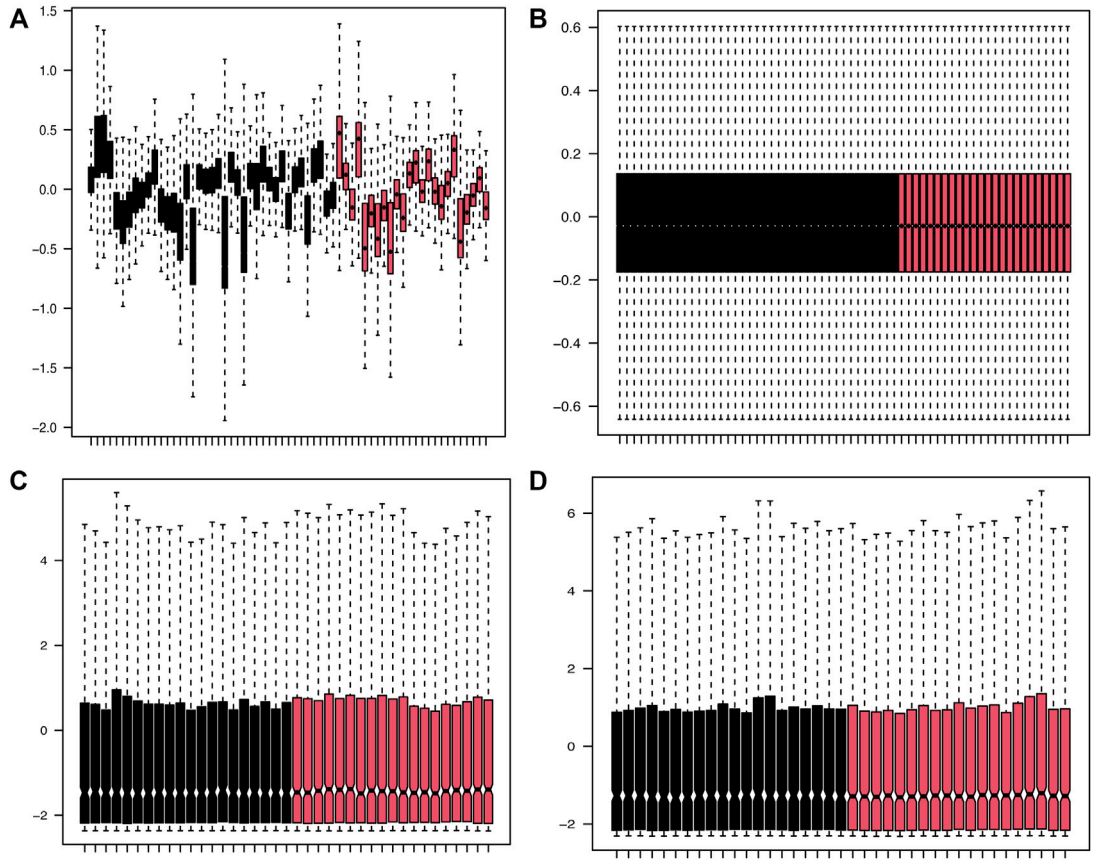


FIGURE 2 | Density plots of the dataset samples before and after correction. **(A,B)** The boxplot of GSE16561 dataset samples before and after correction after removing the inter-batch differences. **(C,D)** The boxplot of GSE110993 dataset samples before and after correction after removing the inter-batch differences.

pre-processed identically for background correction and normalization using the limma package (Ritchie et al., 2015). Finally, batch effects were removed using the sva package (Leek et al., 2012). The corresponding boxplot is shown in **Figure 2**.

Genes Associated With Cerebral Infarction

After normalization, the gene expression matrices of the two datasets were presented as PCA plots (**Figures 3A,B**). The results showed that the two groups of samples clustered more obviously after normalization, indicating that the source of the samples was reliable. We used R software to preprocess the data and performed differential expression analysis on the gene expression matrix of GSE16561 data set to obtain 3698 DEGs, while the differential analysis in the GSE110993 data set yielded 62 DEMirs. The results are shown in the volcano plots (**Figures 3C,D**) and the heat maps (**Figures 3E,F**).

Intersected Differentially Expressed Genes and Target Genes

The starbase database, a non-coding RNA database, was used to find target genes based on miRNA, verify the interaction between miRNA and mRNA, and generate a target gene RNA network. Based on the differential expression results and the prediction results of starbase, the intersection of the DEGs and the miRNA target genes was drawn (**Figure 4A**). The intersection of the differentially expressed miRNAs and target miRNAs of DEGs included 26 genes, as shown in **Figure 4B**.

GO/KEGG Enrichment Analysis, Pathway Diagram (Based on DEGs)

Based on the intersection of the DEGs and the miRNA target genes, GO analysis was performed. The main biological processes involving the DEGs included neutrophil activation, neutrophil degranulation, neutrophil activation involved in immune response. In terms of cellular component, the DEGs were mostly enriched in secretory granule membrane and cytoplasmic vesicle lumen. In terms of molecular functions, the DEGs were linked with amide binding, peptide binding, structural constituent of ribosome, amyloid-beta binding (**Figure 5A**). As shown in **Figure 5B**, KEGG analysis results included Coronavirus disease (Covid-19), Hematopoietic cell lineage, Tuberculosis, etc. **Figure 5C** is an Upset plot of the gene ontology (GO) analysis. The main enrichment can reflect the intersection between different terms. Similarly, **Figure 5D** is

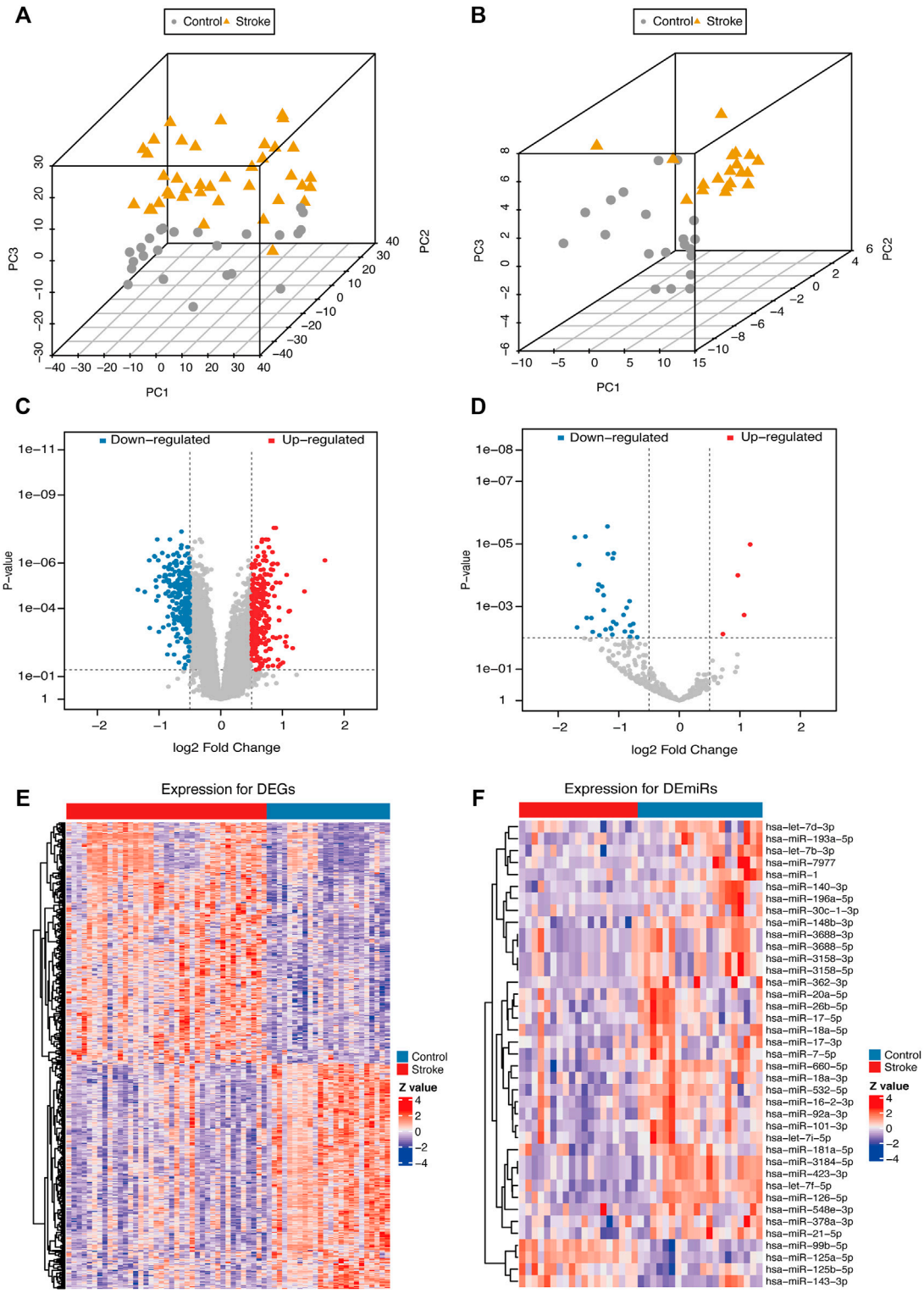


FIGURE 3 | PCA plots and differential expression of the samples of the data sets after correction. PCA plots of the GSE16561 **(A)** and GSE110993 **(B)** datasets after removing the inter-batch differences. Volcano plots of the GSE16561 **(C)** and GSE110993 **(D)** dataset; red plot represented upregulation, and blue plot represented downregulation. Heat maps of the GSE16561 **(E)** and GSE110993 **(F)** datasets. The color scale represented the abundance of gene expression. The darker the color shade, the higher expression level.

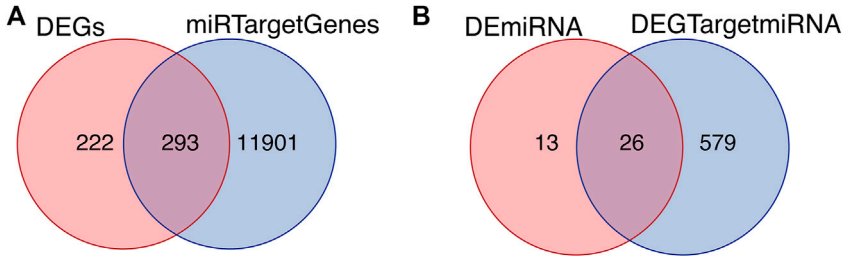


FIGURE 4 | Intersected differentially expressed genes and target genes. **(A)** The intersection of DEGs and miRNA target genes. **(B)** The intersection of differentially expressed miRNAs and target miRNAs of DEGs.

an Upset plot of KEGG data. **Figure 5E** is a circos diagram of GO analysis based on differentially expressed mRNA. **Figure 5F** is a circos diagram of KEGG pathway enrichment analysis. The **Supplementary Table S2** shows the detailed enrichment analysis results. The two key pathway diagrams were constructed according to the above-mentioned differentially expressed mRNA, as shown in **Figure 6**.

GSEA and GSVA

GSEA showed that the pathways were mainly enriched aromatic compound catabolic process, cellular nitrogen compound catabolic process, heterocycle catabolic process, nucleobase-containing compound catabolic process, protein targeting, autophagy-animal, NOD-like receptor signaling pathway, osteoclast differentiation, regulation of actin cytoskeleton, Ribosome pathway (**Figures 7A–J**). The detailed enrichment results are deposited in https://www.ncbi.nlm.nih.gov/gds NCBI: GEO. Accession numbers are GSE16561 and GSE110993 respectively. Furthermore, the gene set variation of each sample in each specific pathway converted the new biological function annotation into a new expression matrix. GSVA analysis is shown in **Figures 8A,B**. There were differences in the grouping of terms such as KEGG-ribosome, GO BP-pyrimidine nucleotide biosynthetic process between the patients with acute cerebral infarction and the control group, or high and low expression groups, which were consistent with **Figures 5A,B**.

Competitive Endogenous RNA Interaction Network Analysis, PPI Network Analysis

We used the GDCRNATools package for data standardization, and then screened for differential mRNAs, miRNAs, and lncRNAs, including different forms of upregulation and downregulation, as shown in **Figure 9A**. Considering the number of miRNAs shared by mRNA, and lncRNA, key mRNAs were locked by combining with the miRcode database, as a database support for the interaction of lncRNA-miRNA and miRNA-mRNA. MiRNAs were identified through the mechanism of mRNA binding to ceRNA, and then a model was generated to evaluate the ceRNA interaction network, visualized using Cytoscape (**Figure 9A**). The ceRNA network was displayed also as a Sankey diagram (**Figure 9B**). The network contained nine specific lncRNAs, 13 miRNAs, and 21 mRNAs. AL360004.1 (degree = 5),has-miR-125a-5p (degree = 16), hsa-miR-125b-5p (degree = 16), and KRT10 (degree = 5) were considered the most important transcripts among the lncRNAs, miRNAs, and mRNAs, respectively. Because hsa-miR-125a-5p, hsa-miR-125b-5p had the highest ceRNA degree (degree = 16), we concluded that this family gene might have an important influence on the pathogenesis of acute cerebral infarction. Next, we explored the interaction relationship between proteins encoded by different genes. PPI network of DEGs was established (**Figure 9C**), and the hub genes relationship was generated (**Figure 9D**). At the same time, according to the existing ceRNA network, interactions between differentially expressed miRNAs and their target genes were analyzed, as shown in **Figure 9E**, and its hub-genes are shown in **Figure 9F**.

ROC Analysis of Key mRNA and miRNA

According to our previous ceRNA analysis (section 3.6), the molecules participating in the interaction network play a key role in the pathogenesis of acute cerebral infarction. Furthermore, we plotted the ROC curve by selecting the key mRNAs and miRNAs (**Figures 10A–G**), and then screened the biomarkers to reflect its predictive ability and accuracy for the disease.

Immune Cell Infiltration Level and Correlation Analysis of Immune Cells

By deconvolution analysis via CIBERSORT of the expression matrix of 21 immune cell subtypes were analyzed, using the limited threshold is $p < 0.05$. We obtained the immune cell infiltration matrix, and the results of immune infiltration distribution in acute cerebral infarction samples. Immune cell infiltration is shown in **Figure 11A**. Compared with other immune cells, T cells CD8, T cells CD4 naïve, and T cells CD4 memory resting were more infiltrated in samples that were not divided into acute cerebral infarction and control samples. However, Macrophage M and resting Dendritic cells infiltration is limited. After further grouping analysis, the significant infiltration of Macrophages M0, activated Mast cells, and Monocytes in the acute cerebral infarction samples difference compared with the control group (**Figure 11B**). In the control group, T cells CD8, B cells naïve, and activated NK cells had statistical increased in number compared with the acute cerebral infarction group. The next step was to analyze the correlation of different immune cells. As shown in

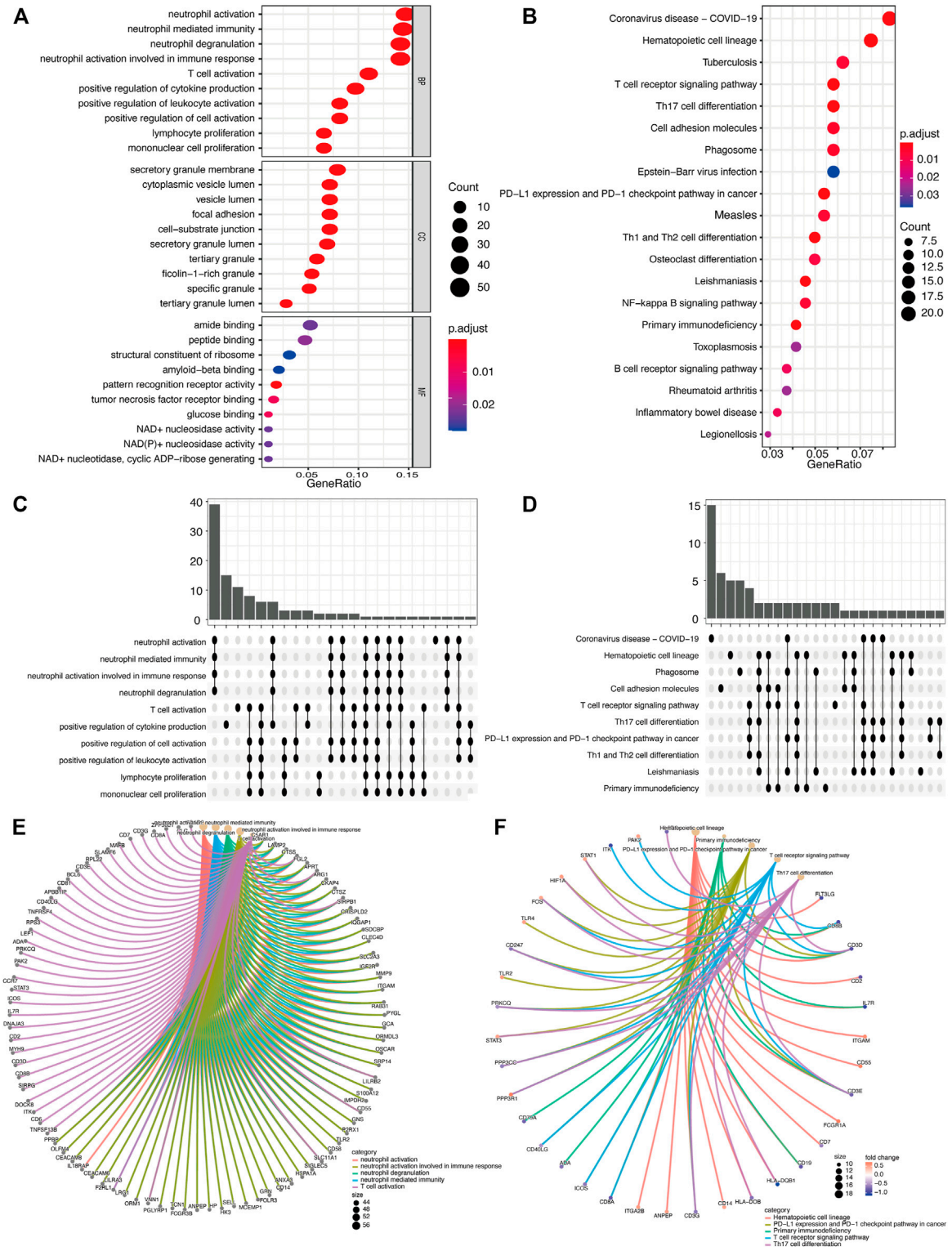


FIGURE 5 | GO/KEGG function enrichment analysis. **(A)** In GO biological function enrichment analysis, the X horizontal axis represents the proportion of DEGs enriched in GO term, and the color of the dot represents the adjusted *p* value: the redder the color, the smaller the adjusted *p* value; the bluer the color, the greater the adjusted *p* value. The size of the dot represents the amount of enriched mRNA. **(B)** In KEGG enrichment analysis, the X horizontal axis represents the proportion of DEGs, and the color of the dot represents the corrected *p* value. **(C)** GO function enrichment analysis upset chart. The horizontal axis represents the categories of term names enriched by DEGs, and the vertical axis represents the number of DEGs in this term. **(D)** KEGG function enrichment analysis Upset plot. **(E)** GO function enrichment analysis circos plot. **(E)** The outer circle is the information of the corresponding entry gene in the enrichment analysis, and the line is the corresponding enrichment term entry. **(F)** KEGG function enrichment analysis circos plot.

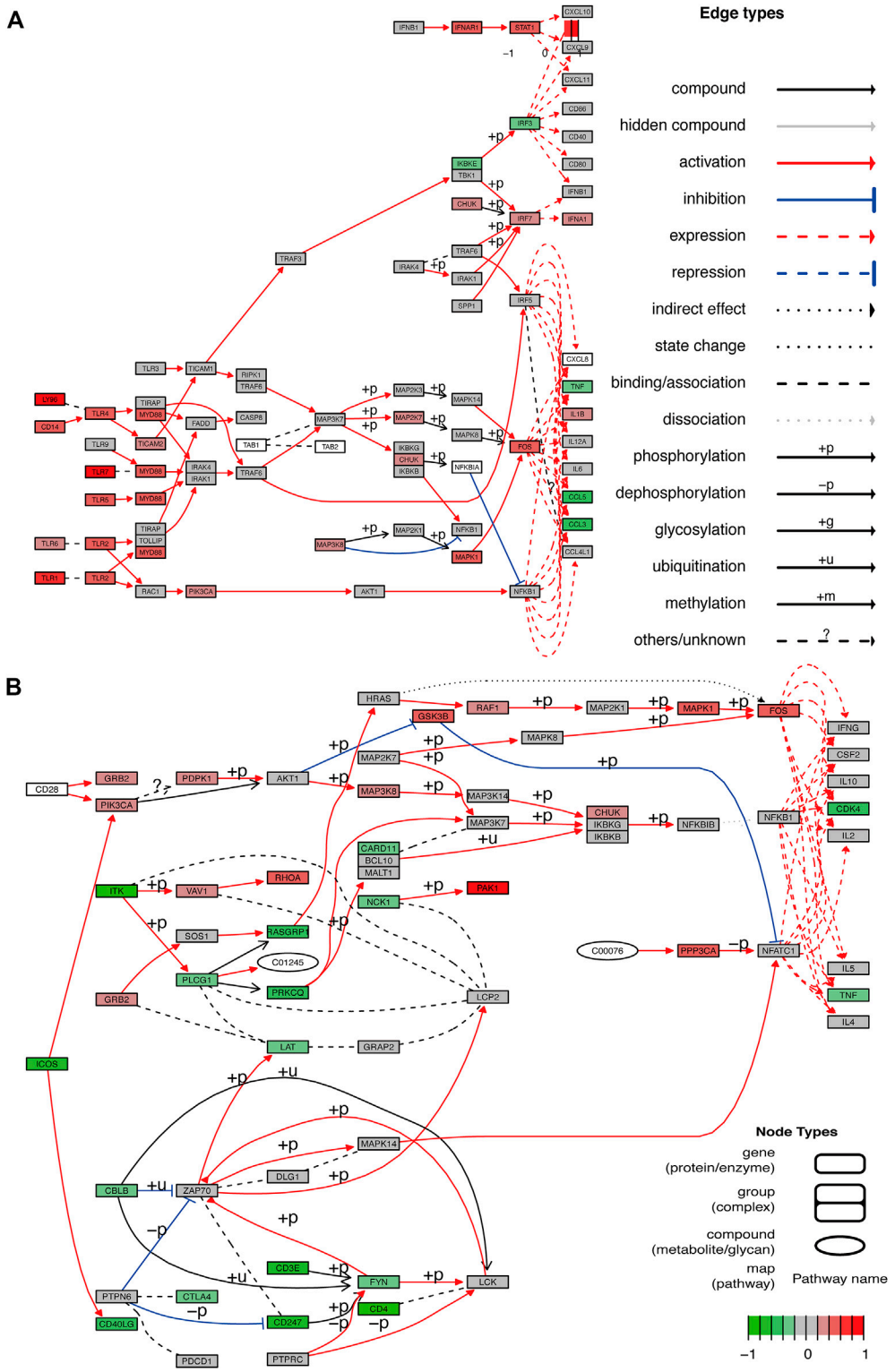


FIGURE 6 | Pathway diagram. **(A,B)** Two pathway diagrams composed of two major networks are constructed using DEGs.

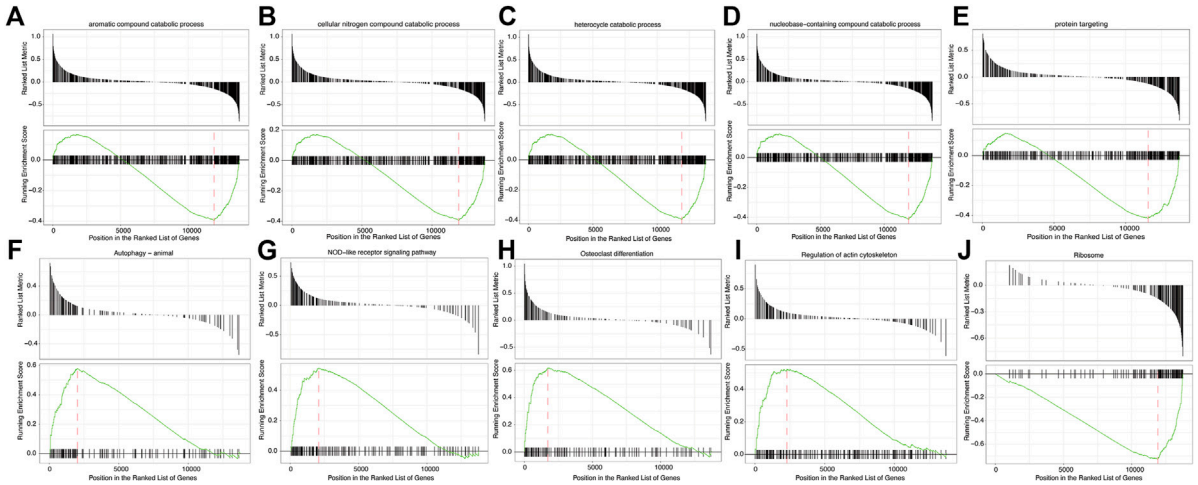


FIGURE 7 | Gene Set Enrichment Analysis (GSEA). **(A–J)** GSEA enrichment analysis result sub-graph. The upper part of the graph represents the distribution of rank values of all genes after sorting, and the Signal2Noise algorithm is used by default. The lower part of the graph represents the line chart of the gene Enrichment Score, the horizontal axis is each gene in the gene set, and the vertical axis is the corresponding result.

Figure 11C, the proportions of different subgroups of infiltrating immune cells were weakly to moderately correlated. Neutrophils had a strong positive correlation with Macrophages M0 and activated Mast cells. T Cells CD8 were positively correlated with activated NK cells and B cells naïve, and Neutrophils were negatively correlated with T cells CD8. **Figure 11D** shows the changes in the proportion of immune cells within and between different groups. Therefore, the abnormal immune infiltration and the heterogeneity of immune infiltration in acute cerebral infarction suggested that different immune cells play a role in the occurrence and development of the disease. The immune signature may be used as prognostic targets for immunotherapy and may have significant clinical significance.

Correlation Between Diagnostic Markers and Infiltrating Immune Cells

In the previous analysis, differentially expressed hub genes were selected, and ROC analysis was performed to screen for diagnostic markers. To determine the correlation between the hug genes and the immune cell infiltrations in acute cerebral infarction, linear regression was performed on the model, and the goodness-of-fit of regression model coefficients were evaluated to clarify the correlation between diagnostic markers and immune cell infiltration (**Figure 12**).

Target Genes and Transcription Factor Network Analysis

Differential expression of mRNAs, miRNA target genes, and transcription factors were intersected to obtain four molecules, CEBPD, MAFB, FOS, and STAT1 (**Figure 13A**). Applying the intersection of DEGs and DEmiRTargetGenes as starting point, a network between these four molecules and miRNAs was generated (**Figure 13B**). The network integrating differential lncRNAs, miRNAs, and mRNAs shows the interaction analysis between target genes and transcription factors (**Figure 13C**).

Drug Sensitivity Analysis

The half-maximal inhibitory concentration (IC50), that is, the concentration of a drug that inhibits cell growth by 50% in different treatments, was predicted using the GDSC database. The value reflects the degree of cell tolerance to the drug. The lower the IC50 value, the more sensitive the cells are to drugs. IC50 and AUC values were obtained from all cell lines and drug combinations through the GDSC database. According to the screening results, we compared acute cerebral infarction with the control group and integrated the data. Finally, we visualized and predicted the IC50 of patients on drugs, and screened out drugs with significant differences between groups. (**Figure 14**).

Validation of the Identified miRNAs and mRNAs

Five Cell viability was assessed through the CCK-8 assay. Oxygen glucose deprivation/re-oxygenation (OGD/R) was used to mimic neural injury. Data demonstrated thatOGD/R insult in model group exhibited decreased cell viability, compared to the normal control group. When compared with the normal control group, the miRNA expression levels of miR-148b-3p ($p < 0.01$), miR-125a-5p ($p < 0.05$), miR-125b-5p ($p < 0.01$) and miR-143-3p ($p < 0.05$) in the model control group were significantly downregulated while the mRNA expression levels of FBL ($p < 0.01$), RPS3 ($p < 0.01$), and RPS15 ($p < 0.01$) were significantly upregulated (**Figure 15**).

DISCUSSION

The incidence of AIS increases with aging and is linked to a poor prognosis. Timely monitoring of undiagnosed strokes is critical to

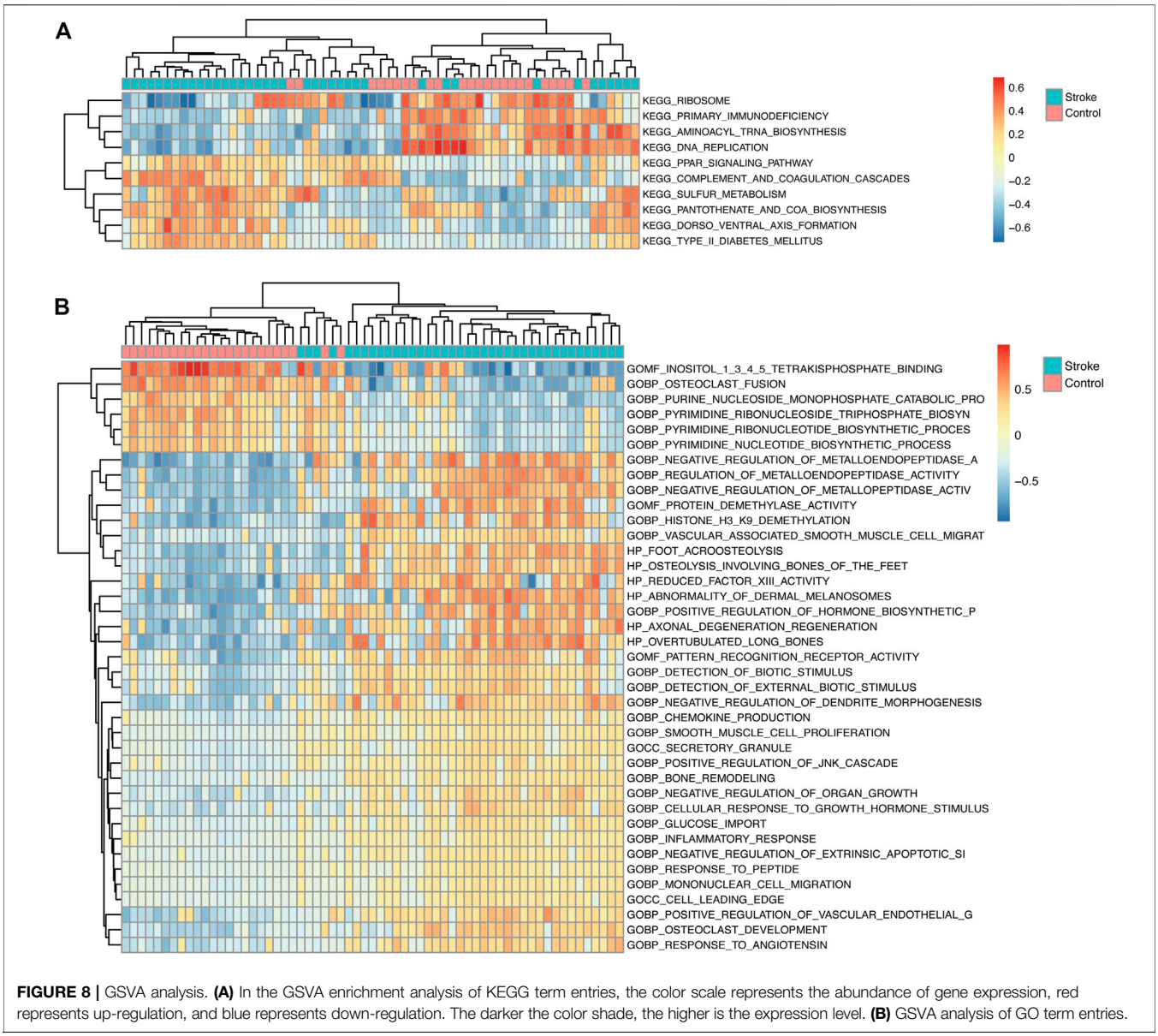


FIGURE 8 | GSVA analysis. **(A)** In the GSVA enrichment analysis of KEGG term entries, the color scale represents the abundance of gene expression, red represents up-regulation, and blue represents down-regulation. The darker the color shade, the higher is the expression level. **(B)** GSVA analysis of GO term entries.

reduce mortality. However, long delays in imaging and treatment initiation and poor functional prognosis represent a significant challenge. Despite remarkable advances in the development of therapeutic strategies, developing effective targeted molecular therapies is limited. Therefore, there has been an increased search for noninvasive and quantitative markers for AIS. This study aimed to identify novel biomarkers with sensitivity and specificity for the diagnosis and treatment of AIS. Through a comprehensive analysis, a total of 3698 DEGs were identified from the mRNA microarray dataset (GSE16561) and a total of 26 DE-miRNAs were identified from the miRNA high-throughput sequencing dataset (GSE110993). GO enrichment, KEGG pathway analysis, GSEA and GSVA were performed to enrich and analyze DEGs. We performed ceRNA interactions analysis and a network of miRNA-mRNA interactions in patients with acute cerebral infarction, allowing us to uncover potential biomarkers associated with AIS. A total of 17 hub genes were identified based on DEGs and miRTargetgenes. We compared the levels of immune cell infiltrates in cerebral infarction group and control group, and we found that screened diagnostic markers correlated with immune cell infiltration. We also identified drugs with significant differences between groups by predicting the patient's IC50 for the drug.

First, functional annotation to DEGs displayed enrichment of GO and KEGG pathway analyses referring to inflammation and immune response. Biological processes analysis in GO annotation indicated that DEGs were primarily enriched in the inflammatory response associated with neutrophils, T cell and lymphocytes. Among them, neutrophils are involved in multiple biological processes, including neutrophil activation, neutrophil degranulation, and neutrophil activation. Neutrophils have received particular attention during recent years of their

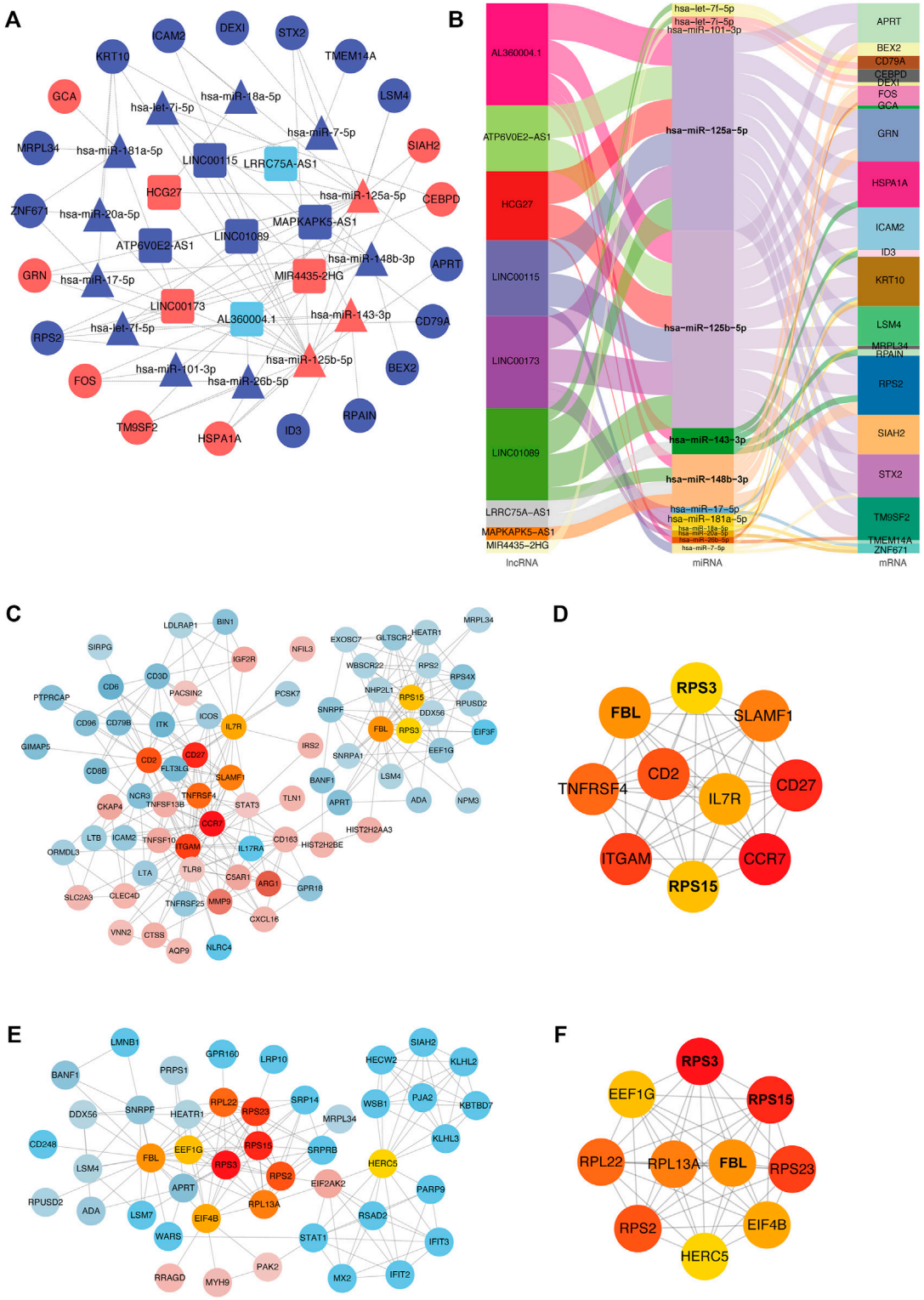


FIGURE 9 | ceRNA interaction and protein-protein interaction analysis. **(A)** ceRNA network diagram. In the network diagram, red indicates upregulation, blue indicates downregulation, squares indicate lncRNA, triangles indicate miRNA, and circles indicate mRNA. Sankey diagram **(B)**. The three columns include lncRNAs, miRNAs, and mRNAs in order from left to right. The line colors represent different types of gene-gene interactions. **(C)** Diagram of interaction of differentially expressed proteins. Red indicates increased expression, blue indicates decreased expression, and color intensity indicates different degrees of u-regulation or downregulation. Orange represents the hub genes. **(D)** Diagram of hub-genes interaction in differentially expressed proteins. **(E)** MiRNAs targeting mRNAs interaction diagram. Red indicates upregulated expression, blue indicates downregulated expression, and color intensity indicates different degrees of upregulation or downregulation. Orange represents hub genes. **(F)** Diagram of hub genes interaction in targeted mRNAs.

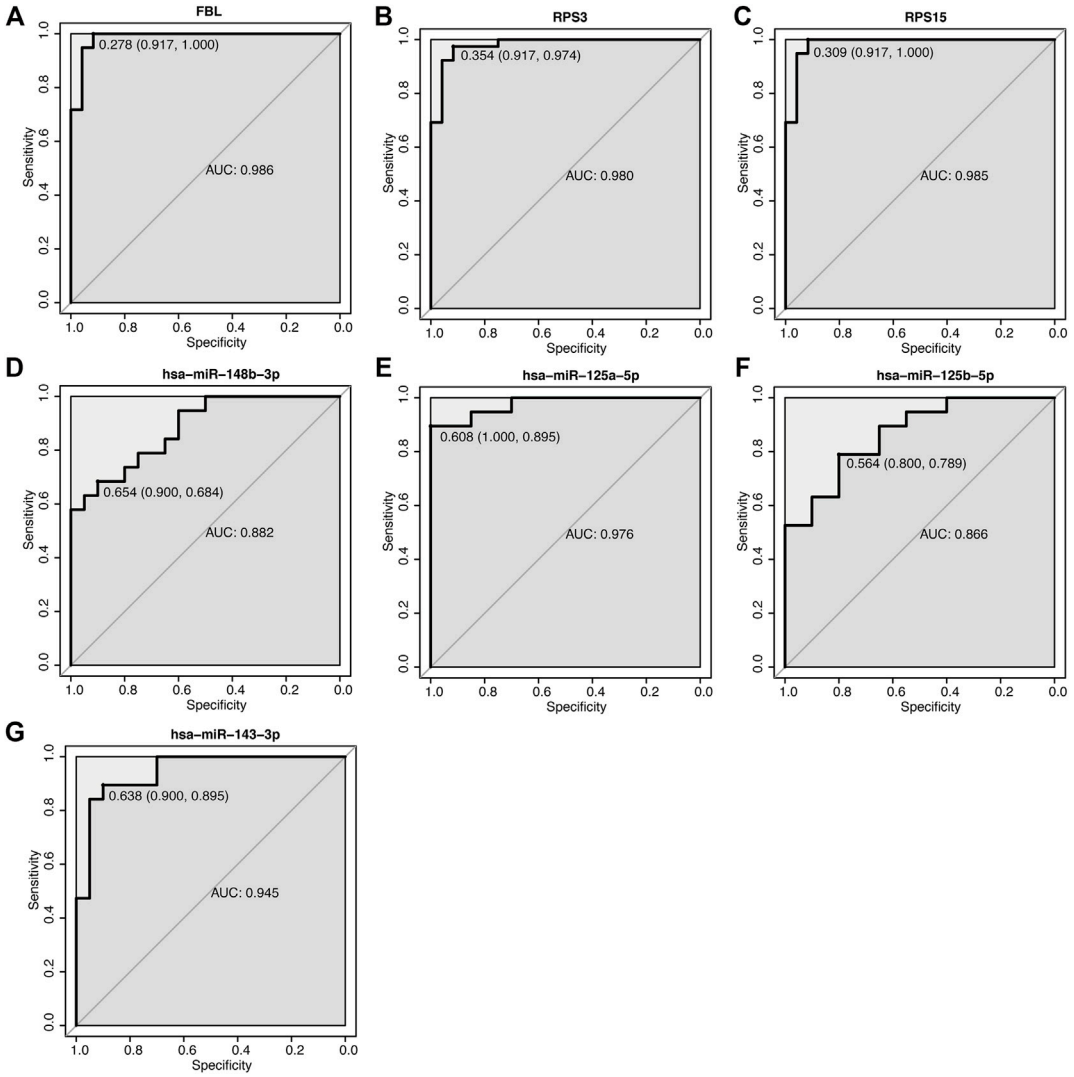


FIGURE 10 | ROC curve of key mRNA and miRNA. **(A–G)** ROC curve of mRNA and miRNA. The abscissa is specificity, and the ordinate is sensitivity (true positive rate), specificity = 1 (false positive rate) AUC is the area under the ROC curve enclosed by the coordinate axis.

significant destructive potential. Experimental studies have shown that neutrophils reach the ischemic area in the first few hours after an ischemic attack (Perez-de-Puig et al., 2015). They can release neurotoxic proteolytic enzymes, accumulate neutrophils in blood vessels, block blood flow in capillaries and cause the no-reflow phenomenon (Allen et al., 2012; Perez-de-Puig et al., 2015).In terms of molecular function annotations, the occurrence of AIS was closely related to amide binding, peptide binding, structural constituent of ribosome, and amyloid-beta binding. KEGG suggested that most of DEGs in subjects were mainly enriched in the toll-like receptor signaling pathway, Cell adhesion molecules, T cell receptor signaling pathway, NF–kappa B signaling pathway, the B cell receptor signaling pathway, which have collectively been confirmed as essential mechanisms in inflammation of ischemic stroke. The theranostic strategy is a combination of diagnosis and therapy which can be furnished through analyzing relevant data in the GEO database and forming an AIS-related lncRNA-miRNA-mRNA regulatory network.

To detect underlying biological functions, GSEA and GSVA were performed. The results of GSEA suggest that the NOD-like receptor signaling pathway and autophagy pathway were significantly enriched pathways. In cerebral ischemic injury, autophagy can be protective (Carloni et al., 2010) or destructive (Koike et al., 2008). If its protective function can be controlled, autophagy may become novel therapeutic targets for ischemic brain (Carloni et al., 2010). Moreover, NOD-like receptor, which regulates innate immunity and inflammatory processes (Shiau et al., 2013), is expected to become as a therapeutic target in ischemic stroke. GSVA analysis confirmed that the most abundant pathways are related to immune response, inflammation and apoptosis.

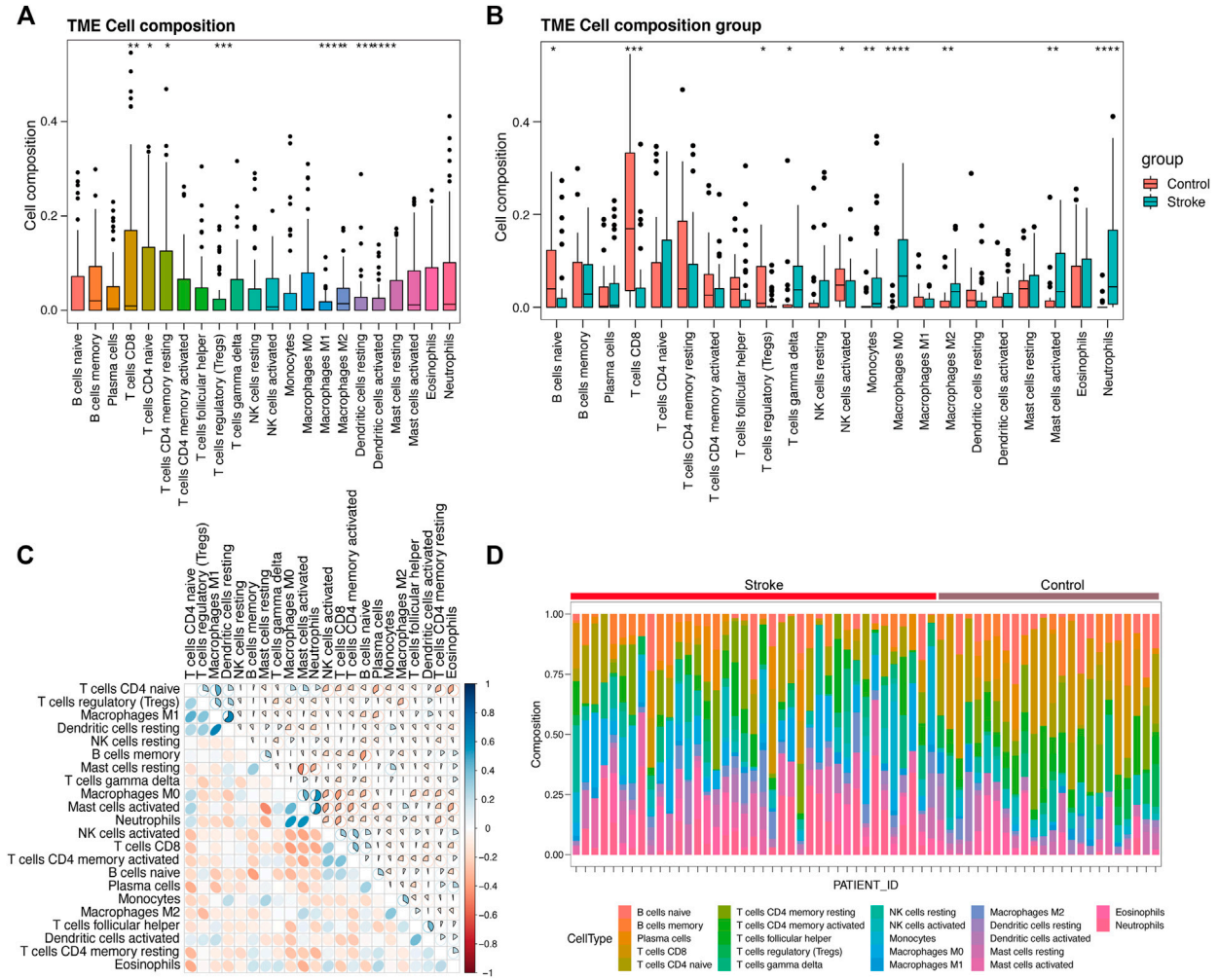


FIGURE 11 | Evaluation and visualization of immune cell infiltration. **(A)** Ungrouped immune cell infiltration map. **(B)** Immune cell infiltration map between acute cerebral infarction group and control group. **(C)** Correlation heat map of 22 types of immune cell infiltration. Blue and red indicate positive and negative correlations, respectively. The darker the color, the stronger is the correlation. **(D)** Immune cell infiltration map between a single sample of acute cerebral infarction group and control group. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

Using GEO, differentially expressed transcripts, including mRNAs, miRNAs and lncRNAs, were identified and the ceRNA network in AIS was constructed. In the ceRNA network, AL360004.1 (degree = 5), hsa-miR-125a-5p (degree = 16), hsa-miR-125b-5p (degree = 16) and KRT10 (degree = 5) were the molecules with the highest connection score. Previous studies have correlated cancer-related functions with AL360004.1, LINC00173, LINC01089 and LINC00115; however, none of them have been linked to the pathogenesis of AIS. LINC00115 expression levels correlate with prognosis in human bladder (Jiang et al., 2018) and lung cancer (Li et al., 2016) patients. A study demonstrated that LINC00115, a novel miRNA sponge of the miR-200 family, can promote ZEB1 signaling in GBM (Tang et al., 2019). Deletion or ectopic expression of LINC00115 affects ZEB1 signaling, neuro-like sphere formation *in vitro*, and animal survival time *in vivo* (Tang et al., 2019). A study reported that LINC00173 plays an important oncogenic role in glioma by activating the miR-765/NUTF2 pathway (Du et al., 2020). And silencing of LINC00173. V1 attenuates vascular endothelial cell proliferation and migration (Chen J. et al., 2020). Thus, they represent a potential novel biological markers for AIS diagnosis and therapy. KRT10 is an intermediate filament (IF) protein belonging to the type I (acidic) cytokeratin family. Keratin expression may affect cell proliferation and differentiation (Paladini and Coulombe, 1998; Paramio et al., 1999). KRT10 impairs cell cycle progression through isolating and inhibiting protein kinase B (PKB; Akt) and atypical PKC, the key effectors of the phosphatidylinositol 3-kinase (PI3K) pathway (Paramio et al., 2001). PI3K pathway plays a central role in neuronal survival. Therefore, KRT10 may also be a potential therapeutic targets for AIS. MiR-125a and miR-125b belong to the same miRNA family, which has identical 'seed sequence'. MiR-125a-5p and miR-125b-5p both have been shown to inhibit angiogenesis (Banerjee et al., 2013; Pan et al., 2015). Overexpression of miR-125a-5p promotes

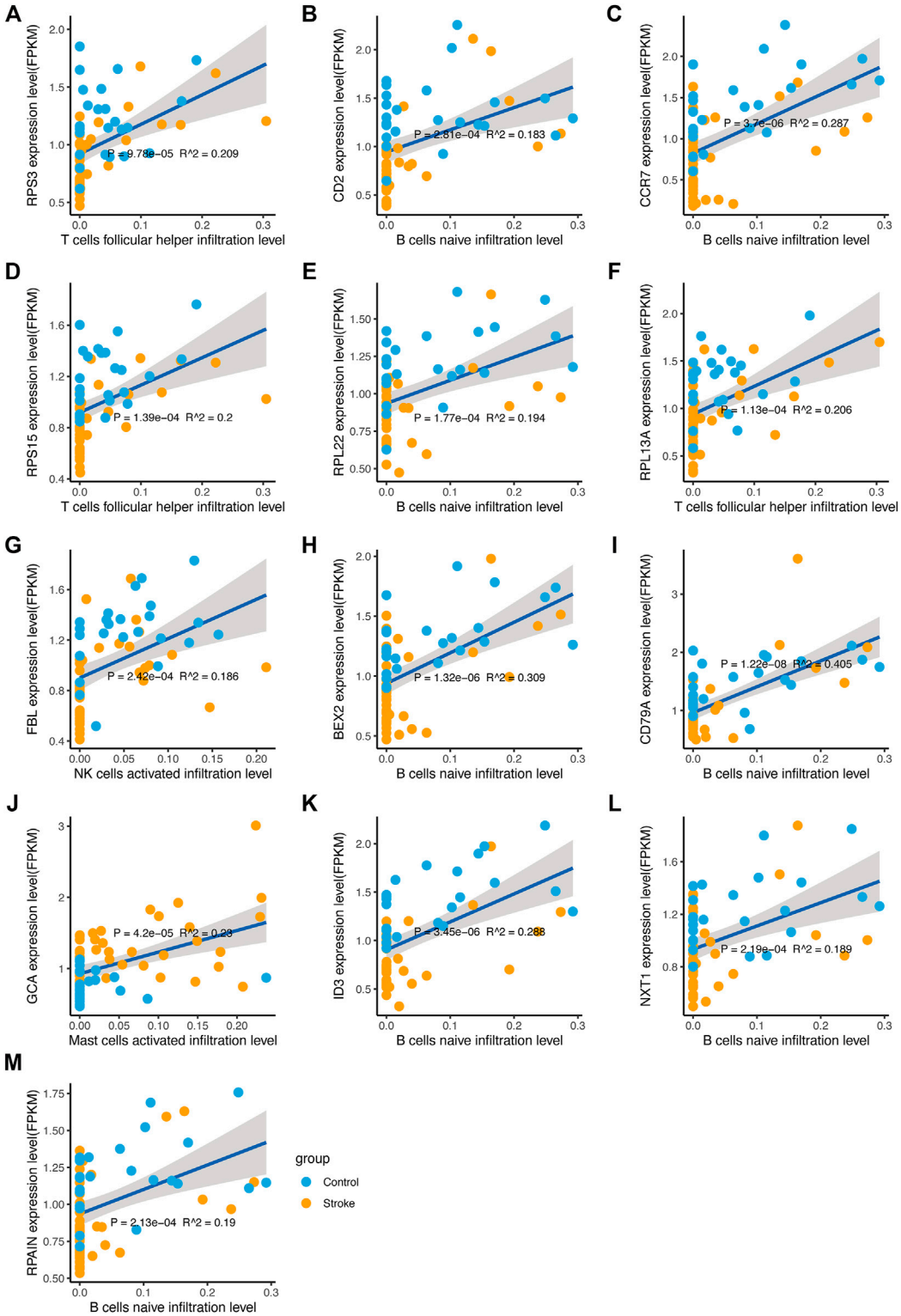


FIGURE 12 | Correlation between diagnostic markers and immune cell infiltration. **(A–M)** The linear regression of diagnostic markers and immune cell infiltration level. The horizontal axis indicates the immune cell infiltration level, and the vertical axis indicates the marker expression. The *p* value is the regression significance level, and R^2 is the goodness-of-fit.

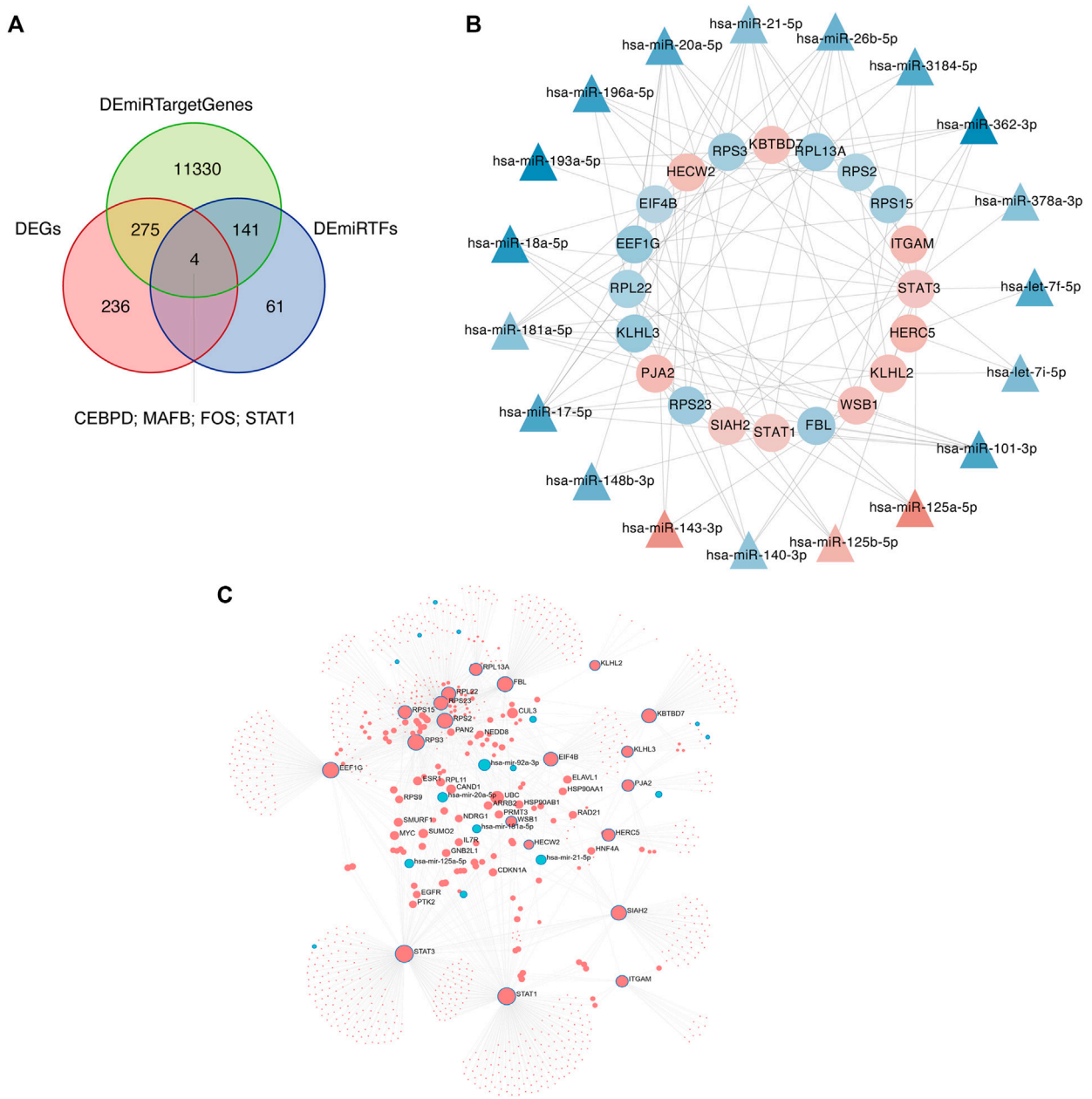


FIGURE 13 | Network analyses of target genes and transcription factors. **(A)** The intersection of differentially expressed genes (DEGs), differential miRNA target genes, and transcription factors is determined using Venn diagram analysis. **(B)** Intersection molecules-miRNA network analysis. The inner ring is the intersection of DEGs and DEmiRTargetGenes, and the outer ring is miRNAs. Red indicates increased expression, blue indicates reduced expression, and color intensity indicates different degrees of upregulation or downregulation. **(C)** Network diagram of differential lncRNAs, miRNAs, and mRNAs. Red indicates upregulated expression; blue indicates downregulated expression.

nitric oxide (NO) production, reduces ROS production, and delays human brain microvascular endothelial cells (HBMECs) senescence through the PI3K/AKT/eNOS signaling pathway (Pan et al., 2017). MiR-125b-5p regulates synaptic morphology and function (Edbauer et al., 2010). MiR-125a-5p is associated with colorectal and liver cancer diseases (Chen L.-Y. et al., 2020; Zhou et al., 2021). Acute kidney injury and triple-negative breast cancer are associated with miR-125b-5p (Lv et al., 2021). According to a report, MiR-148b plays multiple roles in the development of various biological processes (Friedrich et al., 2017). A study confirmed the role of miR-148b in the modulation of proliferation and differentiation of neural stem cell after ischemic stroke (Wang et al., 2017). Therefore, we determined that this group of miRNAs is a promising diagnostic marker. Seventeen DEGs were highly connected as the most significant hub genes in the PPI network and there are multiple interactions

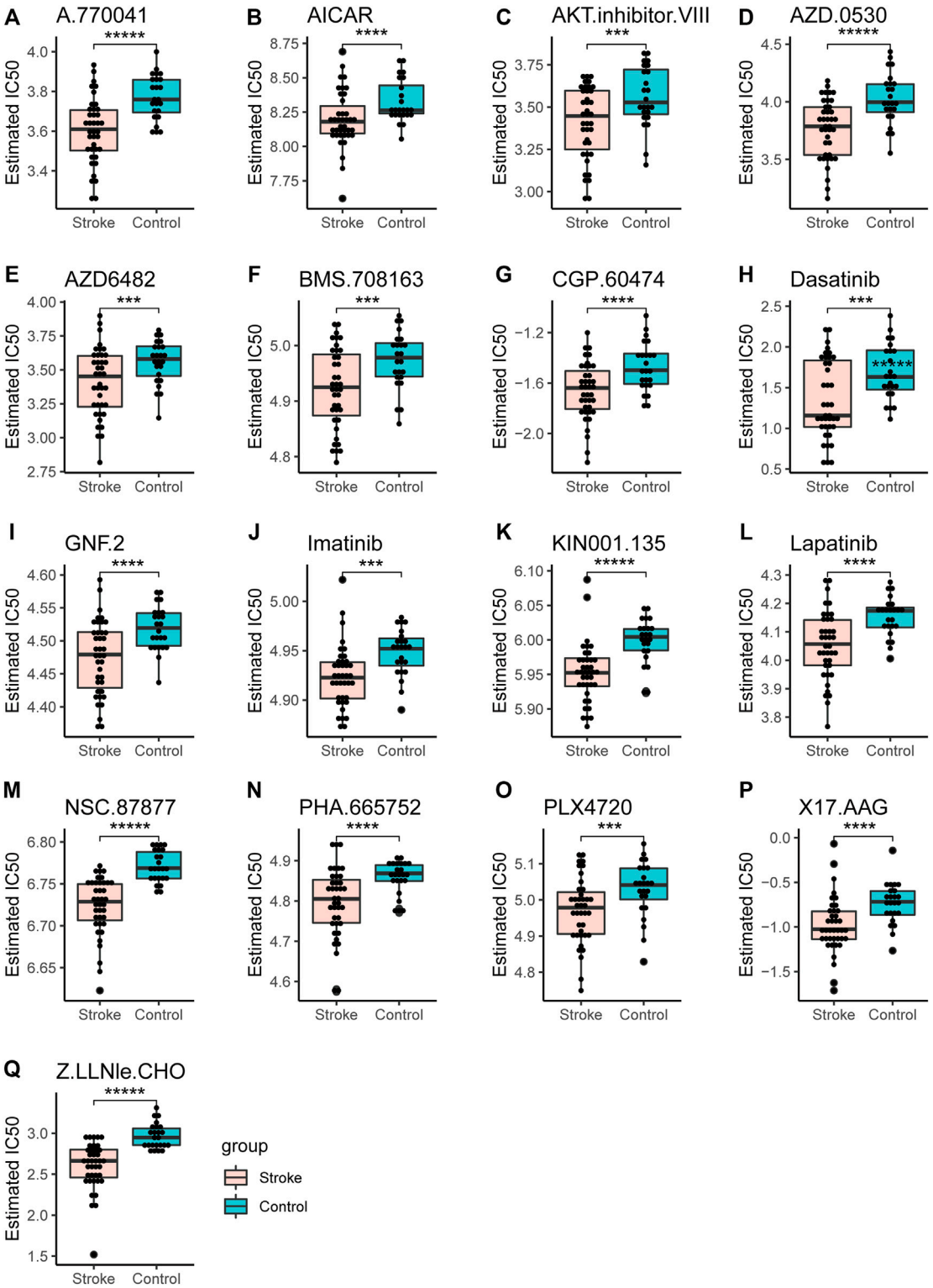


FIGURE 14 | Drug sensitivity analysis. The IC50 values of different drugs were determined in the control group and the acute cerebral infarction group. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$, *****$p < 0.00001$.

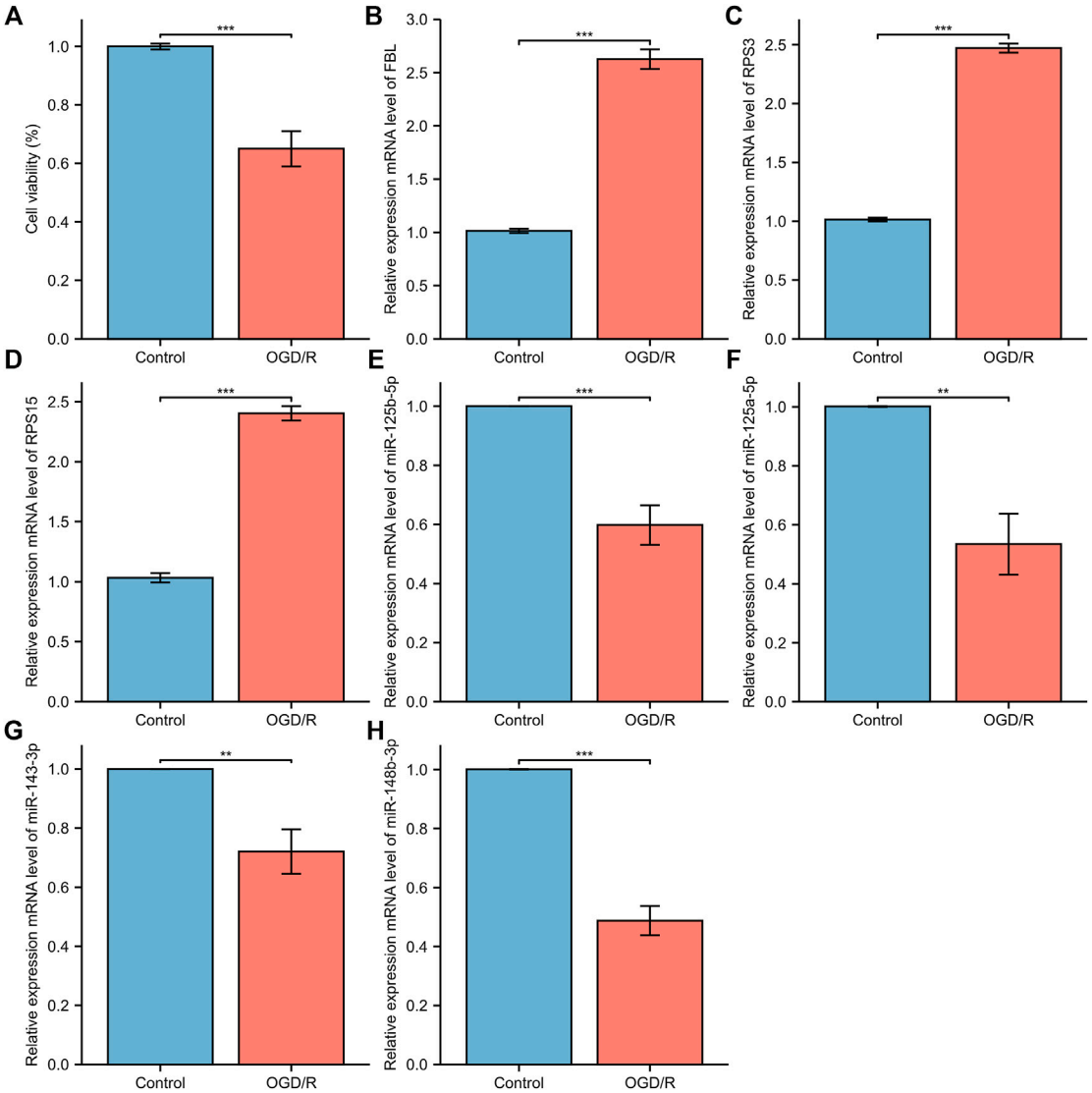


FIGURE 15 | The relative expression of differentially expressed mRNA and miRNA in HT22. **(A)** The CCK-8 cell viability assay; **(B)** FBL **(C)** RPS3; **(D)** RPS15 **(E)** miRNA-143-3p **(F)** miRNA-148b-3p; **(G)** miRNA-125b-5p; **(H)** miRNA-125a-5p. The control group reflects the normal HT22 and the OGE/R group reflects the model group. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

in the network. Moreover, FBL, RPS3 and RPS15 were ranked as the top three proteins of the most potentially serving as key regulators in AIS. Both RPS3 and RPS15 encode a ribosomal protein, which is part of the 40S subunit. Interestingly, GO and KEGG analysis, as well as GSEA and GSVA, showed that ribosome pathway was closely related to the increased incidence of ischemic stroke. RPS3 induces neuronal apoptosis by interacting with the E2F1 transcription factor and inducing the expression of pro-apoptotic proteins BCL2L11/BIM and HRK/Dp5 (Lee et al., 2010). When located in the mitochondria, RPS3 reduces cellular ROS levels and mitochondrial DNA damage (Kim et al., 2013). Diseases associated with RPS3 include Eumycotic Mycetoma and Schopf-Schulz-Passarge Syndrome. A study shows that RPS15 is a critical morbific leucine-rich repeat kinase 2 (LRRK2) substrate in Parkinson's disease models of *drosophila* and human neuron (Martin et al., 2014). Phosphorylation of RPS15 is related to LRRK2 neurodegeneration and neurotoxicity (Martin et al., 2014), suggesting that the genes encoding ribosomal proteins may be potential targets and treatment for early diagnosis of AIS. The components of small nucleolar ribonucleoprotein (snRNP) particles include FBL products, which are required for ribosomal RNAs processing and modification. FBL interacts with small nucleolar RNAs (snoRNAs) and ribosomal RNA (rRNA) to modify the mRNA 2′-O-methylation, thereby regulating ROS and oxidative reactions (Elliott et al., 2019). Diseases associated with FBL include Diffuse Scleroderma and Systemic Scleroderma. In terms of diagnostic value, the AUC of

these three mRNAs and four miRNAs' genes was analyzed. All the AUC values were in the range 0.866–0.989, suggesting that these genes had moderate predictor performances (Akobeng, 2007) in diagnostic examinations. However, in our investigation the qPCR results showed that miR-125a-5p, miR-125b-5p, and miR-143-3p levels in the model control group were significantly downregulated. Thus, these results drive us to explore whether these miRNAs will also play a protective role in AIS when cells viability is around 65%. The validation of identified DEGs and DEmiRs was performed using qPCR, and confirmed the above results, proof of the diagnostic effectiveness of the DEGs and DEMirs.

It was further confirmed that the infiltration levels of Macrophages M0, activated Mast cells, and Monocytes in acute cerebral infarction samples were higher than those in control samples. Contrarily, the infiltration levels of T cells CD8, B cells naïve, and activated NK cells in control samples significantly higher than those in AIS samples. Macrophages infiltrate and promote rapid inflammatory responses in the acute phase of AIS; however, T cells in the late phases of cerebral infarction (Iadecola and Anrather, 2011). Animal models of ischemic stroke results in increased number of activated mast cells (Bot et al., 2020), and mast cells are also involved in arteriogenesis and collateral formation (Chillo et al., 2016). Activation of mast cells plays a proinflammatory role by recruiting immune cells such as neutrophils and monocytes (Bot et al., 2007; Sun et al., 2007; Chillo et al., 2016). After identifying differential expression of hub-genes, we analyzed the correlation between the hub genes of AIS and the level of immune infiltration. It has been suggested that peripheral immune cells such as neutrophils to infiltrate in the ischemic brain region after disruption of the blood-brain barrier (BBB) in ischemic stroke (Qian et al., 2016). In addition, elevated expression level of hub-genes significantly correlated with T cells follicular helper, B cells naive, NK cells, and Mast cells infiltration ($p < 0.05$), facilitating a general increase the levels of infiltrating immune cells.

Four transcription factors were obtained by taking the intersection of DEmRNA, DemiRTargetGenes, and DEmiRTFs:CEBPD, MAFB, FOS, and STAT1. CEBPD is an important TF that regulates the expression of multiple genes and participates in immune and inflammatory responses (Wang et al., 2006). MAFB avoids excess inflammation after ischemic stroke (Shichita et al., 2017). FOS plays a s crucial role in post ischemic inflammation and cell death (Chung, 2015). STAT1 is activated by ROS and contributes to ischemic injury (Takagi et al., 2002). AZD0530 is a small molecule inhibitor of Src family kinases under investigation. Moreover, orally administered AZD0530 is highly CNS penetrable in both mice and humans (Kaufman et al., 2015). Recent a study shows that AZD0530 rescues deficits in memory and restores synapse density in transgenic mouse Alzheimer disease models (van Dyck et al., 2019). Elevated or stable Notch levels can promote neuronal death in ischemic stroke. However, NOTCH signaling pathway is inhibited through a gamma-secretase inhibitor (GSI) (Xu et al., 2016). On the other hand, GSIs have been used for the treatment of Alzheimer's disease to prevent the cleavage of amyloid precursor protein and the subsequent release of amyloid β peptide. Therefore, this suggests that AZD0530 and GSI-I (Z.LLNle.CHO) can be used as therapeutic agents in models of Alzheimer's disease. Whether they can be used as a therapeutic agent for cerebral infarction remains unclear. NSC-87877 is an effective Shp2 inhibitor, but it has a similar inhibitory effect on Shp1. Some studies have reported that NSC-87877 is a potential new treatment for relapsing-remitting multiple sclerosis, MuSK antibody positive myasthenia gravis (MuSK-MG) (Huda et al., 2020) and intracerebral hemorrhage (Liu et al., 2019). In addition, Yinlong et al. reported that NSC 87877 treatment attenuated ICH-induced apoptosis and neuronal death (Liu et al., 2019). In this regard, we speculated whether NSC-87877 could not only promote cerebral neovascularization and the brain vascular restoration after stroke but also exert neuroprotective effects.

Our study presents few limitations. First, a comprehensive analysis of warranted venous blood samples and brain tissue was not performed in this study; however, it is necessary to comprehensively diagnose the dysfunctions in acute ischemic stroke. Second, the study includes a relatively small cohort and, therefore, some of the data failed to reach statistical significance. To determine better accuracy and validation of the hub genes associated with AIS, a larger sample size for further external validation is needed. Third, the results should be further verified by western blot (WB), real-time PCR and immunofluorescence assays. Further experiments are clearly warranted to fully elucidate the role of hub genes and the underlying mechanisms of acute ischemic stroke. Fourth, to investigate potential function and mechanisms related to DEGs and hub genes in AIS, the study of the cell or tissue-type specific gain-of-function and loss-of-function still needs to be performed. Signaling pathways are more diverse in AIS than previously thought, such as Toll-like receptor 4 (TLR4)/NF-κB/NLRP3 signaling pathway, T cell receptor signaling pathway, and PPAR-γ signaling pathway. Although previous studies have identified several signaling pathways, more detail experimental evidence is still needed to improve our understanding of the possible phenotype and pathway regulation of these predicted genes in AIS. Fifth, ceRNA network and interaction among hub genes will be an exciting new field to explore and will shed new light on ischemic cerebrovascular disease. Co-Immunoprecipitation and pull-down assays would provide strong support to the proposed mechanism. Further investigation is needed to explore the intermolecular interactions responsible for the molecular cooperativity in the progression of cerebral infarctions, such as the ribosomal protein family (C and D). Moreover, the contribution of the identified DEGs into the pathogenesis of AIS should be examined in detail. Therefore, it may be necessary to test the efficacy of activation and inactivation in more experiments to molecular interactions. These are very valuable in understanding the mechanisms of protein-protein interactions. Sixth, ArenaIdb database (Bonnici et al., 2018) truly integrates the content of starbase, mircode with other datasets that the authors have not used. We do not currently do analysis of other datasets using arenaidb. In the follow-up research, we will supplement the validation of other datasets.

In conclusions, this study identified several pathways and biomarkers in AIS consistent with current knowledge of the pathology of this disease. We believe that new insights are

provided on the molecular mechanisms underlying the pathogenesis of AIS.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

ZW, HW contributed to conception and design of the study. ZW analysed the database. ZW performed the statistical analysis. ZW performed the experiment. ZW, WW reviewed relevant literature and drafted the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by Grants from the National Natural Science Funds for Young Scholars of China (Grant NO. 81801312), and the Hebei Province Medical Research Project of China (Grant No. 20200715).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.833545/full#supplementary-material

REFERENCES

Adams, H. P., Jr., Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L., et al. (1993). Classification of Subtype of Acute Ischemic Stroke. Definitions for Use in a Multicenter Clinical Trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke* 24 (1), 35–41. doi:10.1161/01.str.24.1.35

Akobeng, A. K. (2007). Understanding Diagnostic Tests 3: Receiver Operating Characteristic Curves. *Acta Paediatr.* 96 (5), 644–647. doi:10.1111/j.1651-2227.2006.00178.x

Allen, C., Thornton, P., Denes, A., McColl, B. W., Pierozynski, A., Monestier, M., et al. (2012). Neutrophil Cerebrovascular Transmigration Triggers Rapid Neurotoxicity through Release of Proteases Associated with Decondensed DNA. *J. Immunol.* 189 (1), 381–392. doi:10.4049/jimmunol.1200409

Banerjee, S., Cui, H., Xie, N., Tan, Z., Yang, S., Icyuz, M., et al. (2013). miR-125a-5p Regulates Differential Activation of Macrophages and Inflammation. *J. Biol. Chem.* 288 (49), 35428–35436. doi:10.1074/jbc.M112.426866

Barr, T. L., Conley, Y., Ding, J., Dillman, A., Warach, S., Singleton, A., et al. (2010). Genomic Biomarkers and Cellular Pathways of Ischemic Stroke by RNA Gene Expression Profiling. *Neurology* 75 (11), 1009–1014. doi:10.1212/WNL.0b013e3181f2b37f

Benakis, C., Brea, D., Caballero, S., Faraco, G., Moore, J., Murphy, M., et al. (2016). Commensal Microbiota Affects Ischemic Stroke Outcome by Regulating Intestinal γδ T Cells. *Nat. Med.* 22 (5), 516–523. doi:10.1038/nm.4068

Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., et al. (2018). Heart Disease and Stroke Statistics-2018 Update: A Report from the American Heart Association. *Circulation* 137 (12), e67–e492. doi:10.1161/CIR.0000000000000558

Bonnici, V., Caro, G. D., Constantino, G., Liuni, S., D'Elia, D., Bombieri, N., et al. (2018). Arena-Idb: A Platform to Build Human Non-coding RNA Interaction Networks. *BMC Bioinformatics* 19 (Suppl. 10), 350. doi:10.1186/s12859-018-2298-8

Bot, I., de Jager, S. C. A., Zernecke, A., Lindstedt, K. A., van Berkel, T. J. C., Weber, C., et al. (2007). Perivascular Mast Cells Promote Atherogenesis and Induce Plaque Destabilization in Apolipoprotein E-Deficient Mice. *Circulation* 115 (19), 2516–2525. doi:10.1161/CIRCULATIONAHA.106.660472

Bot, I., van der Velden, D., Bouwman, M., Kröner, M. J., Kuiper, J., Quax, P. H. A., et al. (2020). Local Mast Cell Activation Promotes Neovascularization. *Cells* 9 (3), 701. doi:10.3390/cells9030701

Carloni, S., Girelli, S., Scopa, C., Buonocore, G., Longini, M., and Balduini, W. (2010). Activation of Autophagy and Akt/CREB Signaling Play an Equivalent Role in the Neuroprotective Effect of Rapamycin in Neonatal Hypoxia-Ischemia. *Autophagy* 6 (3), 366–377. doi:10.4161/auto.6.3.11261

Chen, J., Liu, A., Wang, Z., Wang, B., Chai, X., Lu, W., et al. (2020a). LINC00173.v1 Promotes Angiogenesis and Progression of Lung Squamous Cell Carcinoma by Sponging miR-511-5p to Regulate VEGFA Expression. *Mol. Cancer* 19 (1), 98. doi:10.1186/s12943-020-01217-2

Chen, L.-Y., Wang, L., Ren, Y.-X., Pang, Z., Liu, Y., Sun, X.-D., et al. (2020b). The Circular RNA Circ-ERBIN Promotes Growth and Metastasis of Colorectal Cancer by miR-125a-5p and miR-138-5p/4EBP-1 Mediated Cap-independent HIF-1α Translation. *Mol. Cancer* 19 (1), 164. doi:10.1186/s12943-020-01272-9

Chen, W., Guo, C., Feng, H., and Chen, Y. (2020c). Mitochondria: Novel Mechanisms and Therapeutic Targets for Secondary Brain Injury after Intracerebral Hemorrhage. *Front. Aging Neurosci.* 12, 615451. doi:10.3389/fnagi.2020.615451

Chillo, O., Kleinert, E. C., Lautz, T., Lasch, M., Pagel, J.-I., Heun, Y., et al. (2016). Perivascular Mast Cells Govern Shear Stress-Induced Arteriogenesis by Orchestrating Leukocyte Function. *Cel Rep.* 16 (8), 2197–2207. doi:10.1016/j.celrep.2016.07.040

Chung, L. (2015). A Brief Introduction to the Transduction of Neural Activity into Fos Signal. *Dev. Reprod.* 19 (2), 61–67. doi:10.12717/dr.2015.19.2.061

Davis, S., and Meltzer, P. S. (2007). GEOquery: A Bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23 (14), 1846–1847. doi:10.1093/bioinformatics/btm254

Dennis, M., Mead, G., Forbes, J., Graham, C., Hackett, M., Hankey, G. J., et al. (2019). Effects of Fluoxetine on Functional Outcomes after Acute Stroke (FOCUS): A Pragmatic, Double-Blind, Randomised, Controlled Trial. *Lancet* 393 (10168), 265–274. doi:10.1016/s0140-6736(18)32823-x

Dirnagl, U., Moskowitz, M. A., and Iadecola, C. (1999). Pathobiology of Ischaemic Stroke: An Integrated View. *Trends Neurosciences* 22, 391–397. doi:10.1016/s0166-2236(99)01401-0

Doll, D., Barr, T. L., and Simpkins, J. W. (2014). Cytokines: Their Role in Stroke and Potential Use as Biomarkers and Therapeutic Targets. *Aging Dis.* 5 (5), 294–306. doi:10.14336/AD.2014.0500294

Du, Q., Liu, J., Tian, D., Zhang, X., Zhu, J., Qiu, W., et al. (2020). Long Noncoding RNA LINC00173 Promotes NUTF2 Expression through Sponging miR-765 and Facilitates Tumorigenesis in Glioma. *Cancer Manag. Res.* 12, 7211–7217. doi:10.2147/CMAR.S262279

Edbauer, D., Neilson, J. R., Foster, K. A., Wang, C.-F., Seeburg, D. P., Batterton, M. N., et al. (2010). Regulation of Synaptic Structure and Function by FMRP-Associated microRNAs miR-125b and miR-132. *Neuron* 65 (3), 373–384. doi:10.1016/j.neuron.2010.01.005

Elliott, B. A., Ho, H.-T., Ranganathan, S. V., Vangaveti, S., Ilkayeva, O., Abou Assi, H., et al. (2019). Modification of Messenger RNA by 2′-O-Methylation Regulates Gene Expression *In Vivo*. *Nat. Commun.* 10 (1), 3401. doi:10.1038/s41467-019-11375-7

Fatahzadeh, M., and Glick, M. (2006). Stroke: Epidemiology, Classification, Risk Factors, Complications, Diagnosis, Prevention, and Medical and Dental Management. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodontology* 102 (2), 180–191. doi:10.1016/j.tripleo.2005.07.031

Forrester, S. J., Kikuchi, D. S., Hernandes, M. S., Xu, Q., and Griendling, K. K. (2018). Reactive Oxygen Species in Metabolic and Inflammatory Signaling. *Circ. Res.* 122 (6), 877–902. doi:10.1161/CIRCRESAHA.117.311401

Friedrich, M., Pracht, K., Mashreghi, M.-F., Jäck, H.-M., Radbruch, A., and Seliger, B. (2017). The Role of the miR-148/-152 Family in Physiology and Disease. *Eur. J. Immunol.* 47 (12), 2026–2038. doi:10.1002/eji.201747132

Friendly, M. (2002). Corrgrams. *The Am. Statistician* 56 (4), 316–324. doi:10.1198/000313002533

Guo, D., Ma, J., Yan, L., Li, T., Li, Z., Han, X., et al. (2017). Down-Regulation of Lncrna MALAT1 Attenuates Neuronal Cell Death through Suppressing Beclin1-Dependent Autophagy by Regulating Mir-30a in Cerebral Ischemic Stroke. *Cell Physiol Biochem* 43 (1), 182–194. doi:10.1159/000480337

Hu, X., Leak, R. K., Shi, Y., Suenaga, J., Gao, Y., Zheng, P., et al. (2015). Microglial and Macrophage Polarization-New Prospects for Brain Repair. *Nat. Rev. Neurol.* 11 (1), 56–64. doi:10.1038/nrneurol.2014.207

Huda, S., Cao, M., De Rosa, A., Woodhall, M., Rodriguez Cruz, P. M., Cossins, J., et al. (2020). SHP2 Inhibitor Protects AChRs from Effects of Myasthenia Gravis MuSK Antibody. *Neurol. Neuroimmunol Neuroinflamm* 7 (1), e645. doi:10.1212/NXI.0000000000000645

Iadecola, C., and Anrather, J. (2011). The Immunology of Stroke: From Mechanisms to Translation. *Nat. Med.* 17 (7), 796–808. doi:10.1038/nm.2399

Iadecola, C., Buckwalter, M. S., and Anrather, J. (2020). Immune Responses to Stroke: Mechanisms, Modulation, and Therapeutic Potential. *J. Clin. Invest.* 130 (6), 2777–2788. doi:10.1172/JCI135530

Jeggari, A., Marks, D. S., and Larsson, E. (2012). miRcode: A Map of Putative microRNA Target Sites in the Long Non-coding Transcriptome. *Bioinformatics* 28 (15), 2062–2063. doi:10.1093/bioinformatics/bts344

Jiang, B., Hailong, S., Yuan, J., Zhao, H., Xia, W., Zha, Z., et al. (2018). Identification of Oncogenic Long Noncoding RNA SNHG12 and DUXAP8 in Human Bladder Cancer through a Comprehensive Profiling Analysis. *Biomed. Pharmacother.* 108, 500–507. doi:10.1016/j.biopha.2018.09.025

Karagkouni, D., Paraskevopoulou, M. D., Tastsoglou, S., Skoufos, G., Karavangeli, A., Pierros, V., et al. (2020). DIANA-LncBase V3: Indexing Experimentally Supported miRNA Targets on Non-Coding Transcripts. *Nucleic Acids Res.* 48 (D1), D101–D110. doi:10.1093/nar/gkz1036

Kaufman, A. C., Salazar, S. V., Haas, L. T., Yang, J., Kostylev, M. A., Jeng, A. T., et al. (2015). Fyn Inhibition Rescues Established Memory and Synapse Loss in Alzheimer Mice. *Ann. Neurol.* 77 (6), 953–971. doi:10.1002/ana.24394

Kim, Y., Kim, H. D., and Kim, J. (2013). Cytoplasmic Ribosomal Protein S3 (rpS3) Plays a Pivotal Role in Mitochondrial DNA Damage Surveillance. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1833 (12), 2943–2952. doi:10.1016/j.bbamcr.2013.07.015

Koike, M., Shibata, M., Tadakoshi, M., Gotoh, K., Komatsu, M., Waguri, S., et al. (2008). Inhibition of Autophagy Prevents Hippocampal Pyramidal Neuron Death after Hypoxic-Ischemic Injury. *Am. J. Pathol.* 172 (2), 454–469. doi:10.2353/ajpath.2008.070876

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Soft.* 25 (1), 1–18. doi:10.18637/jss.v025.i01

Lee, S. B., Kwon, I.-S., Park, J., Lee, K.-H., Ahn, Y., Lee, C., et al. (2010). Ribosomal Protein S3, a New Substrate of Akt, Serves as a Signal Mediator between Neuronal Apoptosis and DNA Repair. *J. Biol. Chem.* 285 (38), 29457–29468. doi:10.1074/jbc.M110.131367

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The Sva Package for Removing Batch Effects and Other Unwanted Variation in High-Throughput Experiments. *Bioinformatics* 28 (6), 882–883. doi:10.1093/bioinformatics/bts034

Li, D. S., Ainiwaer, J. L., Sheyhiding, I., Zhang, Z., and Zhang, L. W. (2016). Identification of Key Long Non-coding RNAs as Competing Endogenous RNAs for miRNA-mRNA in Lung Adenocarcinoma. *Eur. Rev. Med. Pharmacol. Sci.* 20 (11), 2285–2295.

Li, H., Tang, C., and Wang, D. (2020). LncRNA H19 Promotes Inflammatory Response Induced by Cerebral Ischemia-Reperfusion Injury through Regulating the miR-138-5p-P65 axis. *Biochem. Cel Biol.* 98 (4), 525–536. doi:10.1139/bcb-2019-0281

Li, R., Qu, H., Wang, S., Wei, J., Zhang, L., Ma, R., et al. (2018). GDCRNATools: an R/Bioconductor Package for Integrative Analysis of lncRNA, miRNA and mRNA Data in GDC. *Bioinformatics* 34 (14), 2515–2517. doi:10.1093/bioinformatics/bty124

Liu, Y., Ma, C., Li, H., Shen, H., Li, X., Fu, X. a., et al. (2019). Nogo-A/Pir-B/TrkB Signaling Pathway Activation Inhibits Neuronal Survival and Axonal Regeneration after Experimental Intracerebral Hemorrhage in Rats. *J. Mol. Neurosci.* 69 (3), 360–370. doi:10.1007/s12031-019-01365-1

Lv, Y., Lv, X., Yang, H., Qi, X., Wang, X., Li, C., et al. (2021). LncRNA SNHG6/miR-125b-5p/BMPR1B Axis: A New Therapeutic Target for Triple-Negative Breast Cancer. *Front. Oncol.* 11, 678474. doi:10.3389/fonc.2021.678474

Martin, I., Kim, J. W., Lee, B. D., Kang, H. C., Xu, J.-C., Jia, H., et al. (2014). Ribosomal Protein S15 Phosphorylation Mediates LRRK2 Neurodegeneration in Parkinson's Disease. *Cell* 157 (2), 472–485. doi:10.1016/j.cell.2014.01.064

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Paladini, R. D., and Coulombe, P. A. (1998). Directed Expression of Keratin 16 to the Progenitor Basal Cells of Transgenic Mouse Skin Delays Skin Maturation. *J. Cel Biol* 142 (4), 1035–1051. doi:10.1083/jcb.142.4.1035

Pan, Q., Liao, X., Liu, H., Wang, Y., Chen, Y., Zhao, B., et al. (2017). MicroRNA-125a-5p Alleviates the Deleterious Effects of Ox-LDL on Multiple Functions of Human Brain Microvessel Endothelial Cells. *Am. J. Physiology-Cell Physiol.* 312 (2), C119–C130. doi:10.1152/ajpcell.00296.2016

Pan, W., Zhu, S., Dai, D., Liu, Z., Li, D., Li, B., et al. (2015). MiR-125a Targets Effector Programs to Stabilize Treg-Mediated Immune Homeostasis. *Nat. Commun.* 6, 7096. doi:10.1038/ncomms8096

Paramio, J. M., Casanova, M. L., Segrelles, C., Mittnacht, S., Lane, E. B., and Jorcano, J. L. (1999). Modulation of Cell Proliferation by Cytokeratins K10 and K16. *Mol. Cel Biol* 19 (4), 3086–3094. doi:10.1128/mcb.19.4.3086

Paramio, J. M., Segrelles, C., Ruiz, S., and Jorcano, J. L. (2001). Inhibition of Protein Kinase B (PKB) and PKCζ Mediates Keratin K10-Induced Cell Cycle Arrest. *Mol. Cel Biol* 21 (21), 7449–7459. doi:10.1128/mcb.21.21.7449-7459.2001

Pathan, M., Keerthikumar, S., Ang, C.-S., Gangoda, L., Quek, C. Y. J., Williamson, N. A., et al. (2015). FunRich: An Open Access Standalone Functional Enrichment and Interaction Network Analysis Tool. *Proteomics* 15 (15), 2597–2601. doi:10.1002/pmic.201400515

Perez-de-Puig, I., Miró-Mur, F., Ferrer-Ferrer, M., Gelpi, E., Pedragosa, J., Justicia, C., et al. (2015). Neutrophil Recruitment to the Brain in Mouse and Human Ischemic Stroke. *Acta Neuropathol.* 129 (2), 239–257. doi:10.1007/s00401-014-1381-0

Qi, X., Shao, M., Sun, H., Shen, Y., Meng, D., and Huo, W. (2017). Long Non-Coding RNA SNHG14 Promotes Microglia Activation by Regulating miR-145-5p/PLA2G4A in Cerebral Infarction. *Neuroscience* 348, 98–106. doi:10.1016/j.neuroscience.2017.02.002

Qian, L., Yuanshao, L., Wensi, H., Yulei, Z., Xiaoli, C., Brian, W., et al. (2016). Serum IL-33 Is a Novel Diagnostic and Prognostic Biomarker in Acute Ischemic Stroke. *Aging Dis.* 7 (5), 614–622. doi:10.14336/AD.2016.0207

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43 (7), e47. doi:10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinformatics* 12, 77. doi:10.1186/1471-2105-12-77

Saenger, A. K., and Christenson, R. H. (2010). Stroke Biomarkers: Progress and Challenges for Diagnosis, Prognosis, Differentiation, and Treatment. *Clin. Chem.* 56 (1), 21–33. doi:10.1373/clinchem.2009.133801

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Shiau, C. E., Monk, K. R., Joo, W., and Talbot, W. S. (2013). An Anti-Inflammatory NOD-Like Receptor Is Required for Microglia Development. *Cel Rep.* 5 (5), 1342–1352. doi:10.1016/j.celrep.2013.11.004

Shichita, T., Ito, M., Morita, R., Komai, K., Noguchi, Y., Ooboshi, H., et al. (2017). MAFB Prevents Excess Inflammation after Ischemic Stroke by Accelerating Clearance of Damage Signals through MSR1. *Nat. Med.* 23 (6), 723–732. doi:10.1038/nm.4312

Sun, J., Sukhova, G. K., Wolters, P. J., Yang, M., Kitamoto, S., Libby, P., et al. (2007). Mast Cells Promote Atherosclerosis by Releasing Proinflammatory Cytokines. *Nat. Med.* 13 (6), 719–724. doi:10.1038/nm1601

Sun, Y., Wang, J., Han, B., Meng, K., Han, Y., and Ding, Y. (2021). Elucidating the Molecular Mechanism of Ischemic Stroke Using Integrated Analysis of miRNA, mRNA, and lncRNA Expression Profiles. *Front. Integr. Neurosci.* 15, 638114. doi:10.3389/fnint.2021.638114

Takagi, Y., Harada, J., Chiarugi, A., and Moskowitz, M. A. (2002). STAT1 Is Activated in Neurons after Ischemia and Contributes to Ischemic Brain Injury. *J. Cereb. Blood Flow Metab.* 22 (11), 1311–1318. doi:10.1097/01.WCB.0000034148.72481.F4

Tang, J., Yu, B., Li, Y., Zhang, W., Alvarez, A. A., Hu, B., et al. (2019). TGF-β-Activated lncRNA LINC00115 Is a Critical Regulator of Glioma Stem-Like Cell Tumorigenicity. *EMBO Rep.* 20 (12), e48170. doi:10.15252/embr.201948170

Tiedt, S., Prestel, M., Malik, R., Schieferdecker, N., Duering, M., Kautzky, V., et al. (2017). RNA-Seq Identifies Circulating miR-125a-5p, miR-125b-5p, and miR-143-3p as Potential Biomarkers for Acute Ischemic Stroke. *Circ. Res.* 121 (8), 970–980. doi:10.1161/CIRCRESAHA.117.311572

van Dyck, C. H., Nygaard, H. B., Chen, K., Donohue, M. C., Raman, R., Rissman, R. A., et al. (2019). Effect of AZD0530 on Cerebral Metabolic Decline in Alzheimer Disease: A Randomized Clinical Trial. *JAMA Neurol.* 76 (10), 1219–1229. doi:10.1001/jamaneurol.2019.2050

Venkat, P., Chen, J., and Chopp, M. (2018). Exosome-Mediated Amplification of Endogenous Brain Repair Mechanisms and Brain and Systemic Organ Interaction in Modulating Neurological Outcome after Stroke. *J. Cereb. Blood Flow Metab.* 38 (12), 2165–2178. doi:10.1177/0271678X18782789

Walter, W., Sánchez-Cabo, F., and Ricote, M. (2015). GOplot: An R Package for Visually Combining Expression Data with Functional Analysis. *Bioinformatics* 31 (17), 2912–2914. doi:10.1093/bioinformatics/btv300

Wang, H., Liao, S., Li, H., Chen, Y., and Yu, J. (2019). Long Non-Coding RNA TUG1 Sponges Mir-145a-5p to Regulate Microglial Polarization after Oxygen-Glucose Deprivation. *Front. Mol. Neurosci.* 12, 215. doi:10.3389/fnmol.2019.00215

Wang, J.-M., Ko, C. Y., Chen, L. C., Wang, W. L., and Chang, W. C. (2006). Functional Role of NF-IL6beta and its Sumoylation and Acetylation Modifications in Promoter Activation of Cyclooxygenase 2 Gene. *Nucleic Acids Res.* 34 (1), 217–231. doi:10.1093/nar/gkj422

Wang, J., Chen, T., and Shan, G. (2017). miR-148b Regulates Proliferation and Differentiation of Neural Stem Cells via Wnt/β-Catenin Signaling in Rat Ischemic Stroke Model. *Front. Cel. Neurosci.* 11, 329. doi:10.3389/fncel.2017.00329

Xu, J., Wang, C., Meng, F., and Xu, P. (2021a). Long Non-Coding RNA H19 Inhibition Ameliorates Oxygen-Glucose Deprivation-Induced Cell Apoptosis and Inflammatory Cytokine Expression by Regulating the microRNA-29b/SIRT1/PGC-1α axis. *Mol. Med. Rep.* 23 (2), 131. doi:10.3892/mmr.2020.11770

Xu, Q., Zhao, B., Ye, Y., Li, Y., Zhang, Y., Xiong, X., et al. (2021b). Relevant Mediators Involved in and Therapies Targeting the Inflammatory Response Induced by Activation of the NLRP3 Inflammasome in Ischemic Stroke. *J. Neuroinflammation* 18 (1), 123. doi:10.1186/s12974-021-02137-8

Xu, R., Shimizu, F., Hovinga, K., Beal, K., Karimi, S., Droms, L., et al. (2016). Molecular and Clinical Effects of Notch Inhibition in Glioma Patients: A Phase 0/I Trial. *Clin. Cancer Res.* 22 (19), 4786–4796. doi:10.1158/1078-0432.CCR-16-0048

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells. *Nucleic Acids Res.* 41 (Database issue), D955–D961. doi:10.1093/nar/gks1111

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, G., Li, T., Chang, X., and Xing, J. (2021a). Long Noncoding RNA SNHG14 Promotes Ischemic Brain Injury via Regulating miR-199b/AQP4 Axis. *Neurochem. Res.* 46 (5), 1280–1290. doi:10.1007/s11064-021-03265-6

Zhang, L., Liu, B., Han, J., Wang, T., and Han, L. (2020). Competing Endogenous RNA Network Analysis for Screening Inflammation-Related Long Non-Coding RNAs for Acute Ischemic Stroke. *Mol. Med. Rep.* 22 (4), 3081–3094. doi:10.3892/mmr.2020.11415

Zhang, S. R., Phan, T. G., and Sobey, C. G. (2021b). Targeting the Immune System for Ischemic Stroke. *Trends Pharmacol. Sci.* 42 (2), 96–105. doi:10.1016/j.tips.2020.11.010

Zhao, H., Han, Z., Ji, X., and Luo, Y. (2016). Epigenetic Regulation of Oxidative Stress in Ischemic Stroke. *Aging Dis.* 7 (3), 295–306. doi:10.14336/AD.2015.1009

Zhong, Y., Yu, C., and Qin, W. (2019). LncRNA SNHG14 Promotes Inflammatory Response Induced by Cerebral Ischemia/Reperfusion Injury through Regulating miR-136-5p/ROCK1. *Cancer Gene Ther.* 26 (7-8), 234–247. doi:10.1038/s41417-018-0067-5

Zhou, J., Wang, L., Cui, Y., and Tang, L. (2021). miR-125a-5p-Targeted Regulation of TRA2β Expression Inhibits Proliferation and Metastasis of Hepatocellular Carcinoma Cells. *Am. J. Transl Res.* 13 (12), 14074–14080.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Wu, Wei, Fan, Gu, Li and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 08 April 2022
doi: 10.3389/fgene.2022.832331

Exploration of Reduced Mitochondrial Content–Associated Gene Signature and Immunocyte Infiltration in Colon Adenocarcinoma by an Integrated Bioinformatic Analysis

Jinlin Kang[1,2,3†], *Na li*[1,2,3,4†], *Fen Wang*[1,2,3†], *Yan Wei*[1,2,3], *Yangyang Zeng*[1,2,3], *Qifan Luo*[1,2,3], *Xuehua Sun*[1,2,3], *Hui Xu*[1,2,3], *Jin Peng*[1,2,3*] *and Fuxiang Zhou*[1,2,3*]

[1]*Department of Radiation and Medical Oncology, Zhongnan Hospital Wuhan University, Wuhan, China,* [2]*Hubei Province Key Laboratory of Tumor Biological Behaviors, Wuhan, China,* [3]*Hubei Cancer Clinical Study Center, Wuhan, China,* [4]*Renmin Hospital of Wuhan University, Wuhan, China*

Purpose: Mitochondrial dysfunction refers to cancer immune evasion. A novel 7-gene prognostic signature related to the mitochondrial DNA copy number was utilized to evaluate the immunocyte infiltration in colon cancer according to the risk scores and to predict the survival for colon cancer.

Experimental design: We performed an integrated bioinformatic analysis to analyze transcriptome profiling of the EB-treated mitochondrial DNA–defected NCM460 cell line with differentially expressed genes between tumor and normal tissues of COAD in TCGA. The LASSO analysis was utilized to establish a prognostic signature. ESTIMATE and CIBERSORT validated the differences of immunocyte infiltration between colon cancer patients with high- and low-risk scores.

Results: Our study identified a 7-gene prognostic signature (*LRRN2*, *ANKLE1*, *GPRASP1*, *PRAME*, *TCF7L1*, *RAB6B*, and *CALB2*). Patients with colon cancer were split into the high- and low-risk group by the risk scores in TCGA (training cohort: HR = 2.50 $p < 0.0001$) and GSE39582 (validation cohort: HR = 1.43 $p < 0.05$). ESTIMATE and CIBERSORT revealed diverseness of immune infiltration in the two groups, especially downregulated T-cell infiltration in the patients with high-risk scores. Finally, we validated the colon patients with a low expression of the mitochondrial number biomarker TFAM had less $CD3^+$ and $CD8^+$ T-cell infiltration in clinical specimens.

Conclusion: An mtDNA copy number-related 7-gene prognostic signature was investigated and evaluated, which may help to predict the prognosis of colon cancer patients and to guide clinical immunotherapy *via* immunocyte infiltration evaluation.

Keywords: mitochondiral DNA, Colon Cancer, Immunocyte infiltration, Bioinformatic analysis, Gene signature

OPEN ACCESS

Edited by:
William C. Cho, QEH, Hong Kong SAR, China

Reviewed by:
Zengjun Wang, Nanjing Medical University, China
Zaisheng Ye, Fujian Provincial Cancer Hospital, China

***Correspondence:**
Jin Peng
pjxp888@126.com
Fuxiang Zhou
happyzhoufx@sina.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *09 December 2021*
Accepted: *14 March 2022*
Published: *08 April 2022*

Citation:
Kang J, li N, Wang F, Wei Y, Zeng Y, Luo Q, Sun X, Xu H, Peng J and Zhou F (2022) Exploration of Reduced Mitochondrial Content–Associated Gene Signature and Immunocyte Infiltration in Colon Adenocarcinoma by an Integrated Bioinformatic Analysis. Front. Genet. 13:832331. doi: 10.3389/fgene.2022.832331

INTRODUCTION

Colon cancer is the fourth most common malignant tumor, which caused approximately 247,563 deaths in China in 2018 (Feng et al., 2019). The incidence and mortality of colorectal cancer increased rapidly with the economic development in recent years (Chen et al., 2016). Molecular testing plays an increasingly important role in estimating prognosis and deciding the best therapy for colon cancer patients with worse prognosis (Van Cutsem et al., 2016). We focused on mitochondrial dysfunction, which has been proved to play diverse roles in cancer metabolism, immune response, and cell signaling pathways (Hutson et al., 1988; West et al., 2011; Wallace, 2012). Mitochondria have their own 16-kilobase mitochondrial genome independent of the nucleus genome, which encodes two rRNAs, 22 tRNAs, and 13 polypeptides. The alterations of mitochondrial DNA (mtDNA) contribute to aberrant mitochondrial respiration, metabolism, and other cellular functions (Yu, 2011). Metabolism alteration has been commonly acknowledged as one of the hallmarks of cancer, which is associated with tumorigenesis, tumor microenvironment, cell progression, immune infiltration, and treatment response (Elia and Fendt, 2016; Pavlova and Thompson, 2016; Lunt and Fendt, 2017). Notably, increasing evidence demonstrated the relationship between immune cell infiltration and dysregulated metabolic pathways (Peteret al., 2017; Li et al., 2019). The alterations of metabolism promoted reprogramming of the tumor microenvironment (TME), herein enhanced immunostimulation (Cheng et al., 2014; Arts et al., 2016).

Following the inspiring success in melanoma and lung cancer, immunotherapy has been applied in the treatment of colon cancer since 2017. However, current immune checkpoint inhibitors (ICIs) merely showed less effect in several restricted subtypes of colon cancer (Ganesh et al., 2019). Accumulating evidence suggests that less infiltration of immunocytes showed the contribution of immune resistance (Herrera et al., 2013). The immune system recognizes itself through binding T-cell receptors (TCRs) to the human leukocyte antigen (HLA). This process is modulated by co-stimulatory factors named immune checkpoints, such as programmed cell death 1 (PD1) and cytotoxic T-lymphocyte antigen 4 (CTLA4) (Sharma and Allison, 2015). However, there were only few studies that clarified the relationship between mtDNA copy number reduction and tumor immunocyte infiltration. The value of aforementioned seven signature genes in predicting the immunocyte infiltration in colon cancer was assessed. Subsequently, the results were validated in GSE39582 (n = 550) and estimated using the ESTIMATE algorithm. Taken together, we utilized integrated bioinformatic analysis to select seven significant genes positively related to mitochondrial content reduction and confirmed their potential values in colon cancer to predict the prognosis and evaluate immune infiltration.

MATERIAL AND METHODS

Data Acquisition

HTSeq-FPKM workflow type transcriptome data and clinical data of colon adenocarcinoma (COAD) were obtained from TCGA (https://portal.gdc.cancer.gov). GEO datasets (GSE39582) were obtained from GEO (https://www.ncbi.nlm.nih.gov/geo/). Only those patients whose overall survival days were more than 30 days were collected.

Cell Lines and Culture Condition

The human immortalized colon cell line NCM460 was acquired from Procell Company. (Wuhan, China). The cells were cultured in Dulbecco's modified Eagle medium (DMEM; Thermo Fisher, United States) with 10% fetal bovine serum (Gibco, United States) and 1% pen-strep (100 U/ml penicillin and 100 mg/ml streptomycin) (Gibco, United States). The mtDNA copy number knockdown was induced by maintaining in 50 ng/ml ethidium bromide (EB) with 1 mM sodium pyruvate (Sigma-Aldrich, St. Louis, MO, United States) and 50 μg/ml uridine (Sigma-Aldrich, St. Louis, MO, United States) ethidium bromide (EB) maintained for over 4 months.

RNA-Seq and Data Analysis

Total RNA was extracted from NCM460 using TRIzol. RNA samples were taken from EB-treated NCM460 and normal NCM460 cells; RNA transcriptome analysis was conducted by Seqhealth Technology Company (Wuhan, China).

Differentially Expressed Gene Selection

The transcriptome data of NCM460 and EB-treated NCM460 cells, COAD in TCGA dataset, and GSE39582 were extracted for further analysis. Differences between tumor tissue and adjacent normal tissue were collected using the limma package. |logFC| more than 0.5 and the false discovery rate (FDR) less than 0.05 were set down as the cut-off standard.

Functional Enrichment Analysis

The functional analysis of DEGs was implemented using clusterProfiler.

PPI Network Analysis

The analysis of the PPI network with the confidence score >0.9 was performed using the Metascape website (https://metascape.org) and Cytoscape. In order to visualize the PPI network and screen out valuable hub genes, we utilized the MCODE plug-in of Cytoscape software with the degree as 5, node score as 0.2, k-core as 2, and maximum depth as 100 for cut-off criteria

Estimation of Immune Cell Proportions and Infiltration

The correlation scores of individual samples were evaluated using the ESTIMATE algorithm. The expressions of the HLA family genes, CTLA4, and 22 immune-related subtypes were analyzed

between the low-risk and high-risk groups using the limma and CIBERSORT package.

Survival Analysis

The Kaplan–Meier survival analysis was performed to assess the correlation between the patient survival probability and different groups, which were characterized by the risk score related to the 7-gene feature expression in TCGA calculated by LASSO. The results were validated in GSE39582 (n = 550).

Identification of Prognostic Characteristics of Immune-Related Genes

All of the 475 DEGs ($p < 0.05$) were screened by overlapping transcriptome DEGs from tumor vs. adjacent normal tissue and DEGS from NCM460 cells treated with ethidium bromide (EB) or not. Among them, 12 genes were assayed by LASSO Cox analysis with 10-round cross-validation. Stepwise multiple Cox regression analysis was performed to select and optimize prognostic features. Univariate and multivariate Cox regression analyses assessed the independence of the prognostic signature from clinical factors. Time-dependent ROC analysis was performed using the survival package. Nomogram and calibration curves were accomplished using the rms package.

Immunohistochemistry

Patients pathologically diagnosed by biopsy in Zhongnan hospital, Wuhan University, in the past 3 years were included in our research. The paraffin sections of surgical tissues of 39 colon cancer patients were used for the immunohistochemical analyses. These sections were put into xylene, absolute ethanol, 85% alcohol, and 75% alcohol for deparaffinzation and rehydration. The tissue sections were filled with citric acid (PH6.0) in a microwave oven and heated on medium power for antigen retrieval. The sections were put into 3% hydrogen peroxide and incubated at 4°C in darkness. To cover the tissues, the sections were sealed for 30 min in 3%BSA at 4°C. The sections were placed flat in a wet box and incubated with primary antibodies against transcription factor A (TFAM, Proteintech), CD3 (Servicebio GB13014), and CD8 (Servicebio GB13429 overnight at 4°C. The samples were then incubated with the secondary antibody (HRP-labeled) for 50 min. The samples were visualized and counterstained with DAB and hematoxylin and then dehydrated in ethanol and cleared in xylene. The sections were placed and shaken on the decolorization shaker three times completely for 15 min in PBS(PH7.4) after each step. The intensities of CD3, CD8, and TFAM were visualized by Image-Pro Plus 6.0 and were analyzed by Halo (India labs, United States).

Quantitative Reverse Transcription PCR

Total RNA of SW480 cells was extracted using the RNeasy Mini Kit, while reverse transcription was performed using the Prime Script RT Reagent Kit. SYBR Premix Ex Taq (Aidlab Biotechnologies) was utilized for real-time PCR at an ABI Prism 7900 instrument (Applied Biosystems).

The primer sequences used in this research are as follows:

LRRN2: F CGAGGCTACCACTGTGGAC
LRRN2: R GGGCATCCGAAAAGCTGTTC
ANKLE1: F GACCCCAACGCTCGATCTG
ANKLE1: R TCGGGCTCCTGAGTCTCTG
GPRASP1: F AGGAGGAGACCAATATGGGGT
GPRASP1: R GGACCTAGACATGGTATTAGCCT
PRAME: F TGGAATTAACTTGTGGCAACCT
PRAME: R TCTGACAGCCCTCTAACACGA
TCF7L1: F TCGTCCCTGGTCAACGAGT
TCF7L1: R ACTTCGGCGAAATAGTCCCG
RAB6B: F TGTACGACAGCTTCGACAACA
RAB6B: R CTGCGGAACCTCTCCTGAC
CALB2: F ACTTTGACGCAGACGGAAATG
CALB2: R GAAGTTCTCTTCGGTTGGCAG
β-Actin: F CATGTACGTTGCTATCCAGGC
β-Actin: R CTCCTTAATGTCACGCACGAT

Statistical Analysis Method

Statistical analysis was performed in R version 4.0.3. The correlations of protein expression in pathological data were analyzed using prism 7.

RESULTS

Identification and Functional Analysis of DEGs

The gene expression profile of the mtDNA-reduced NCM460 cell line, which was derived from the normal human colon mucosal epithelium, was analyzed using the "limma" package. A total of 2075 DEGs were screened, of which 794 genes were upregulated while 1281 genes were downregulated. The results are shown in **Figures 1A,B**. To first realize the function of DEGs of EtBr-treated colon cells, the alternated DEGs were transmitted to the "clusterProfiler" for functional annotations. GO and KEGG term enrichment analyses showed that the upregulated DEGs were fairly enriched, as shown in (**Figures 1C–F**). Furthermore, we found that DEGs in the biological process (BP) group were enriched in "signal transduction, cell adhesion, oxidation–reduction process, positive regulation of GTPase activity, and positive regulation of cell proliferation" based on the results of GO analysis of all DEGs. In the cellular composition (CC) group, all DEGs were associated with the "integral component of membrane, plasma membrane, cytosol, extracellular exosome, and integral component of plasma membrane." In addition, the terms "calcium ion binding, receptor binding, kinase activity, heparin binding, and oxidoreductase activity" were enriched in the molecular function (MF) module. These apparently elevated terms of DEGs supported the notion that they functioned in carcinogenesis and progression for colon cancer in patients with mtDNA content reduction.

Intersection of mtDNA Content Reduction–Associated DEGs and Prognostic-Related DEGs

Comparing the transcriptome of NCM460 and the EtBr-treated NCM460 cell line, 2075 DEGs were identified, which were composed of 794 upregulated genes and 1281 downregulated genes. We selected 475 DEGs using the Venn diagram of

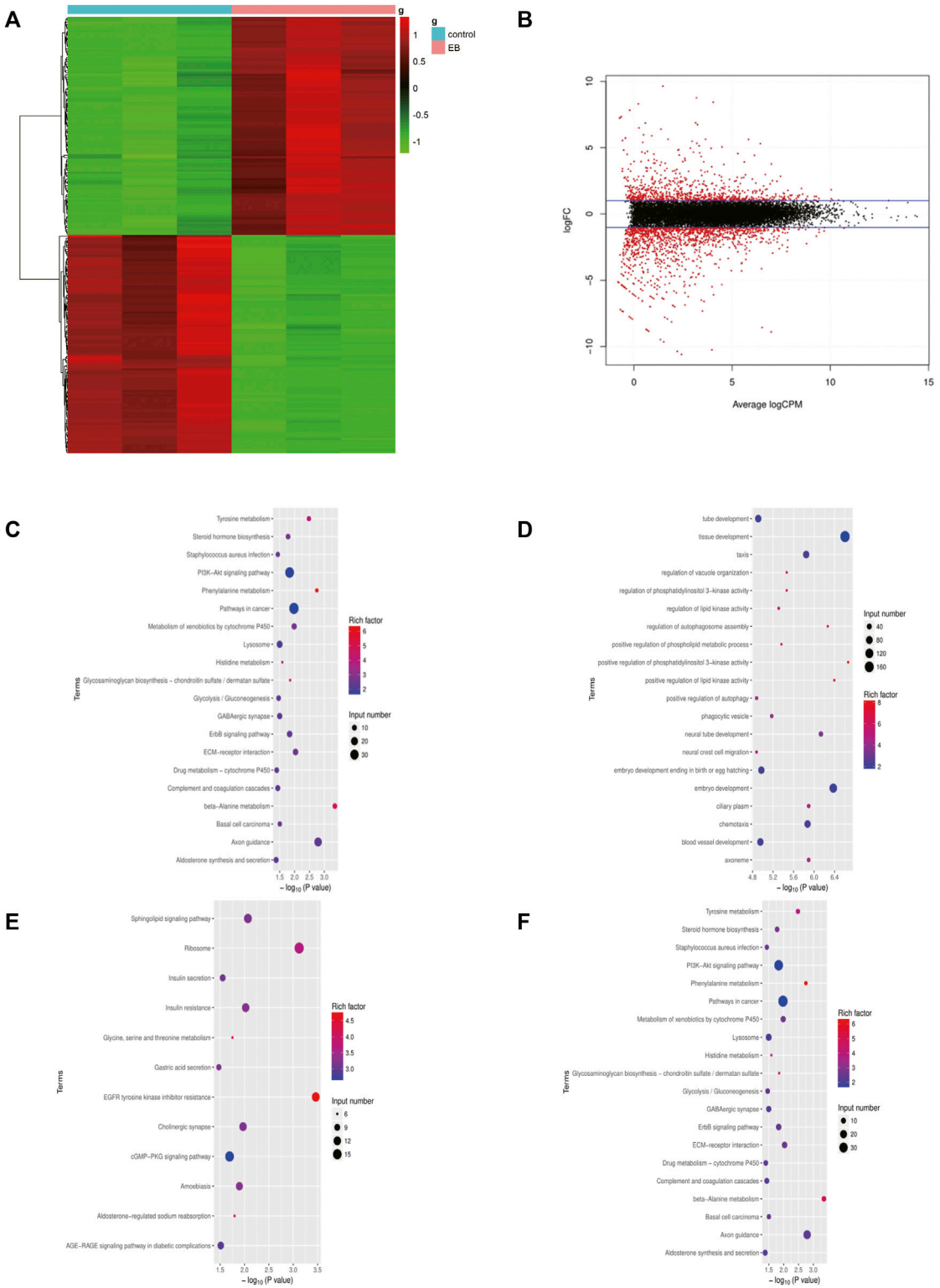


FIGURE 1 | DEGs identified in the EB-treated NCM460 cell line. **(A)** Heatmap of the top 200 DEGs by |logFC|. The color from blue to red indicates low to high expression level. **(B)** Volcano map between EB-treated and control groups of the NCM460 cell line. **(C,D)** Enriched terms in GO analysis of DEGs in EB-treated NCM460, respectively. **(E,F)** Enriched terms in KEGG analysis of DEGs in EB-treated NCM460 DEGs, respectively; DEGS means differentially expressed genes; logFC means log fold change.

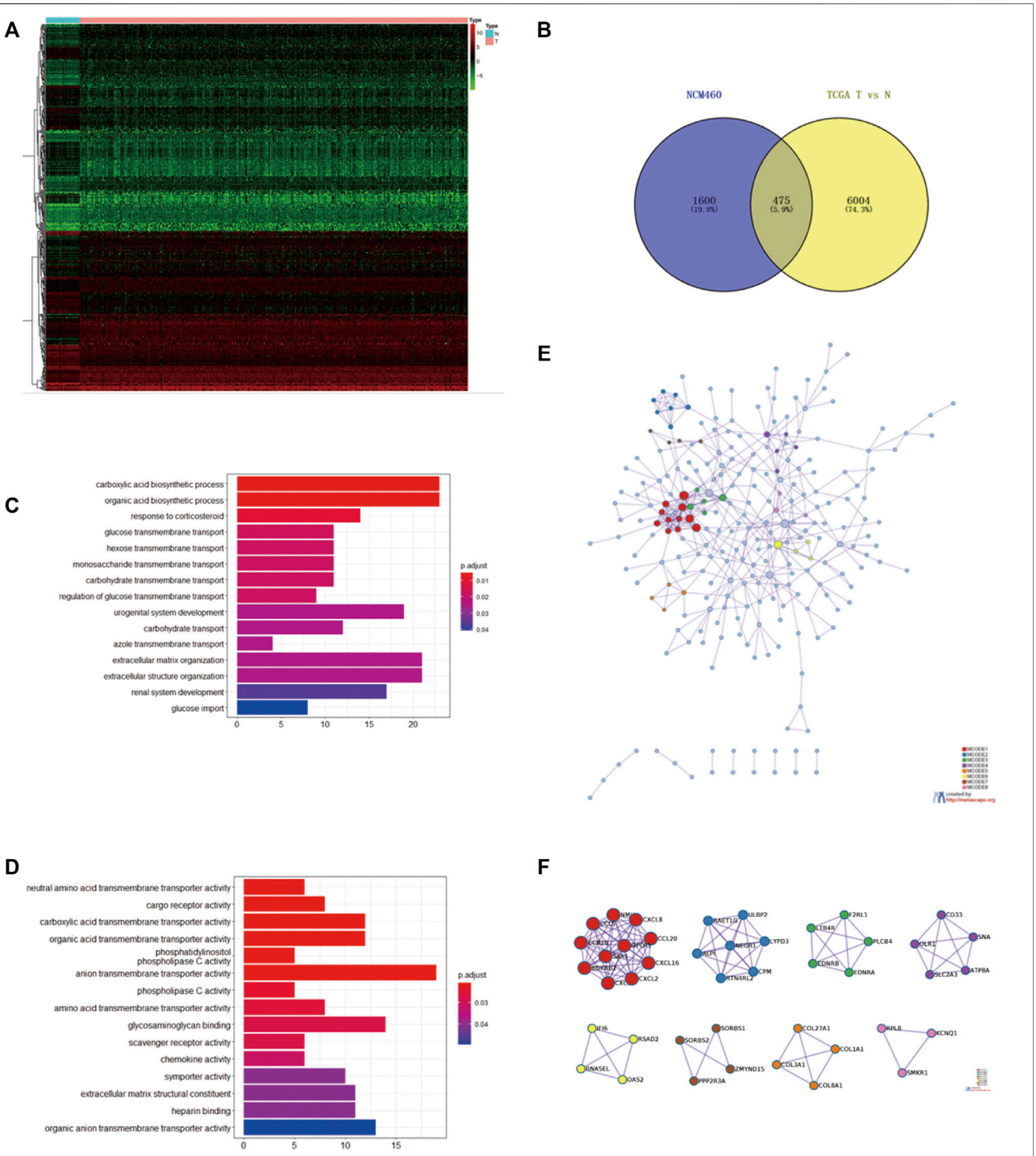


FIGURE 2 | Intersection of the mtDNA-related gene and prognostic gene in TCGA **(A)**. Heatmap of the top 200 DEGs identified between tumor and compared normal tissues from COAD in TCGA based on the value of |logFC| **(B)**. Intersection of 475 common DEGs with the prognostic value in TCGA and EB-treated NCM460 using the Venn diagram. Top three hub modules were classified in MODE. Red circles represented upregulated genes, while blue circles represented downregulated genes. **(C,D)** Heatmap of GO enriched terms and biological process (BP) GO enriched terms. **(E)** PPI network of the common 475 genes calculated by STRING software from the Metascape website **(F)** Top eight modules of the PPI network; GO, gene ontology; PPI, protein–protein interaction [Color figure can be viewed at wileyonlinelibrary.com].

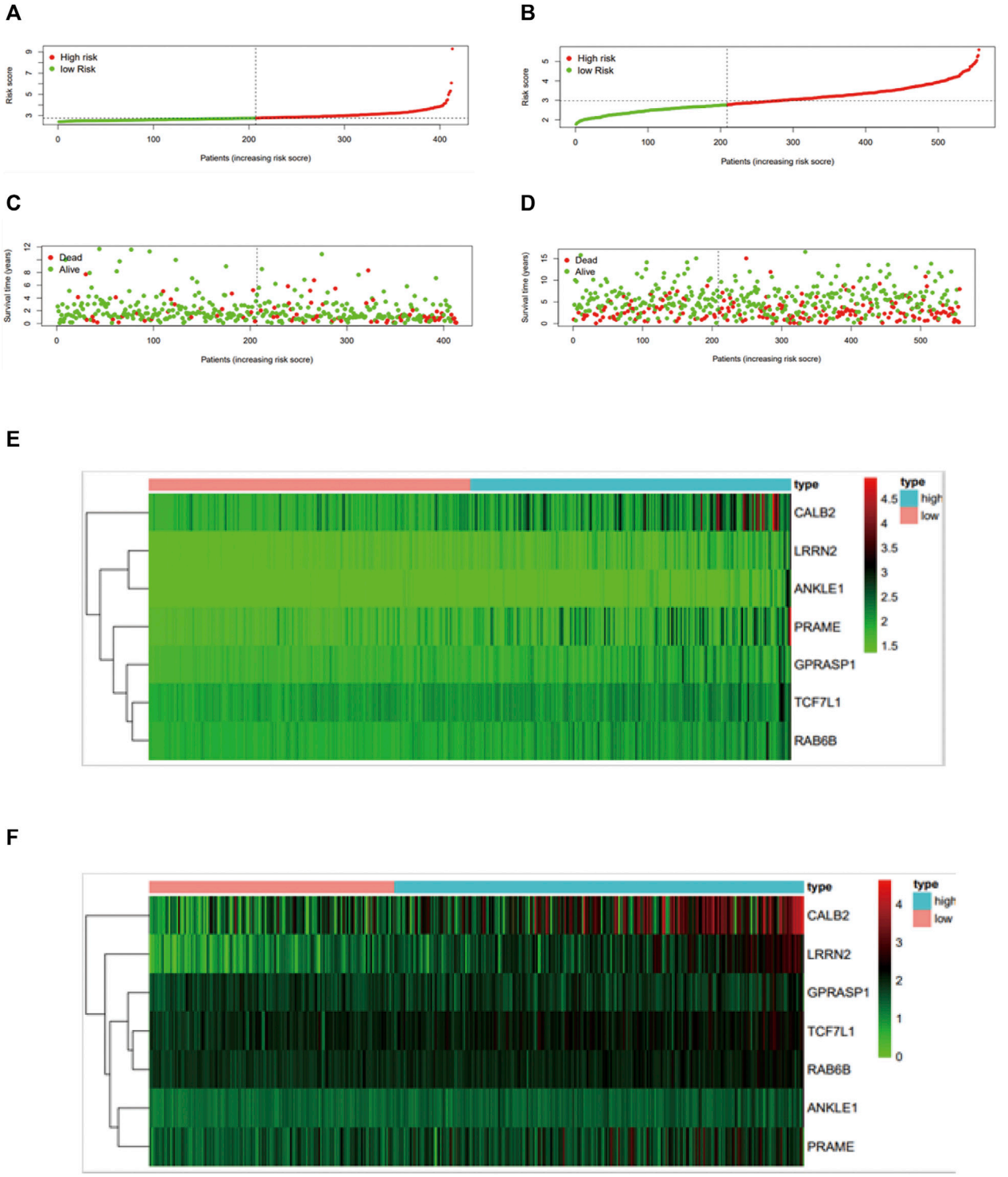


FIGURE 3 | Construction and validation of the mtDNA-related prognostic signature. Assessment of LASSO regression analysis of 475 DGEs is shown here. **(A)** Risk score, **(C)** survival status, and **(E)** profiles of mtDNA-associated gene expression. The value of the mtDNA-related prognostic signature was validated in GSE39582 by the **(B)** risk score, **(D)** survival status, and **(F)** mtDNA-related gene expression profiles, and GSE39582 was displayed in **Figure 4**.

prognostic genes in TCGA datasets and EB-treated NCM460 DEGs (**Figures 2A,B**). Functional analysis was performed *via* clusterProfiler in R. The top GO terms included the carboxylic acid biosynthetic process and organic acid biosynthetic process. The network of GO enriched terms is displayed in **Figures 2C,D**. Anion transmembrane transporter activity and glycosaminoglycan binding are enriched in the biological processes (BPs). Interestingly, there are no items enriched in the cellular composition (CC), and the top 15 terms are the same in all GO and MF (**Figure 2C**). It seemed DEGs are mostly correlated with the biosynthetic process and transmembrane transport. In addition, 475 DEGs were analyzed using the Search Tool for the Retrieval of Interacting Genes database (STRING) and performed to formulate the protein–protein interaction (PPI) network (**Figure 2E**). Central modules were instituted using Molecular Complex Detection (MCODE) (**Figure 2F**). The results gained from STRING showed that the PPI network of DEGs consisted of 252 nodes and 444 edges. The top 30 distinguished proteins were classified as hub genes, which might play vital roles in the mtDNA content reduction related to tumor initiation through PPI analysis. A total of eight clusters were generated in MCODE, and the top three clusters were selected as hub modules by the scores evaluated in MCODE (**Supplementary Table S2**). The 11 genes in MCODE1 (*CCR10*, *CXCL16*, *CXCL2*, *CXCL3*, *CXCL8*, *GPER1*, *NMU*, *SAA1*, *BDKRB2*, *CCL20*, and *CCL5*) were associated with cell chemotaxis, which induced the directional migration of cells including cancer cells and immune cells.

Construction and Validation of a 7-Gene Prognostic Signature

We implemented a prognostic analysis of 475 DEGs by the overlap of prognostic genes in TCGA datasets and NCM460 EB-related DEGs. The significant prognostic factors revealed by univariate Cox proportional hazards regression ($p < 0.001$) were SNAP25, LRRN2, ANKLE1, GPRASP1, CD37, PRAME, PDZD4, TCF7L1, RAB6B, CALB2, DUSP9, and SUSD5 (**Supplementary Table S1**).From the least absolute shrinkage and selection operator (LASSO) Cox regression analysis, 12 genes were filtered for stepwise multivariate Cox regression analysis (**Supplementary Table S1**). Finally, seven key genes (*LRRN2*, *ANKLE1*, *GPRASP1*, *PRAME*, *TCF7L1*, *RAB6B*, and *CALB2*) were generated to manufacture the prognostic signature: risk score = 0.231*LRRN2+ 0.448*ANKLE1+0.086*GPRASP1+ 0.118*PRAME+ 0.104*TCF7L1+ 0.171*RAB6B+ 0.049*CALB2. The distribution of the risk score, survival status, and gene expression profiles between the two groups in TCGA and GSE39582 are displayed in Fig.4.

Colon cancer patients in TCGA were split into a high-risk group and a low-risk group by the median risk score and then validated in GSE39582 dataset (**Figures 3A–F**). As a result, the overall survival time of patients in the low-risk group was significantly longer than that of the high-risk group ($n = 413$, HR = 2.5,$p < 0.0001$), which was also validated in GSE39582 ($n = 550$,HR = 1.4, $p < 0.05$) (**Figures 4A,B**). The distribution of risk scores with age, gender, stage, pathologic stage(T), pathologic lymph node status, and metastasis were also analyzed by univariate and multivariate cox proportional hazards regression in the training dataset and validated dataset (**Figures 4C–J**).

ESTIMATE Evaluation and Immune Infiltration Analysis

Recent studies suggest that the reprogramming of tumor metabolism led to local immunosuppression in the tumor microenvironment. An integrated bioinformatic analysis was applied to evaluate the differences of immune infiltration between the two groups with high- and low-risk scores established previously. As a result (**Figure 5**), the high-risk score group shows a higher ESTIMATE ($p < 0.001$), immune ($p < 0.05$), and stromal scores ($p < 0.001$) with a lower tumor purity ($p < 0.001$), which indicated that these patients with high risk scores may have strong infiltration of immune cells and stromal cells. Patients in the high-risk score group had a low-level expression in 12 HLA genes of the whole 24 HLA families and a higher expression of CTLA4 ($p < 0.01$) but no difference in the PDL1 expression.

Compared to the low-risk group, CIBERSORT displayed different immune cell infiltrations in the high-risk group: downregulated dendritic cells ($p = 0.006$), $CD8^+$ T cells ($p = 0.008$), and $CD4^+$ activated memory T cells ($p < 0.0001$), while upregulated naïve B cells ($p = 0.032$), M0 macrophages ($p < 0.0001$), T regulatory cells ($p = 0.024$), and eosinophils ($p = 0.011$). These results suggested that patients with the high-risk score might benefit less from immunotherapy, which showed the value of the 7-gene prognostic signature in predicting the efficacy of immune checkpoint inhibitor immunotherapy. Above results are shown in **Figure 6**

$CD8^+$ T-Cell Infiltration and Mitochondrial Copy Number Variation

As previously mentioned, our findings had demonstrated that patients with lower mtDNA content relatively were less immune cell infiltrated, especially $CD8^+$ T cells indicated using the ESTIMATE algorithm. To investigate whether mtDNA depletion was associated with a significant decrease of T-cell infiltration, we examined the abundance of the mitochondrial biomarker TFAM and T-cell infiltrations by CD3 and CD8 by immunohistochemistry (IHC) in the pathological section of colon cancer carcinoma in Zhongnan Hospital (**Figures 7A,B,D,E**). As a result, we found a significant positive relationship ($p = 0.0031$ and 0.0237, respectively) between $CD3^+$ and $CD8^+$ T cells and the expression of the mitochondrial biomarker TFAM (**Figures 7C,F**).

TFAM is critical for mitochondrial DNA replication, transcription, and stability, which has been confirmed to be related with poor prognosis in several cancers (Toki et al., 2010; Yoshida et al., 2011; Kurita et al., 2012; Kunkel et al., 2016). To investigate whether TFAM expression could be responsible for the decrease of T-cell infiltration in the mtDNA content-reduced condition, TFAM expression was knocked down by the TFAM siRNA lentivirus vector to knockdown in SW480 cells. Furthermore, five genes of the seven signature genes were significantly upregulated in TFAM-knockdown cells by qPCR (**Figure 7G**).

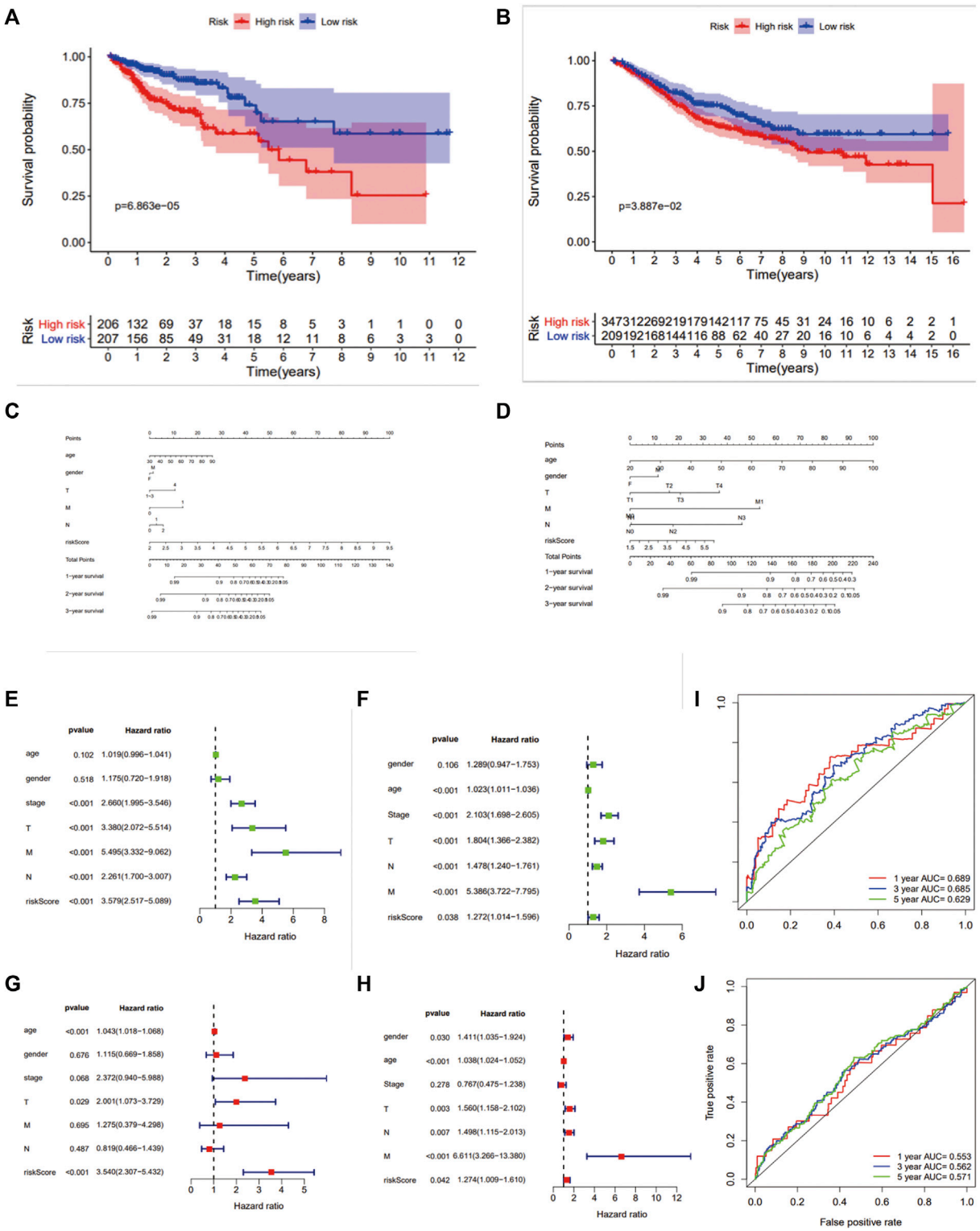


FIGURE 4 | Survival analysis and ROC analysis of the 7-gene prognostic signature. Kaplan–Meier survival curves of OS **(A,B)**, OS predictive nomogram **(C,D)**, univariate **(E,F)** and multivariate Cox proportional hazards regression **(G,H)** and ROC curves **(I,J)** in patients with colon carcinoma in TCGA and GSE39582.

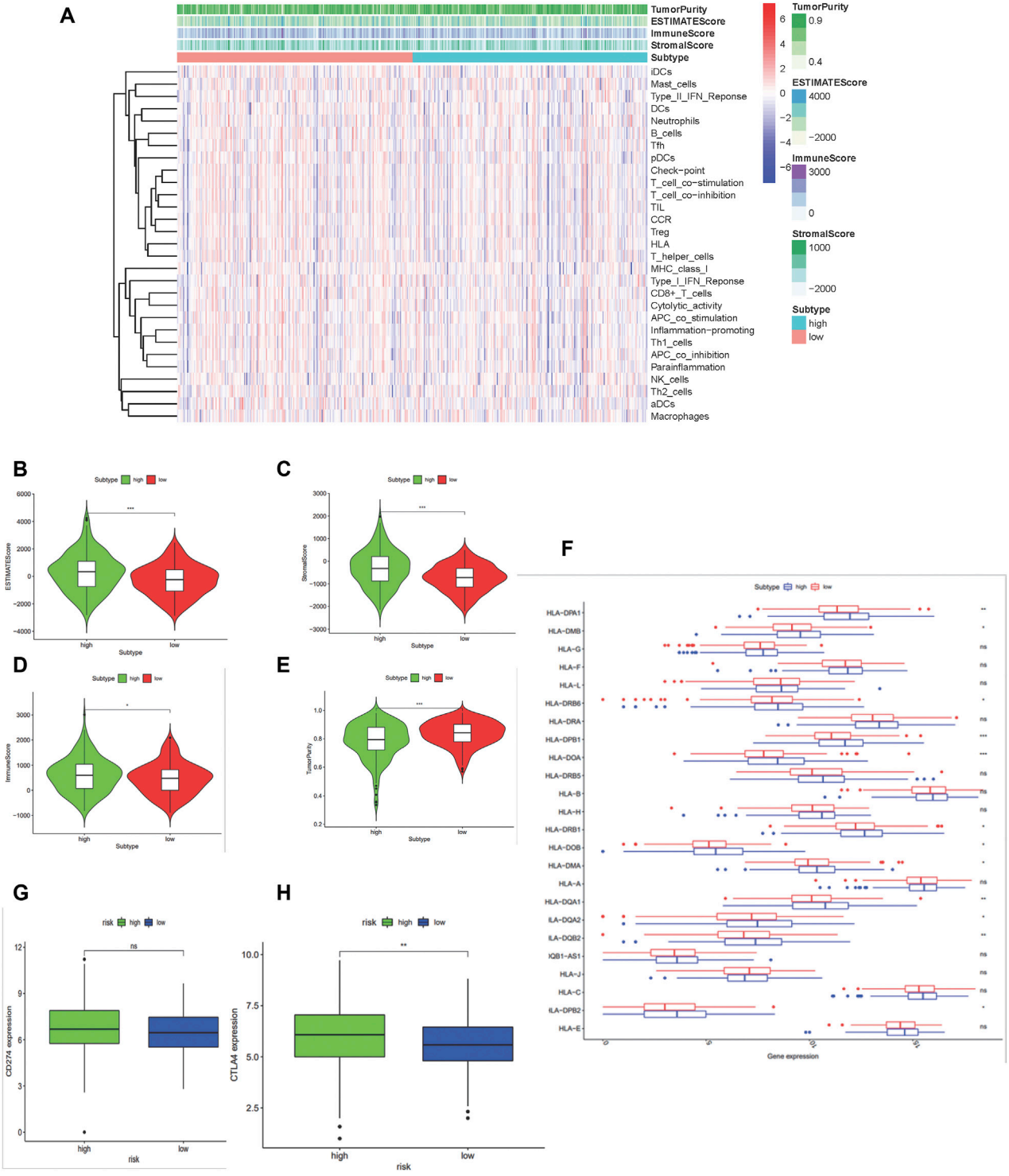


FIGURE 5 | Construction and certification of mtDNA-related patients groups in colon carcinoma samples. **(A)** Binary risk group constructed by LASSO regression analysis of 475 DGEs and further validated using the ESTIMATE algorithm. **(B–E)** Differences of scores calculated by the ESTIMATE between the high-risk and low-risk group. **(F)** Boxplot shows the expression of HLA family genes. **(G–H)** Expressions of PD-L1 (CD274) and CTLA4. (*:$p < 0.05$ **:$p < 0.01$ ***: $p < 0.001$).

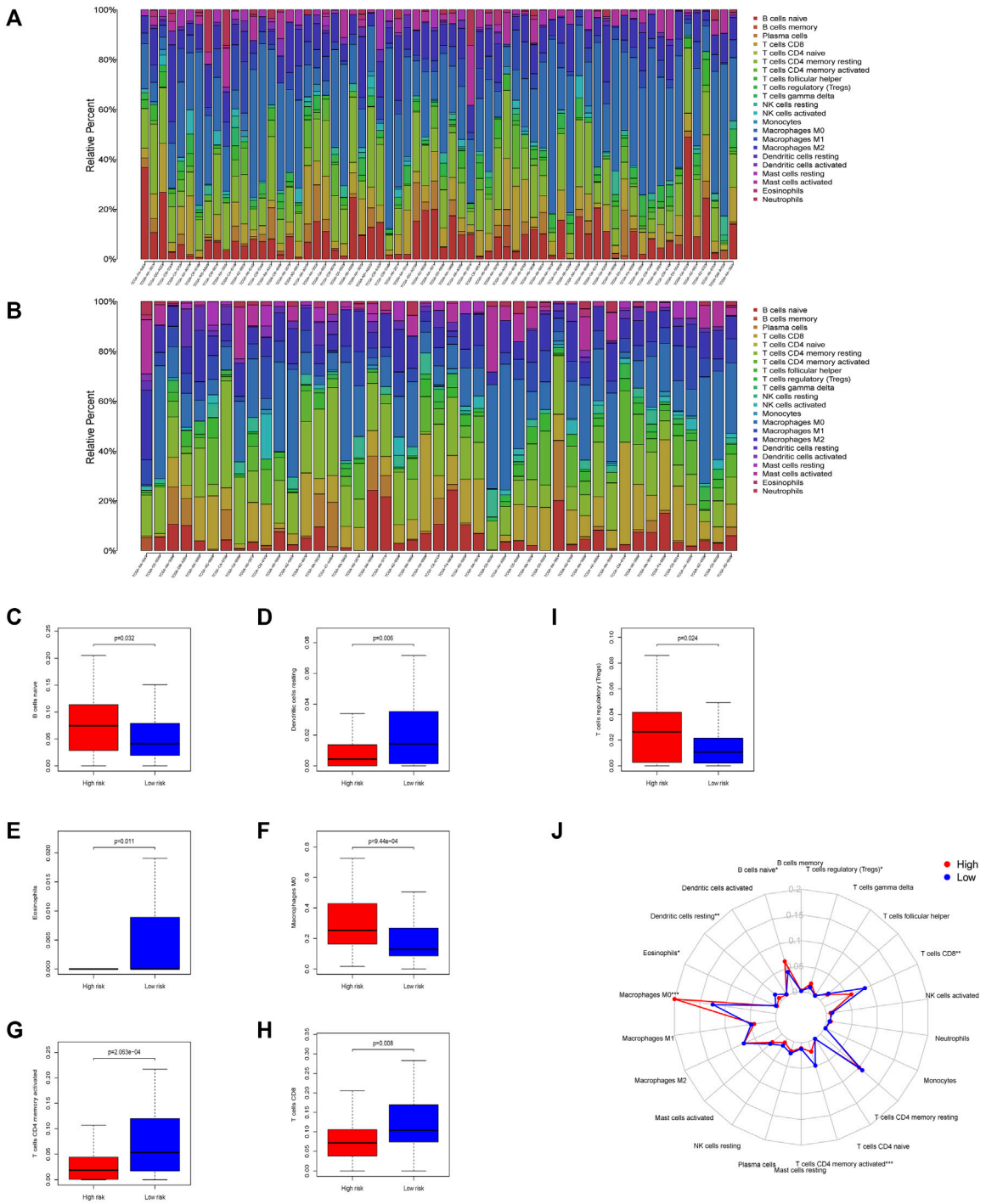


FIGURE 6 | Immune infiltration conditions between the two mtDNA-related groups in colon carcinoma samples. Immune infiltration of high- **(A)** and low-**(B)** risk group calculated using the CIBERSORT algorithm was shown in the two barplot, respectively. **(C–I)** Differences of tumor-infiltrating immune cells during the two groups. Seven subtypes of tumor-infiltrating immune cells between the high-risk and low-risk group showed significant differences (p <0.05). **(J)** Radar plot shows the whole survey of CIBERSORT analysis (*:p < 0.05 **:p < 0.01 ***: and p < 0.001).

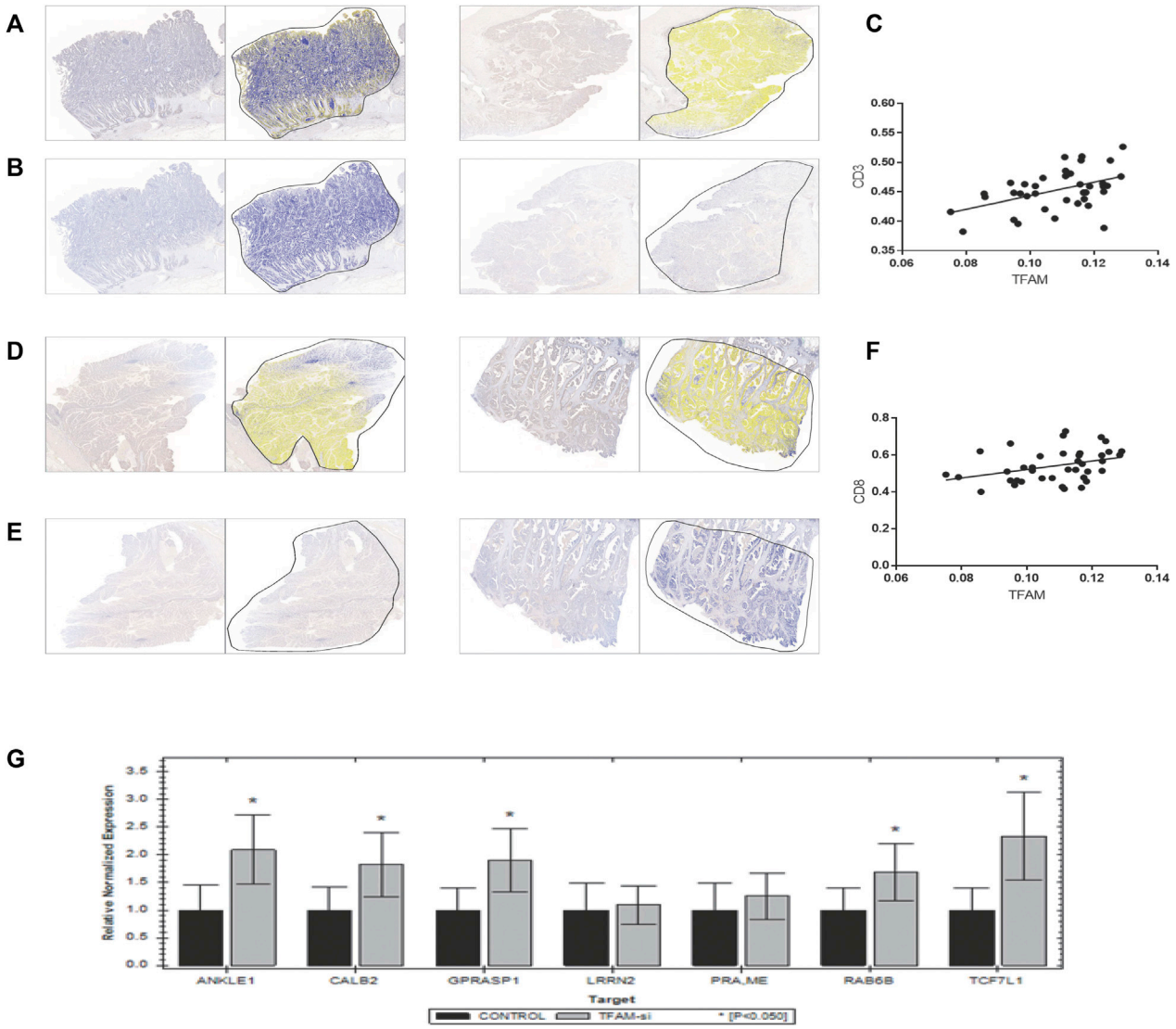


FIGURE 7 | Verification of the correlation between T-cell infiltration and mitochondrial copy number. Results of immunohistochemistry of 39 colon patients are shown here. **(A,B,D,E)** were paired IHC tissues. As shown in results, tissues with a higher TFAM expression [left of **(A,B,D,E)**] have a stronger CD3 [right of **(A,B)**] and CD8 [right of **(D,E)**] expression. **(C,F)** shows positive correlation relationship of CD3 and CD8 with the TFAM expression (p = 0.0031 and 0.0237, respectively). **(G)** Results of the mRNA expression of seven signature genes in TFAM-knockdown SW480 cells by qPCR.

DISCUSSION

As per our results, patients with colon adenocarcinoma were split into two different prognostic groups based on the risk scores calculated according to the expression of mtDNA content reduction-related signature genes in TCGA. Although previous studies have demonstrated that mtDNA content reduction played a vital role in tumor initiation, progression, and drug resistance in colon adenocarcinoma, few studies investigated the relationship of mtDNA content and immune infiltration (Buckner et al., 2014; Wei et al., 2018; Chen et al., 2020). It has been commonly acknowledged that mitochondrial DNA content reduction induced tumorigenesis, metabolic reprogram, and biological alterations, which influenced the immunocyte distribution (Lee et al., 2005; Lee and John, 2015; Mou et al., 2018). Hence, we investigated the transcriptome profiling of EtBr-treated NCM460 (human immortalized colon cell line) to assess the biological alterations induced by mtDNA content reduction.

Among the whole 2,075 DEGs, there were 794 upregulated genes and 1281 downregulated genes. All these DEGs were evaluated by GO term enrichment and KEGG pathway analysis. As the results of GO enrichment analysis, most of these terms were associated with metabolic reprogramming which was one of the hallmarks in tumorigenesis (Srinivasan et al., 2017). Furthermore, the GO and KEGG analyses also proved that alternations induced by mtDNA content reduction were mainly proficient in the metabolic pathways in COAD, which deserves further exploration.

From the 2,075 mtDNA-related DEGs, 475 DEGs were selected using the Venn diagram of prognostic genes in TCGA datasets. A total of eight clusters were generated in MCODE. MCODE 1 was the most notable cluster among them. The 11 genes in MCODE1 (*CCR10*, *CXCL16*, *CXCL2*, *CXCL3*, *CXCL8*, *GPER1*, *NMU*, *SAA1*, *BDKRB2*, *CCL20*, and *CCL5*) were associated with cell chemotaxis, which induced the directional migration of cells including cancer cells and immune cells (Buckner et al., 2014; AbdelMageed et al., 2019; Chen et al., 2019; Zhao Q. Q. et al., 2020). The findings indicated that the defected mtDNA content in cancer cells might be associated with tumor progression and immune cell infiltrations in COAD.

Finally, a mtDNA content-related 7-gene signature was identified according to prognostic genes from colon adenocarcinoma transcriptome data. The efficacy of the risk scores and the integrated nomograms to predict the outcome of patients was evaluated in TCGA dataset (COAD) and validated in GSE39582 datasets. Therefore, our mtDNA content-related signature risk scores maybe a prognostic biomarker in colon adenocarcinoma.

The question followed was why the mtDNA content-related signature played such roles in overall survival. As indicated in results, the 7-gene prognostic signature showed significant differences of HLA family genes and immune checkpoint expression between two groups, which played critical roles in the antitumor immune system (Fridman et al., 2012; McGranahan et al., 2017; Topalian et al., 2020). In total, 12 HLA genes were significantly downregulated among the whole 24 HLA families in patients with high-risk scores, which indicated that these patients may benefit less from immunotherapy since immune evasion. We investigated the expression of major immune checkpoints PDL1, CTLA4, LAG3, and PDCD1. Regrettably, there was no significant difference in PDL1, LAG3, and PDCD1. However, the high-risk score group had a higher expression of CTLA4, which predominantly inhibits T-cell activation and immune response (Fiegle et al., 2019). CTLA-4 blockade was shown to inhibit tumor progression by upregulating effector T-cell activity and suppress regulatory T cells (Tregs), which suggest that the patients with high-risk score of the 7-gene prognostic signature might benefit less from CTLA-4 blockade therapy. Thus, it also demonstrated the value of the 7-gene prognostic signature in immunotherapy.

In this study, seven genes (*LRRN2*, *ANKLE1*, *GPRASP1*, *PRAME*, *TCF7L1*, *RAB6B*, and *CALB2*) were identified. *LRRN2* was the first to be reported upregulated in glioma and was related to cell adhesion and signal transduction (Almeida et al., 1998). *ANKLE1* was known as a new hotspot for the predisposition of breast cancer, which played a vital role in DNA damage response and DNA repair (Bakshi et al., 2020). *GPRASP1* was mainly localized in the cell cytoplasm and could translocate to the nucleus, which indicates it may be involved in transcription regulation and tumorigenesis by retrograde signaling (Abu-Helo and Simonin, 2010). *PRAME* is considered as a repressor of retinoic receptor, which is recognized by cytolytic T lymphocytes (Xu et al., 2020). *TCF7L1* encodes T-cell family factors and regulates cell senescence *via* the Wnt/β-catenin signaling pathway (Shan et al., 2019). *Rab6B*, whose family member affects the regulation of intracellular transport routes, has been known as an oncogene in colon cancer (Zhao L. et al., 2020). Moreover, it has been convinced that *CALB2* is associated with apoptosis, ECM, and poor clinical outcomes *via* the mitochondrial pathway (Stevenson et al., 2011; Ojasalu et al., 2020). In brief, the role of the seven genes in colon cancer and mitochondrial-associated tumorigenesis also need further exploration, especially *PRAME* and *TCF7L1*.

On the other hand, the high-risk group had a higher ESTIMATE, immune, and stromal scores with lower tumor purity, which indicated the patients with high-risk scores might have more immune infiltration. Previous studies proved that reprogramming of the tumor immune microenvironment was a valuable prognostic signature, such as a lack of T cells, microphage phenotype, the number of B cells, $CD4^+$ T cells, $CD8^+$ T cells, DC cells, and the ablation of eosinophils (Shankaran et al., 2001; Galon et al., 2006; Bindea et al., 2013; Herrera et al., 2013; Arnold et al., 2020; Petitprez et al., 2020). The abundance of $CD8^+$, activated $CD4^+$ memory T cells, and DC cells was downregulated, while M0 microphages, B naïve cells, and Tregs were \ significantly upregulated in the high-risk group, which showed a pro-tumor tendency. M0 cells were considered as another type of TAM or an incompletely differentiated M2 and strongly associated with the poor outcome in many different types of cancers (Yang et al., 2019; Zhang et al., 2020). It was fairly controversial for the roles of tumor-infiltrating B cells, which was both tumor-promoting and tumor-suppressing (Martinet et al., 2012; Dieu-Nosjean et al., 2014; Colbeck et al., 2017; Sautès-Fridman et al., 2019). As is known, Tregs depletion promotes tumor growth and tumor immune evasion (Wei et al., 2004). $CD4^+$ and $CD8^+$ T-cell responses play a key role in the process of malignant cell elimination (Ostroumov et al., 2018). B naïve cells had been confirmed as a biomarker to predict the immunotherapy response (Petitprez et al., 2020). In tissues from colon cancer patients, less $CD3^+CD8^+$ T cells and the higher expression of the mitochondrial biomarker TFAM were found. It could be exclaimed that the alterations of the gene expression induced by mtDNA content defects brought about TME reprogramming, which indicated the potential of our risk score to forecast the effect of immunotherapy. Taken together, the 7-gene signature might work in the process of carcinogenesis, proliferation, and cancer immune evasion. Further exploration of these findings was required.

In conclusion, by multiple bioinformatics analysis, the hub genes of the DEGs intersected from TCGA and the transcriptome data of EB-treated NCM460 were screened, and the 7-gene prognostic signature was established in colon cancer initiation and progression. The predicting efficacy of the 7-gene prognostic signature could be attributed to the variations of immune infiltration and immune checkpoint in the tumor microenvironment. Our finding provides novel insights into the roles of mitochondrial DNA content reduction in TME reprogramming.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Ethics Committee, Zhongnan Hospital of Wuhan University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

JK, NL and FW designed this work. YW and QL performed the experiment and wrote the manuscript. YZ and XS performed the bioinformatics analysis. All the authors performed data review and approved the manuscript.

FUNDING

The National Natural Science Foundation (81773238) supported our study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.832331/full#supplementary-material

REFERENCES

AbdelMageed, M., Ali, H., Olsson, L., Lindmark, G., Hammarström, M. L., Hammarström, S., et al. (2019). The Chemokine CXCL16 Is a New Biomarker for Lymph Node Analysis of Colon Cancer Outcome. *Int. J. Mol. Sci.* 20 (22). 5793. doi:10.3390/ijms20225793

Abu-Helo, A., and Simonin, F. (2010). Identification and Biological Significance of G Protein-Coupled Receptor Associated Sorting Proteins (GASPs). *Pharmacol. Ther.* 126 (3), 244–250. doi:10.1016/j.pharmthera.2010.03.004

Almeida, A., Zhu, X. X., Vogt, N., Tyagi, R., Muleris, M., Dutrillaux, A.-M., et al. (1998). GAC1, a New Member of the Leucine-Rich Repeat Superfamily on Chromosome Band 1q32.1, Is Amplified and Overexpressed in Malignant Gliomas. *Oncogene* 16 (23), 2997–3002. doi:10.1038/sj.onc.1201828

Arnold, I. C., Artola-Boran, M., Gurtner, A., Bertram, K., Bauer, M., Frangez, Z., et al. (2020). The GM-CSF-IRF5 Signaling axis in Eosinophils Promotes Antitumor Immunity through Activation of Type 1 T Cell Responses. *J. Exp. Med.* 217 (12), e20190706. doi:10.1084/jem.20190706

Arts, R. J. W., Novakovic, B., ter Horst, R., Carvalho, A., Bekkering, S., Lachmandas, E., et al. (2016). Glutaminolysis and Fumarate Accumulation Integrate Immunometabolic and Epigenetic Programs in Trained Immunity. *Cel Metab.* 24 (6), 807–819. doi:10.1016/j.cmet.2016.10.008

Bakshi, D., Katoch, A., Chakraborty, S., Shah, R., Sharma, B., Bhat, A., et al. (2020). ANKLE1 as New Hotspot Mutation for Breast Cancer in Indian Population and Has a Role in DNA Damage and Repair in Mammalian Cells. *Front. Genet.* 11, 609758. doi:10.3389/fgene.2020.609758

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* 39 (4), 782–795. doi:10.1016/j.immuni.2013.10.003

Buckner, C. M., Moir, S., Kardava, L., Ho, J., Santich, B. H., Kim, L. J. Y., et al. (2014). CXCR4/IgG-expressing Plasma Cells Are Associated with Human Gastrointestinal Tissue Inflammation. *J. Allergy Clin. Immunol.* 133 (6), 1676–1685. doi:10.1016/j.jaci.2013.10.050

Chen, L., Peng, J., Wang, Y., Jiang, H., Wang, W., Dai, J., et al. (2020). Fenofibrate-induced Mitochondrial Dysfunction and Metabolic Reprogramming Reversal: the Anti-tumor Effects in Gastric Carcinoma Cells Mediated by the PPAR Pathway. *Am. J. Transl Res.* 12 (2), 428–446.

Chen, M. C., Baskaran, R., Lee, N. H., Hsu, H. H., Ho, T. J., Tu, C. C., et al. (2019). CXCL2/CXCR2 axis Induces Cancer Stem Cell Characteristics in CPT-11-resistant LoVo colon Cancer Cells *via* Gαi-2 and Gαq/11. *J. Cell Physiol.* 234 (7), 11822–11834. doi:10.1002/jcp.27891

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer Statistics in China, 2015. *CA: A Cancer J. Clinicians* 66 (2), 115–132. doi:10.3322/caac.21338

Cheng, S.-C., Quintin, J., Cramer, R. A., Shepardson, K. M., Saeed, S., Kumar, V., et al. (2014). mTOR- and HIF-1α-Mediated Aerobic Glycolysis as Metabolic Basis for Trained Immunity. *Science* 345 (6204), 1250684. doi:10.1126/science.1250684

Colbeck, E. J., Ager, A., Gallimore, A., and Jones, G. W. (2017). Tertiary Lymphoid Structures in Cancer: Drivers of Antitumor Immunity, Immunosuppression, or Bystander Sentinels in Disease? *Front. Immunol.* 8, 1830. doi:10.3389/fimmu.2017.01830

Dieu-Nosjean, M.-C., Goc, J., Giraldo, N. A., Sautès-Fridman, C., and Fridman, W. H. (2014). Tertiary Lymphoid Structures in Cancer and beyond. *Trends Immunol.* 35 (11), 571–580. doi:10.1016/j.it.2014.09.006

Elia, I., and Fendt, S. M. (2016). Vivo Cancer Metabolism Is Defined by the Nutrient Microenvironment. *Translational Cancer Res.* 5 (S6). doi:10.21037/tcr.2016.11.53

Feng, R.-M., Zong, Y.-N., Cao, S.-M., and Xu, R.-H. (2019). Current Cancer Situation in China: Good or Bad News from the 2018 Global Cancer Statistics? *Cancer Commun.* 39 (1), 22. doi:10.1186/s40880-019-0368-6

Fiegle, E., Doleschel, D., Koletnik, S., Rix, A., Weiskirchen, R., Borkham-Kamphorst, E., et al. (2019). Dual CTLA-4 and PD-L1 Blockade Inhibits Tumor Growth and Liver Metastasis in a Highly Aggressive Orthotopic Mouse Model of Colon Cancer. *Neoplasia* 21 (9), 932–944. doi:10.1016/j.neo.2019.07.006

Fridman, W. H., Pagès, F., Sautès-Fridman, C., and Galon, J. (2012). The Immune Contexture in Human Tumours: Impact on Clinical Outcome. *Nat. Rev. Cancer* 12 (4), 298–306. doi:10.1038/nrc3245

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., et al. (2006). Type, Density, and Location of Immune Cells within Human Colorectal Tumors Predict Clinical Outcome. *Science* 313 (5795), 1960–1964. doi:10.1126/science.1129139

Ganesh, K., Stadler, Z. K., Cercek, A., Mendelsohn, R. B., Shia, J., Segal, N. H., et al. (2019). Immunotherapy in Colorectal Cancer: Rationale, Challenges and Potential. *Nat. Rev. Gastroenterol. Hepatol.* 16 (6), 361–375. doi:10.1038/s41575-019-0126-x

Herrera, M., Herrera, A., Domínguez, G., Silva, J., García, V., García, J. M., et al. (2013). Cancer-associated Fibroblast and M2 Macrophage Markers Together Predict Outcome in Colorectal Cancer Patients. *Cancer Sci.* 104 (4), 437–444. doi:10.1111/cas.12096

Hutson, S. M., Fenstermacher, D., and Mahar, C. (1988). Role of Mitochondrial Transamination in Branched Chain Amino Acid Metabolism. *J. Biol. Chem.* 263 (8), 3618–3625. doi:10.1016/s0021-9258(18)68969-0

Kunkel, G. H., Chaturvedi, P., and Tyagi, S. C. (2016). Mitochondrial Pathways to Cardiac Recovery: TFAM. *Heart Fail. Rev.* 21 (5), 499–517. doi:10.1007/s10741-016-9561-8

Kurita, T., Izumi, H., Kagami, S., Kawagoe, T., Toki, N., Matsuura, Y., et al. (2012). Mitochondrial Transcription Factor A Regulates BCL2L1 Gene Expression and Is a Prognostic Factor in Serous Ovarian Cancer. *Cancer Sci.* 103 (2), 239–444. doi:10.1111/j.1349-7006.2011.02156.x

Lee, H.-C., Yin, P.-H., Lin, J.-C., Wu, C.-C., Chen, C.-Y., Wu, C.-W., et al. (2005). Mitochondrial Genome Instability and mtDNA Depletion in Human Cancers. *Ann. N. Y Acad. Sci.* 1042, 109–122. doi:10.1196/annals.1338.011

Lee, W. T. Y., and John, J. S. (2015). The Control of Mitochondrial DNA Replication during Development and Tumorigenesis. *Ann. N.Y. Acad. Sci.* 1350, 95–106. doi:10.1111/nyas.12873

Li, X., Wenes, M., Romero, P., Huang, S. C.-C., Fendt, S.-M., and Ho, P.-C. (2019). Navigating Metabolic Pathways to Enhance Antitumour Immunity and Immunotherapy. *Nat. Rev. Clin. Oncol.* 16 (7), 425–441. doi:10.1038/s41571-019-0203-7

Lunt, S. Y., and Fendt, S. M. (2017). Metabolism - A Cornerstone of Cancer Initiation, Progression, Immune Evasion and Treatment Response. *Curr. Opin. Syst. Biol.* 8, 67–72.doi:10.1016/j.coisb.2017.12.006

Martinet, L., Le Guellec, S., Filleron, T., Lamant, L., Meyer, N., Rochaix, P., et al. (2012). High Endothelial Venules (HEVs) in Human Melanoma Lesions. *Oncoimmunology* 1 (6), 829–839. doi:10.4161/onci.20492

McGranahan, N., Rosenthal, R., Hiley, C. T., Rowan, A. J., Watkins, T. B. K., Wilson, G. A., et al. (2017). Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* 171 (6), 1259–e11. e11. doi:10.1016/j.cell.2017.10.001

Mou, J.-j., Peng, J., Shi, Y.-y., Li, N., Wang, Y., Ke, Y., et al. (2018). Mitochondrial DNA Content Reduction Induces Aerobic Glycolysis and Reversible Resistance to Drug-Induced Apoptosis in SW480 Colorectal Cancer Cells. *Biomed. Pharmacother.* 103, 729–737. doi:10.1016/j.biopha.2018.04.099

Ojasalu, K., Brehm, C., Hartung, K., Nischak, M., Finkernagel, F., Rexin, P., et al. (2020). Upregulation of Mesothelial Genes in Ovarian Carcinoma Cells Is Associated with an Unfavorable Clinical Outcome and the Promotion of Cancer Cell Adhesion. *Mol. Oncol.* 14 (9), 2142–2162. doi:10.1002/1878-0261.12749

Ostroumov, D., Fekete-Drimusz, N., Saborowski, M., Kühnel, F., and Woller, N. (2018). CD4 and CD8 T Lymphocyte Interplay in Controlling Tumor Growth. *Cell. Mol. Life Sci.* 75 (4), 689–713. doi:10.1007/s00018-017-2686-7

Pavlova, N. N., and Thompson, C. B. (2016). The Emerging Hallmarks of Cancer Metabolism. *Cel Metab.* 23 (1), 27–47. doi:10.1016/j.cmet.2015.12.006

Peter, S.J., Beckermann, K. E., Mason, F. M., Andrejeva, G., Greenplate, A. R., et al. (2017). Mitochondrial Dysregulation and Glycolytic Insufficiency Functionally Impair CD8 T Cells Infiltrating Human Renal Cell Carcinoma. *JCI Insight* 2 (12), e93411. doi:10.1172/jci.insight.93411

Petitprez, F., de Reyniès, A., Keung, E. Z., Chen, T. W.-W., Sun, C.-M., Calderaro, J., et al. (2020). B Cells Are Associated with Survival and Immunotherapy Response in Sarcoma. *Nature* 577 (7791), 556–560. doi:10.1038/s41586-019-1906-8

Sautès-Fridman, C., Petitprez, F., Calderaro, J., and Fridman, W. H. (2019). Tertiary Lymphoid Structures in the Era of Cancer Immunotherapy. *Nat. Rev. Cancer* 19 (6), 307–325. doi:10.1038/s41568-019-0144-6

Shan, J., Shen, J., Wu, M., Zhou, H., Feng, J., Yao, C., et al. (2019). Tcf7l1 Acts as a Suppressor for the Self-Renewal of Liver Cancer Stem Cells and Is Regulated by IGF/MEK/ERK Signaling Independent of β-Catenin. *Stem Cells* 37 (11), 1389–1400. doi:10.1002/stem.3063

Shankaran, V., Ikeda, H., Bruce, A. T., White, J. M., Swanson, P. E., Old, L. J., et al. (2001). IFNγ and Lymphocytes Prevent Primary Tumour Development and Shape Tumour Immunogenicity. *Nature* 410 (6832), 1107–1111. doi:10.1038/35074122

Sharma, P., and Allison, J. P. (2015). The Future of Immune Checkpoint Therapy. *Science* 348 (6230), 56–61. doi:10.1126/science.aaa8172

Srinivasan, S., Guha, M., Kashina, A., and Avadhani, N. G. (2017). Mitochondrial Dysfunction and Mitochondrial Dynamics-The Cancer Connection. *Biochim. Biophys. Acta (Bba) - Bioenerg.* 1858 (8), 602–614. doi:10.1016/j.bbabio.2017.01.004

Stevenson, L., Allen, W. L., Proutski, I., Stewart, G., Johnston, L., McCloskey, K., et al. (2011). Calbindin 2 (CALB2) Regulates 5-fluorouracil Sensitivity in Colorectal Cancer by Modulating the Intrinsic Apoptotic Pathway. *PLoS One* 6 (5), e20276. doi:10.1371/journal.pone.0020276

Toki, N., Kagami, S., Kurita, T., Kawagoe, T., Matsuura, Y., Hachisuga, T., et al. (2010). Expression of Mitochondrial Transcription Factor A in Endometrial Carcinomas: Clinicopathologic Correlations and Prognostic Significance. *Virchows Arch.* 456 (4), 387–393. doi:10.1007/s00428-010-0895-7

Topalian, S. L., Taube, J. M., and Pardoll, D. M. (2020). Neoadjuvant Checkpoint Blockade for Cancer Immunotherapy. *Science* 367 (6477), 367. doi:10.1126/science.aax0182

Van Cutsem, E., Cervantes, A., Adam, R., Sobrero, A., Van Krieken, J. H., Aderka, D., et al. (2016). ESMO Consensus Guidelines for the Management of Patients with Metastatic Colorectal Cancer. *Ann. Oncol.* 27 (8), 1386–1422. doi:10.1093/annonc/mdw235

Wallace, D. C. (2012). Mitochondria and Cancer. *Nat. Rev. Cancer* 12 (10), 685–698. doi:10.1038/nrc3365

Wei, W.-Z., Morris, G. P., and Kong, Y.-c. M. (2004). Anti-tumor Immunity and Autoimmunity: a Balancing Act of Regulatory T Cells. *Cancer Immunol. Immunother.* 53 (2), 73–78. doi:10.1007/s00262-003-0444-1

Wei, Y., Chen, L., Xu, H., Xie, C., Zhou, Y., and Zhou, F. (2018). Mitochondrial Dysfunctions Regulated Radioresistance through Mitochondria-To-Nucleus Retrograde Signaling Pathway of NF-κB/PI3K/AKT2/mTOR. *Radiat. Res.* 190 (2), 204–215. doi:10.1667/rr15021.1

West, A. P., Shadel, G. S., and Ghosh, S. (2011). Mitochondria in Innate Immune Responses. *Nat. Rev. Immunol.* 11 (6), 389–402. doi:10.1038/nri2975

Xu, Y., Zou, R., Wang, J., Wang, Z. W., and Zhu, X. (2020). The Role of the Cancer Testis Antigen PRAME in Tumorigenesis and Immunotherapy in Human Cancer. *Cell Prolif* 53 (3), e12770. doi:10.1111/cpr.12770

Yang, S., Liu, T., Cheng, Y., Bai, Y., and Liang, G. (2019). Immune Cell Infiltration as a Biomarker for the Diagnosis and Prognosis of Digestive System Cancer. *Cancer Sci.* 110 (12), 3639–3649. doi:10.1111/cas.14216

Yoshida, Y., Hasegawa, J., Nezu, R., Kim, Y.-K., Hirota, M., Kawano, K., et al. (2011). Clinical Usefulness of Mitochondrial Transcription Factor A Expression as a Predictive Marker in Colorectal Cancer Patients Treated with FOLFOX. *Cancer Sci.* 102 (3), 578–582. doi:10.1111/j.1349-7006.2010.01835.x

Yu, M. (2011). Generation, Function and Diagnostic Value of Mitochondrial DNA Copy Number Alterations in Human Cancers. *Life Sci.* 89 (3-4), 65–71. doi:10.1016/j.lfs.2011.05.010

Zhang, Q., Li, H., Mao, Y., Wang, X., Zhang, X., Yu, X., et al. (2020). Apoptotic SKOV3 Cells Stimulate M0 Macrophages to Differentiate into M2 Macrophages and Promote the Proliferation and Migration of Ovarian Cancer Cells by Activating the ERK Signaling Pathway. *Int. J. Mol. Med.* 45 (1), 10–22. doi:10.3892/ijmm.2019.4408

Zhao, Q. Q., Jiang, C., Gao, Q., Zhang, Y. Y., Wang, G., Chen, X. P., et al. (2020). Gene Expression and Methylation Profiles Identified CXCL3 and CXCL8 as Key Genes for Diagnosis and Prognosis of colon Adenocarcinoma. *J. Cel Physiol* 235 (5), 4902–4912. doi:10.1002/jcp.29368

Zhao, L., Xue, M., Zhang, L., Guo, B., Qin, Y., Jiang, Q., et al. (2020). MicroRNA-4268 Inhibits Cell Proliferation *via* AKT/JNK Signalling Pathways by Targeting Rab6B in Human Gastric Cancer. *Cancer Gene Ther.* 27 (6), 461–472. doi:10.1038/s41417-019-0118-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Kang, li, Wang, Wei, Zeng, Luo, Sun, Xu, Peng and Zhou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GLOSSARY

ANKLE1 ankyrin repeat and LEM domain containing 1

AUC area under the curve

BP biological process

CALB2 calbindin 2

CC cellular composition

CIBERSORT Cell-type Identification by Estimating Relative Subsets of RNA Transcripts HR Hazard ratio

CTLA4 cytotoxic T lymphocyte-associated antigen-4

DEGs differentially expressed genes

ESTIMATE Estimation of Stromal and Immune cells in Malignant Tumor tissues using the Expression data

FDR false discovery rate

FPKM fragments per kilobase per million

GEO Gene Expression Omnibus

GO gene ontology analysis

GPRASP1 G protein-coupled receptor associated sorting protein 1

HLA human leukocyte antigen

HTSeq high-throughput sequence

ICIs immune checkpoint inhibitors

KEGG Kyoto Encyclopedia of Genes and Genomes

LASSO least absolute shrinkage and selection operator

LRRN2 leucine rich repeat neuronal 2

MCODE Molecular Complex Detection

MF molecular functions

mtDNA mitochondrial deoxyribonucleic acid

PD-L1 programmed death-ligand 1

PPI protein–protein interaction

PRAME PReferentially expressed Antigen in MElanoma

OS overall survival

RAB6B member RAS oncogene family 6B

ROC receiver operating characteristic

STRING Search Tool for the Retrieval of Interacting Genes database

TCF7L1 transcription factor 7-like 1

TFAM transcription factor A, Mitochondrial; TCGA The Cancer Genome Atlas

TIICs tumor-infiltrating immune cell

TME tumor microenvironment

frontiers | Frontiers in Neuroscience

REVIEW
published: 25 April 2022
doi: 10.3389/fnins.2022.843794

Genetic Regulation of Vertebrate Forebrain Development by Homeobox Genes

*Ryan F. Leung[1,2], Ankita M. George[1], Enola M. Roussel[1], Maree C. Faux[1,2,3], Jeffrey T. Wigle[4,5] and David D. Eisenstat[1,2,6,7,8]**

[1] Murdoch Children's Research Institute, The Royal Children's Hospital Melbourne, Parkville, VIC, Australia, [2] Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia, [3] Department of Surgery, Royal Melbourne Hospital, The University of Melbourne, Parkville, VIC, Australia, [4] Department of Biochemistry and Medical Genetics, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada, [5] Institute of Cardiovascular Sciences, St. Boniface Hospital Albrechtsen Research Centre, Winnipeg, MB, Canada, [6] Department of Medical Genetics, University of Alberta, Edmonton, AB, Canada, [7] Department of Pediatrics, University of Alberta, Edmonton, AB, Canada, [8] Department of Oncology, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB, Canada

Forebrain development in vertebrates is regulated by transcription factors encoded by homeobox, bHLH and forkhead gene families throughout the progressive and overlapping stages of neural induction and patterning, regional specification and generation of neurons and glia from central nervous system (CNS) progenitor cells. Moreover, cell fate decisions, differentiation and migration of these committed CNS progenitors are controlled by the gene regulatory networks that are regulated by various homeodomain-containing transcription factors, including but not limited to those of the *Pax* (paired), *Nkx*, *Otx* (orthodenticle), *Gsx/Gsh* (genetic screened), and *Dlx* (distal-less) homeobox gene families. This comprehensive review outlines the integral role of key homeobox transcription factors and their target genes on forebrain development, focused primarily on the telencephalon. Furthermore, links of these transcription factors to human diseases, such as neurodevelopmental disorders and brain tumors are provided.

Keywords: forebrain, development, homeobox, bHLH factor, forkhead (Fkh) transcription factors, DNA binding domain

OPEN ACCESS

Edited by:
Anthony LaMantia, Virginia Tech, United States

Reviewed by:
Tomasz Nowakowski, University of California, San Francisco, United States
Kenneth Campbell, Cincinnati Children's Hospital Medical Center, United States

***Correspondence:**
David D. Eisenstat
david.eisenstat@mcri.edu.au

Specialty section:
This article was submitted to Neurodevelopment, a section of the journal Frontiers in Neuroscience

Received: *30 December 2021*
Accepted: *14 March 2022*
Published: *25 April 2022*

Citation:
Leung RF, George AM, Roussel EM, Faux MC, Wigle JT and Eisenstat DD (2022) Genetic Regulation of Vertebrate Forebrain Development by Homeobox Genes. Front. Neurosci. 16:843794. doi: 10.3389/fnins.2022.843794

INTRODUCTION

Overview of Forebrain Development

Early brain development is marked by the formation of different compartments through the segmentation of the neural tube that is guided and defined by specific regional expression of transcription factors. The developing brain is sectioned into three contiguous parts, the prosencephalon in the most anterior area, which then matures into the forebrain; the mesencephalon following posteriorly, which give rises to the midbrain; and further posteriorly the rhombencephalon, the early form of the hindbrain. These areas further partition, where the prosencephalon separates into primary prosencephalon (diencephalon) and secondary prosencephalon (telencephalon) (Puelles, 2013, 2018), and the rhombencephalon divides into the metencephalon and myelencephalon. In contrast to the other two regions, the mesencephalon

does not divide (Stiles, 2008). Within the forebrain, the prosomeric model depicts the division of this area into 7 segments called the prosomeres (Rubenstein et al., 1994; Puelles and Rubenstein, 2003). The diencephalon develops into 3 prosomeres (p1, p2, p3), which are then recognized as the pretectum, thalamus and pre-thalamus. The secondary prosencephalon develops into two hypothalamo-telencephalic prosomeres (hp1, hp2), later giving rise to the hypothalamus and telencephalon. The mesencephalon contributes to two prosomeres (m1, m2) (Puelles, 2018).

The regions adjacent to the ventricular surface in the brain are the ventricular zone (VZ), followed by the subventricular zone (SVZ), and the mantle zone (MZ) (**Figure 1A**). The VZ contains radial glia, which then differentiate into intermediate neural progenitors that populate the SVZ, where both of these cell types can give rise to neurons (Miyata et al., 2001; Noctor et al., 2001, 2004; Haubensak et al., 2004). The telencephalon can be divided into the dorsal (pallium) and ventral (subpallium) telencephalon, where the neocortex and the ganglionic eminences (GE) are located, respectively. The anatomic region separating the dorsal and ventral telencephalon is often referred to as the pallio-subpallial boundary (PSB). The GE is divided into lateral, medial, and caudal GE (LGE; MGE; CGE), and ventral to the MGE is the preoptic area (PoA) (**Figure 1A**). The LGE can be further separated in the ventral LGE (vLGE), where striatal projection neurons originate, and the dorsal LGE (dLGE) that gives rise to intercalated cells of the amygdala and neurons in the olfactory bulb along with the lateral LGE wall (Yun et al., 2001; Stenman et al., 2003; Waclaw et al., 2010). The LGE is a local source of retinoic acid, a morphogen that regulates cortical patterning and regionalization (see Shibata et al., 2021; Ziffra et al., 2021 for more details) (Toresson et al., 1999; Molotkova et al., 2007; Shibata et al., 2021; Ziffra et al., 2021).

Origin of Cortical and Striatal Neurons

Excitatory and inhibitory neuronal activities need to be balanced in order for the nervous system to maintain homeostasis and to optimally process information; these are governed by projection and inhibitory neurons in the brain, respectively. Neuronal progenitor cells (NPC) are produced in both dorsal and ventral telencephalon; NPCs from the dorsal telencephalon give rise to projection neurons (glutamatergic) and NPCs from the ventral telencephalon differentiate into inhibitory interneurons (γ-amino butyric (GABA)-ergic) (Anderson et al., 1997a, 2002b). These neuronal origin sites are conserved amongst mammals, as shown through studies in primates, rodents, and humans, in which some cortical interneurons could be generated locally in the dorsal telencephalon (Letinic et al., 2002; Hansen et al., 2013; Ma et al., 2013). Glutamatergic neurons make up ∼ 70% of the neuronal population in the mouse, with the remaining ∼ 30% being GABAergic interneurons (Hendry et al., 1987). Within the ventral telencephalon, GABAergic interneurons are produced mainly from *Nkx2.1* expressing progenitor cells in the MGE and PoA (Fogarty et al., 2007; Gelman et al., 2009), and migrate tangentially to reach the neocortex (Marín and Rubenstein, 2003). These ventral telencephalic interneurons mainly consist of parvalbumin (pva^+), somatostatin (sst^+), and *5ht3a*+ interneurons subtypes (Rudy et al., 2011). Many *sst*+ interneurons arise and migrate from the CGE, while other interneuron subtypes arise from progenitor cells in the LGE and CGE, including the vasoactive intestinal peptide and cholecystokinin expressing interneurons which reside in the MZ (Anderson et al., 2001; Nery et al., 2002; Miyoshi et al., 2010). The main population of striatal projection neurons comprises the GABAergic medium spiny neurons (MSNs) which arise from progenitors in the LGE, and account for ∼ 80% of the striatal neuron population in primates and rodents (Graveland and DiFiglia, 1985). Some key marker genes for MSN differentiation include *Foxp1/2*, *Ascl1*, *Ebf1,* and *Meis2* (Garel et al., 1999; Toresson et al., 1999; Carri et al., 2013). The differentiation of MSNs is dependent on the temporal expression of a set of transcription factors, particularly the repressive function of *Dlx1/2* on *Ascl1* at specific timepoints, to promote differentiation and migration of striatal neurons (Anderson et al., 1997b; Yun et al., 2002). EBF1 then controls later differentiation and migration from the SVZ to the MZ (Garel et al., 1999).

Olfactory Bulb Neurogenesis

In mice, olfactory bulb neurogenesis occurs from embryonic until early postnatal stages, and is dependent on the neuronal types (Alvarez-Buylla and Lim, 2004; Tucker et al., 2006; Figueres-Oñate and López-Mascaraque, 2016). Initially, projection neurons are generated by E12.5, followed by the development of inhibitory interneurons by E14.5 (Bayer, 1983; Tucker et al., 2006). The olfactory bulb projection neurons, mitral/tufted (M/T) cells, originate from progenitor cells in the pallium and are differentiated from *Pax6*+ radial glia (Whitman and Greer, 2009; Imamura and Greer, 2013). M/T cells can adopt both radial and tangential migration. Earlier born neurons predominantly migrate radially and populate the deeper cortical layers, while later born projection neurons are more likely to migrate tangentially to the superficial cortical layer (Imamura et al., 2011). Migration of these projection neurons is regulated by a number of transcription factors, such as PAX6 and LHX2, which are also crucial for cortical neuron migration (Nomura et al., 2007; Saha et al., 2007). Transcription factors specific for olfactory bulb projection neuron migration include *Ap2-epsilon*, *Arx*, and *FezF1*, which are all important for proper orientation of M/T cells, as well as the expression of *Tbr1/2* (Yoshihara et al., 2005; Feng et al., 2009; Shimizu and Hibi, 2009; Imamura and Greer, 2013).

Olfactory bulb interneurons, in contrast to cortical interneurons, are derived from the dLGE, and postnatally in the SVZ, with the exception of *Emx1*+ pallial progenitors (Wichterle et al., 2001; Stenman et al., 2003). Subsequently, these interneurons tangentially migrate through to the olfactory bulb, postnatally through the rostral migratory stream (Kriegstein and Alvarez-Buylla, 2009). Although born in neuroanatomic regions distinct from cortical interneurons, olfactory bulb interneuron migration is regulated by a similar set of factors. Some of these include *Dlx1/2*, *Ascl1*, and *Robo-Slit* (Andrews et al., 2006; Long et al., 2007). Upon reaching the olfactory bulb, the interneurons differentiate into GABAergic interneurons and subsequently, subtype specification takes place (Lois and Alvarez-Buylla, 1994;

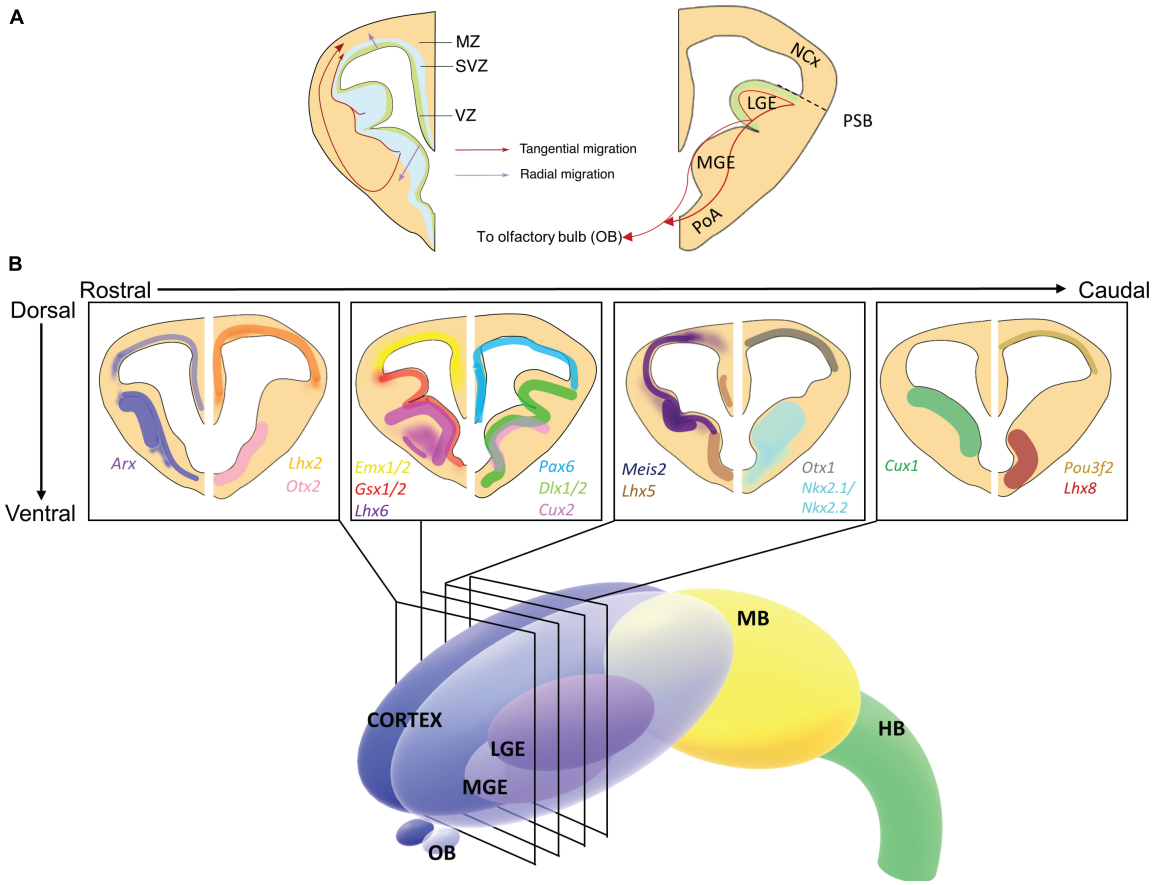


FIGURE 1 | Expression of homeobox genes in the developing embryonic mouse forebrain. **(A)** Schematic illustration of coronal section of E13.5 forebrain depicting ventricular zone (VZ), subventricular zone (SVZ), and mantle zone (MZ) on the left-hand side and neocortex (NCx), lateral ganglionic eminence (LGE), and medial ganglionic eminence (MGE) on the right-hand side. The VZ and SVZ are the proliferative zones, comprised of progenitor cells. Depending on the identity of these differentiated cells, the cells migrate either tangentially (red arrows) or radially (purple arrows) into the MZ and proceed to mature (Left-hand side). Migration toward the olfactory bulb from the VZ of the LGE also occurs (Right-hand side). **(B)** 3-dimensional schematic of the developing forebrain. The LGE and MGE are contained within the cortex, above the olfactory bulbs (OB). The midbrain (MB) and hindbrain (HB) are also labeled. Insets show schematic representations of 4 coronal sections taken from the forebrain depicting the expression of key homeobox gene expression patterns from rostral to caudal at embryonic time point E13.5. Gene name colors correspond to the expression color shown in the section. Transcription factor expression can be overlapping or structurally distinct and is related to the function of the individual transcription factor (Allen Institute for Brain Science, 2019). *Arx* and *Meis2*, to an extent, are expressed throughout the forebrain, whereas *Lhx2, Emx1/2, Pax6, Otx1,* and *Pou3f2* are expressed in the neocortex and pallium. *Dlx1/2, Gsx1, Otx2*, and *Cux1* are expressed in the GE, *Gsx2* is expressed specifically in the LGE, and *Nkx2.1, Cux2, Lhx6*, and *Lhx8* in the MGE. *Irx3* is not depicted here as it is expressed in the thalamus (not shown). For detailed depictions of gene expression patterns, readers are encouraged to review the cited primary references or the Allen Brain Atlas: Developing Mouse Brain (Allen Institute for Brain Science, 2019). NCx, neocortex; LGE, lateral ganglionic eminence; MGE, medial ganglionic eminence; V, ventricle; VZ, ventricular zone; SVZ, subventricular zone; MZ, mantle zone.

Sequerra, 2014) which is itself dependent on the developmental stage, i.e., whether born at an embryonic or postnatal stage (De Marchis et al., 2007; Batista-Brito et al., 2008). Examples of transcription factors that regulate interneuron development are *Sp8/Sp9* which are essential for olfactory bulb development (Li et al., 2017). For a more in-depth discussion about olfactory bulb development refer to a recent review from Tufo et al. (2022).

Radial and Tangential Migration of Neurons

There are two modes of neuronal migration, radial and tangential, classified by the axis of migration (**Figure 1B**). Cells move from the VZ toward the MZ generally by radial migration, and can descend within the VZ before migrating toward the MZ. *Radial migration* occurs during the development of the cerebral cortex, spinal cord, striatum and thalamus (Ayala et al., 2007). Morphological changes of interneurons mark the start of radial migration, whereas restriction of such changes also impairs the migration of these interneurons (LoTurco and Bai, 2006). Two different modes of movements are adopted during radial migration (Nadarajah et al., 2001). Interneurons migrate by somal translocation, by attaching to the outer surface of the developing brain (pial surface) and as microtubules shorten, the nucleus is pulled forward (Franco et al., 2011). Locomotion, on the other hand, allows interneurons to be guided by radial glial

cells toward the destination during the radial migration through complex forebrain structures (Rakic, 1972).

Tangential migration is adopted by cortical interneurons born in the GE, as these cells need to migrate from the GE to the neocortex while avoiding movement toward the striatum (DeDiego et al., 1994). Despite being derived in different areas, interneurons arising from the MGE, CGE and preoptic area have a similar transcriptome (Mayer et al., 2018), which could contribute to the similar migration pattern these interneurons adopt. Transcription factors tightly regulate the migration fate of interneurons, such as the expression or repression of *Nkx2.1* determines whether interneurons migrate into the striatum or

TABLE 1 | Summary of selected transcription factors required for forebrain development.

Gene family	Gene symbol	Human chromosome location[#]	Forebrain expression at E13.5	Forebrain gene function
Homeobox	*Arx*	Xp22.13	Cortex VZ; GE SVZ (Miura et al., 1997)	Promotes GABAergic interneuron tangential migration (Friocourt and Parnavelas, 2010; Olivetti and Noebels, 2012)
	Cux1	7q22.1	GE VZ and SVZ (Nieto et al., 2004; Zimmer et al., 2004)	Represses dendritic arborization (Coqueret et al., 1998)
	Cux2	12q24.11-q24.12	MGE SVZ	Controls neuronal specification and differentiation in the upper cortical layers (Zimmer et al., 2004)
	Dlx1	2q33.1 (Stock et al., 1996)	GE VZ and SVZ (Pleasure et al., 2000)	Regulates GABAergic interneuron specification and migration (Simeone et al., 1992; Anderson et al., 1997a,b)
	Dlx2	2q31.1 (Stock et al., 1996)	GE VZ and SVZ (Pleasure et al., 2000)	Regulates GABAergic interneuron specification and migration (Anderson et al., 1997a,b)
	Emx1	2p13.2	Cortex VZ (Simeone et al., 1992; Yoshida et al., 1997)	Dorsal forebrain specification and patterning (Yoshida et al., 1997; Stocker and O'Leary, 2016)
	Emx2	10q26.11	Cortex VZ and SVZ (Simeone et al., 1992; Yoshida et al., 1997)	Dorsal forebrain specification and patterning (Yoshida et al., 1997; Hamasaki et al., 2004)
	Gsx1	13q12.2	dLGE VZ (Toresson and Campbell, 2001)	Promote OPC proliferation (Chapman et al., 2018)
	Gsx2	4q12	vLGE VZ (Yun et al., 2001)	Promote neuron, oligodendrocyte, and glia specification (Kessaris et al., 2006; Fogarty et al., 2007; Chapman et al., 2018)
	Irx3	16q12.2	Thalamus (Robertshaw et al., 2013)	Promotes differentiation in the thalamus and neurogenesis at the paraventricular nucleus of the hypothalamus (Robertshaw et al., 2013; Smemo et al., 2014)
	Lhx2	9q33.3	Cortex VZ and SVZ (Roy et al., 2014)	Progenitor cell proliferation; dorsal patterning (Godbole et al., 2018)
	Lhx5	12q24.13	Ventral forebrain (Sheng et al., 1997)	Hippocampal neuron differentiation and migration (Abellán et al., 2010)
	Lhx6	9q33.2	MGE SVZ (Matsumoto et al., 1996)	Regulates GABAergic interneuron differentiation and migration (Alifragis et al., 2004; Zhao et al., 2008; Neves et al., 2013)
	Lhx8	1p31.1	MGE MZ (Matsumoto et al., 1996)	Regulates cholinergic interneuron differentiation and specification (Zhao et al., 2003; Fragkouli et al., 2005; Lopes et al., 2012)
	Meis2	15q14	Cortex VZ; LGE, MGE, and CGE (Cecconi et al., 1997; Toresson et al., 1999; Agoston et al., 2014)	Controls gene expression and promotes differentiation and migration of neurons (Agoston et al., 2014)
	Nkx2.1	14q13.3	MGE and PoA (Xu et al., 2005)	Ventral forebrain specification and patterning (Nóbrega-Pereira et al., 2008; Kanatani et al., 2015)
	Nkx2.2	20p11.22	MGE (Ericson et al., 1997)	Promotes GABAergic interneuron specification (Briscoe et al., 1999; Robertshaw et al., 2013)
	Otx1	2p13 (Kastury et al., 1994)	Cortex VZ (Hoch et al., 2015)	Dorsal forebrain specification and patterning (Larsen et al., 2010b)
	Otx2	14q21-22 (Kastury et al., 1994)	GE VZ (Hoch et al., 2015)	Ventral forebrain specification and patterning (Larsen et al., 2010b)
	Pax6	11q13	Cortex VZ (Hirata et al., 2002)	Dorsal forebrain specification and patterning (Scardigli et al., 2003)
	Pou3f2	6q16.1	Cortex VZ (Nakai et al., 1995; Dominguez et al., 2012)	Regulates neuronal differentiation and radial migration in the telencephalon (Artavanis-Tsakonas et al., 1999)
bHLH	*Ascl1*	12q23.2	GE VZ (Fode et al., 2000; Britz et al., 2006)	Interneuron specification from neural progenitor cells (Nieto et al., 2001; Bertrand et al., 2002)
	Olig1	21q22	GE VZ and SVZ (Takebayashi et al., 2000)	Promotes oligodendrocyte differentiation and specification (Tekki-Kessaris et al., 2001; Anderson et al., 2002a; Lu et al., 2002)
	Olig2	21q22	GE VZ and SVZ (Takebayashi et al., 2000)	Promotes oligodendrocyte differentiation and interneuron specification (Tekki-Kessaris et al., 2001; Anderson et al., 2002a; Lu et al., 2002)
	Olig3	6q24	Dorsal thalamus (Takebayashi et al., 2000)	Promotes interneuron specification (Takebayashi et al., 2002; Lowenstein et al., 2021)
Forkhead	*Foxg1*	14q12	Ventral forebrain (Tao and Lai, 1992)	Ventral forebrain specification; regulates neuron migration and specification (Martynoga et al., 2005; Kumamoto and Hanashima, 2017)

[#]*According to NCBI database (NCBI Datasets, 2021).*

neocortex, respectively (Nóbrega-Pereira et al., 2008). There are two major paths for interneurons to migrate from the GE to the developing neocortex, through a superficial route that bypasses the MZ or a deeper route that passes through the SVZ (**Figure 1A**; Wichterle et al., 2001). These migration paths are guided by signaling molecules such as the chemokine CXCL12, which attract interneurons, and its receptor CXCR4. Studies have shown that disruption of CXCL12 or its receptor CXCR4 led to interneuronal mislocalization (Stumm et al., 2003; López-Bendito et al., 2008; Wang et al., 2011b). Furthermore, *Tbr2*+ cortical intermediate progenitor cells may actively attract interneuron migration into the cortex, which is concurrently modulated by CXCL12 signaling (Sessa et al., 2010). Another chemokine, Neuregulin 3 (*Nrg3*), mediated by ErbB4 attracts and regulates the final destination of GABAergic interneurons in the cortex (Rakić et al., 2015). Similarly, repulsive guidance cues Semaphorin 3A and 3F also play a role in guiding interneuron tangential migration, where their expression in the LGE prevents interneuron migration toward the basal area (Chen et al., 2008). This repulsion is achieved by the interactions between these molecules and their receptors neuropilin-1 (*Nrp1*) and neuropilin-2 (*Nrp2*), which are expressed in migrating interneurons (Marín et al., 2001). Some other extrinsic factors act as mitogens to provide motility and control the rate of migration for interneurons, such as the hepatocyte growth factor/scatter factor (Powell et al., 2001). Furthermore, GABA itself can act as a motogen and accelerate tangential migration (Inada et al., 2011). These processes that direct neuron fate determination are ultimately regulated by members of the homeobox and basic helix-loop-helix (bHLH) transcription factor families (**Table 1**).

Homeobox Genes

Homeobox genes are an important gene family for embryonic development, defined by a conserved homeodomain (HD) containing a helix-loop-helix-turn-helix structure (Gehring et al., 1994; Noyes et al., 2008). The 60 amino acid HD is commonly located at the carboxyl terminal end of the protein, and binds DNA primarily through the 50th residue, usually being a glutamine, allowing homeobox genes to function as transcription factors (**Figure 2**; Kappen, 2000). This DNA binding motif is located in the second and third helices, which recognizes and binds to the major groove of DNA at specified consensus sites (**Table 2**). Further, the N-terminal arm contributes to the binding strength through interactions with the DNA minor groove, typically through a basic residue such as arginine at the 5^{th} residue in the HD (Rohs et al., 2009). Apart from the consensus binding sequence, other important factors for DNA binding specificity include cofactors and additional DNA binding domains, such as the paired domain (PRD) in PAX superfamily members. Water molecules have been shown to be crucial for the HD to bind DNA (Billeter et al., 1996). Protein-protein interactions driven by the flanking regions around HD also increase the specificity of DNA binding (Li et al., 1995; Amin et al., 2015; Merabet and Lohmann, 2015). Homeobox proteins often contain other domains apart from the HD, which provide additional DNA specificity for these proteins, and have allowed characterization of homeobox proteins into 11 different classes, such as the Antennapedia (ANTP), Paired (PRD), LIM and NK classes (Holland et al., 2007), and can be further divided into different families within these classes. Large functional and comparative genomics studies have enabled analyses of these proteins, and allowed accurate annotation, naming and classification of homeobox genes (Holland et al., 2007).

Basic Helix-Loop-Helix Genes

Basic helix-loop-helix (bHLH) proteins are another superfamily of transcription factors present in most eukaryotes, with critical functions during embryonic development, such as neurogenesis and myogenesis. bHLH domains contains two alpha helices, helix 1 and helix 2. These helices are connected by a short loop, and at the amino-terminal end of helix 1 is a basic region (Murre et al., 1989). This basic region binds DNA by recognizing a core CANNTGG motif, known as an *E-box motif*, and is specific for different transcription factors (**Table 2**). Upon binding, the basic region is fitted into the major grove of the DNA. The HLH domain interacts with other proteins, forming different homo- or hetero-dimeric complexes that are required for DNA binding (Ellenberger et al., 1994). The unique combinations of these bindings give rise to the diverse transcriptional regulatory functions of bHLH proteins during development. bHLH proteins can be roughly divided into those that are either cell-type specific or widely expressed where the group of transcription factors governing neuron development are often referred to as proneural proteins (Lee, 1997; Srivastava et al., 1997).

SELECTED TRANSCRIPTION FACTORS ENCODED BY HOMEOBOX GENES

In the following major section of this comprehensive review, detailed summaries of 21 homeobox genes (in alphabetical order) that encode homeodomain containing transcription factors are provided. These genes were selected due to their essential role in forebrain development. However, we acknowledge that this selection of genes excludes several other important homeobox genes as well as key bHLH (*Ascl1*, *Olig1*, *Olig2,* and *Olig3*) and forkhead (*Foxg1*) genes required for neurodevelopment. For this reason, we have included *Ascl1*, *Olig1/2/3,* and *Foxg1* in **Figure 2** and the Tables.

A brief note about gene and protein nomenclature is useful. By consensus: mouse gene, *Dlx*; zebrafish gene, *dlx*; human gene, *DLX*; mouse and human protein, DLX.

Aristaless Related Homeobox Gene

The Aristaless related homeobox (*Arx*) paired-like HD transcription factor is the vertebrate homolog of the *Drosophila* aristaless (*al*) gene, which is essential for appendage formation (Miura et al., 1997). The gene is located on human chromosome Xp22.13 and is reported to be involved in neurological disorders such as X-linked intellectual disabilities (**Table 3**; Bienvenu et al., 2002; Friocourt and Parnavelas, 2010). In vertebrate embryogenesis, *Arx* transcriptionally regulates interneuron specification and migration (Fulp et al., 2008; Friocourt and Parnavelas, 2010; Olivetti and Noebels, 2012). ARX contains

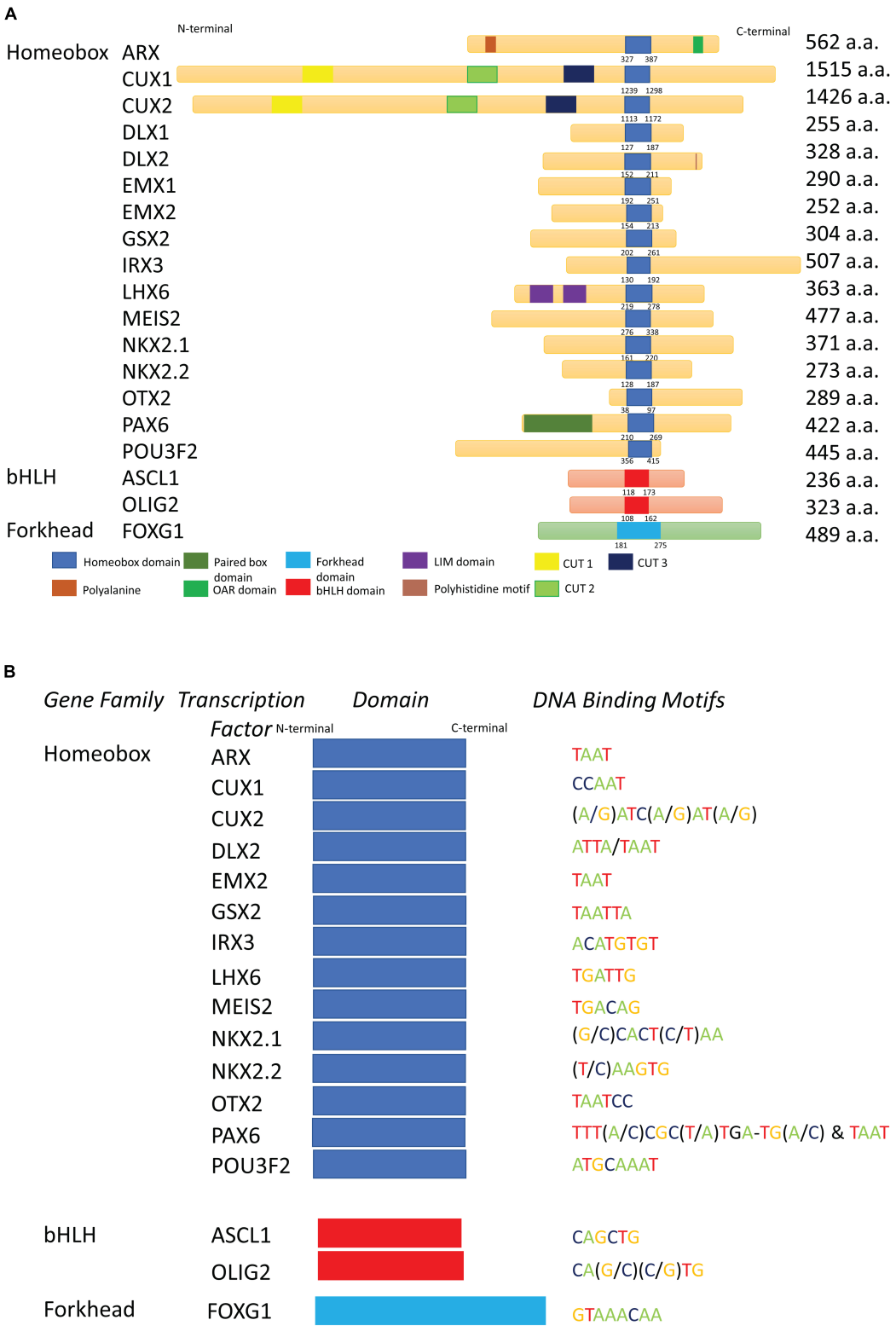


FIGURE 2 | Homeobox transcription factors and their key functional domains. **(A)** Schematic depiction of the domain structure of selected homeobox, bHLH and forkhead transcription factors illustrating the highly conserved nature of the homeodomains and other DNA-binding domains of these transcription factors. Other significant functional domains are also shown. DLX2, but not DLX1, contains a polyhistidine motif in its C-terminus. PAX6 contains both a paired domain, as well as the HD. ARX, along with a HD, also contains a polyalanine repeat and an Aristaless domain. a.a., amino acid. **(B)** The DNA binding sites for Homeobox, bHLH and Forkhead transcription factors from **(A)**. A consensus DNA binding motif for DLX1 and EMX1 is not available.

TABLE 2 | DNA binding motifs and selected target genes.

Gene family	Gene symbol	DNA binding motif	Target genes
Homeobox	*Arx*	TAAT (Cho et al., 2012)	*Cxcr4; Cxcr7; Dlx2; Ebf3; Lhx8* (Fulp et al., 2008; Colasante et al., 2009; Quille et al., 2011)
	Cux1	CCAAT (Moon et al., 2000)	*Nfib; Fesf2; Pou6f2; Sox5* (Gray et al., 2017)
	Cux2	(A/G)ATCAAT (Conforto et al., 2012)	*Xlr3b; Xlr4b* (Cubelos et al., 2010)
	Dlx2	ATTA/TAAT (Zhou et al., 2004)	*Dlx5/6; Gad1/2; Gsx2; Lhx6/8; Nrp2; Olig2; Otx2; Pax6* (Petryniak et al., 2007; Long et al., 2009; Lindtner et al., 2019)
	Emx1		*Nrp1* (Lim et al., 2015)
	Emx2	TAAT (Beckmann et al., 2011)	*Gsx2; Sox2; Ten-1* (Mariani et al., 2012; Desmaris et al., 2018)
	Gsx2	TAATTA (Salomone et al., 2021)	*Dbx1; Dlx1/2* (Corbin et al., 2000; Toresson et al., 2000; Yun et al., 2001)
	Irx3	ACATGTGT (Bilioni et al., 2005)	*Sox14; Gbx2* (Robertshaw et al., 2013; Smemo et al., 2014)
	Lhx8	TGATTG (Park et al., 2012)	*Lhx6; Shh* (Zhao et al., 2003; Flandin et al., 2011)
	Meis2	TGACAG (Chang et al., 1997)	*Dlx1/2; Dlx5/6* (Ghanem et al., 2003)
	Nkx2.1	(G/C)CACT(C/T)AA (Manoli and Driever, 2014)	*Gbx1/2; Gli2; Lhx6/8; Pax6; Nrp1/2* (Nóbrega-Pereira et al., 2008; Kanatani et al., 2015; Sandberg et al., 2016)
	Otx2	TAATCC/T (Briata et al., 1999)	*Arx; Dbx1; Dlx1/2; Fgf8; Hes1; Nkx2.1; Olig1/2; Pax3; Ten-*C (Gherzi et al., 1997; Hoch et al., 2015)
	Pax6	TTT(A/C)CGC(T/A)TGA-TG(A/C) and TAAT (Sun et al., 2015)	*Ascl1; Dlx2; Emx1/2; Ngn2; Pax6* (Scardigli et al., 2003; Sun et al., 2015)
	Pou3f2	ATGCAAAT (Herr and Cleary, 1995)	*Ascl1; Trim8; Vrk2* (Artavanis-Tsakonas et al., 1999; Chen et al., 2018; Pearl et al., 2019)
bHLH	*Ascl1*	CAGCTG (Webb et al., 2013)	*Ccng2; Cdk1/2Dlx2; EphB2; E2f1; Gadd45g; Hipk2; NeuroD; Ngn1* (Cau et al., 2002; Webb et al., 2013; Park et al., 2017)
	Olig1	CA(G/A)NTG (Li et al., 2007)	*Dlx1/2* (Silbereis et al., 2014)
	Olig2	CA(G/C) (C/G)TG (Küspert et al., 2011)	*Irx3; Ngn2; Nkx2.2; Pax6; Sox10; Zep2* (Küspert et al., 2011; Emery and Lu, 2015)
Forkhead	*Foxg1*	GTAAACAA (Dai et al., 2020)	*Ascl1; Cxcr4/7; Ccnd1; Dlx1/2; Eph44; Fgf8; Prox1; Robo1; Sema3A/F* (Bulstrode et al., 2017; Yang et al., 2017; Hou et al., 2019)

multiple structural domains and motifs, including the HD, a PRD-like domain, an N-terminal octapeptide domain, a central acidic domain and the C-terminal aristaless domain as well as three nuclear localization sequences and four polyalanine (polyA) tracts (Miura et al., 1997; **Figure 2**). ARX binds the transcriptional co-repressor TLE1, through the TLE1 octapeptide domain, and recognizes DNA at TAAT sites (Jennings et al., 2006; McKenzie et al., 2007; Cho et al., 2012). *In vitro* assays show that although ARX can be phosphorylated at multiple sites, it is unclear whether ARX functions are regulated by its phosphorylation state (Mattiske et al., 2018; Shi et al., 2020).

Aristaless Related Homeobox is expressed in various parts of the developing forebrain, such as the SVZ in developing GE and the VZ in the neocortex (**Figure 1B**; Miura et al., 1997; Colombo et al., 2004). *Arx* expression in the neocortex is limited to the proliferating neural progenitor cells, and is suppressed in cells radially migrating from the VZ (Friocourt et al., 2006), whereas in the GE *Arx* is continually expressed after neuronal differentiation and migration. Embryonic mice with homozygous *Arx* mutations have small brains with a thin neocortex and die upon birth, which may be related to defective tangential migration of cortical interneurons (Kitamura et al., 2002; Colombo et al., 2007; Friocourt et al., 2008). Targeted conditional deletion of *Arx* in the neocortex results in intermediate progenitor cell proliferation, with a reduced population of cortical neural progenitors. ARX also directly regulates cortical progenitor cell expansion through transcriptional regulation of CDKN1C, a cell cycle progression inhibitor in cortical VZ and SVZ (Colasante et al., 2015). The expression pattern of *Arx* also reveals its contribution to establishing the dorsoventral identity of the developing brain, where *Arx* suppresses ventralization in the dorsal forebrain by repressing *Olig2* expression. *Olig2* is a ventral specific gene and its expression is induced through Sonic Hedgehog (SHH) signaling. The expression of SHH downstream targets, *Gli1* and *Ptch3*, are increased in *Arx* cKO mice dorsal telencephalon. Thus ARX represses these SHH downstream signals, and in turn represses *Olig2* expression (Lim et al., 2019). Both inactivation, through shRNA, and overexpression of *Arx* impact GABAergic interneuron tangential migration to the neocortex from the MGE (Colombo et al., 2007). Furthermore, *Arx* is a direct regulatory target of DLX2, another homeobox transcription factor that regulates tangential migration, where overexpression of *Dlx2* increases *Arx* levels and reduction of *Dlx2* expression reduces *Arx* in the GE (Cobos et al., 2005a). By gain- and loss-of-function analysis, *Arx* was demonstrated to mediate the tangential interneuronal migration driven by DLX2, but not GABAergic neuron specification (Colasante et al., 2008). Conditional deletion of *Arx* in the ventral telencephalon further supports a role for *Arx* in tangential migration resulting in an overall reduction in the number of mature interneurons (Marsh et al., 2016). Additionally, ARX has been shown to transcriptionally regulate genes important for migration, such as *Cxcr4, Cxcr7*, *Ebf3*, and *Lhx7* (Fulp et al., 2008; Colasante et al., 2009; Quille et al., 2011).

Aristaless Related Homeobox mutations can lead to severe neurological diseases, including X-linked intellectual disability,

TABLE 3 | Forebrain mutant phenotypes and related diseases.

Gene family	Gene symbol	Forebrain mutant phenotype description	Related neural diseases
Homeobox	*Arx*	Disrupted GABAergic interneuron migration; structural brain malformation (Kitamura et al., 2002; Colombo et al., 2007)	Epilepsy; X-linked lissencephaly or intellectual disability (Kato et al., 2004; Friocourt et al., 2008)
	Dlx1/2	Disrupted GABAergic interneuron differentiation and migration (Anderson et al., 1997a,b; MacDonald et al., 2013)	Down Syndrome; epilepsy; Rett Syndrome; schizophrenia (Cobos et al., 2005b; Poitras et al., 2007)
	Emx1	Absence of corpus callosum; postnatal cKO disrupted cortex patterning (Qiu et al., 1996; Stocker and O'Leary, 2016)	Tumor suppressor for glioblastoma (Jimenez-García et al., 2021b)
	Emx2	Reduced cortex size; impaired radial migration (Falcone et al., 2016; Monnier et al., 2018)	Tumor suppressor for glioblastoma (Jimenez-García et al., 2021a)
	Gsx2	Reduced LGE size; reduced amount of GABAergic interneurons (Yun et al., 2003)	Basal ganglia malformation; Parkinson's Disease; Huntington's Disease (Zuccoli et al., 2015; De Mori et al., 2019)
	Lhx5	Impaired hippocampus formation (Abellán et al., 2010)	
	Lhx6	Reduced GABAergic interneuron subtype amount; disrupted interneuron migration (Liodis et al., 2007; Neves et al., 2013)	Tourette Syndrome; schizophrenia (Volk et al., 2014; Donegan et al., 2020; Pagliaroli et al., 2020)
	Lhx8	Impaired interneuron differentiation (Manabe et al., 2007, 2008)	Tourette Syndrome (Pagliaroli et al., 2020)
	Nkx2.1	Increased amount of GABAergic interneuron (Sussel et al., 1999; Fragkouli et al., 2009)	Schizophrenia; impaired learning and memory (Sussel et al., 1999; Fragkouli et al., 2009; Malt et al., 2016; Magno et al., 2017)
	Otx1	Reduced cortex size; reduced cell population (Acampora et al., 1996; Pantò et al., 2004)	Medulloblastoma; spontaneous epilepsy and seizures (Boon et al., 2005; Zakrzewska et al., 2013)
	Otx2	Disrupted septum and cortex formation (Acampora et al., 1997)	Medulloblastoma (Boon et al., 2005; Zagozewski et al., 2020)
	Pax6	Disrupted cortex formation; thinned cortex; small eyes (Hill et al., 1991; Tyas et al., 2003; Quinn et al., 2007)	Autism; impaired audition; intellectual disability (Malandrini et al., 2001; Davis et al., 2008)
bHLH	*Ascl1*	Reduced *Dlx1/2* expression; impaired interneuron migration (Nieto et al., 2001; Bertrand et al., 2002)	Parkinson's Disease (Ide et al., 2005)
	Olig1	Increased amount of GABAergic interneuron (Lu et al., 2000; Silbereis et al., 2014)	Down Syndrome (Haydar and Reeves, 2012)
	Olig2	Absence of OPCs (Furusho et al., 2006; Petryniak et al., 2007; Ono et al., 2008)	Down Syndrome; DMG (Lu et al., 2000; Filbin et al., 2018)
Forkhead	*Foxg1*	Reduced cortex size; impaired cortical cell proliferation (Xuan et al., 1995; Hanashima et al., 2002)	Autism; FoxG1 Syndrome; Rett Syndrome; schizophrenia; seizures; West's Syndrome variants (Neul et al., 2010; Florian et al., 2011; Striano et al., 2011; Mariani et al., 2015; Won et al., 2016)

epilepsy, as well as structural brain malformations (**Table 3**; Friocourt and Parnavelas, 2010), and these mutations have been studied extensively using mouse models (Kitamura et al., 2002, 2009; Marsh et al., 2009; Price et al., 2009). The phenotypes related to *ARX* mutations can be grouped based on whether there is a corresponding malformation. Disorders in the malformation group include X-linked lissencephaly associated with abnormal genitalia (Kitamura et al., 2002) and Proud syndrome (Kato et al., 2004), whereas the non-malformation group includes epilepsy, non-syndromic X-linked intellectual disability, and X-linked Infantile Spasms Syndrome (Bienvenu et al., 2002; Kitamura et al., 2009; Price et al., 2009) and different epilepsy syndromes such as West syndrome (Strømme et al., 2002; Kato et al., 2003). Many mutations in *ARX* have been found in the first two polyA tracts, where the polyA tracts are expanded by insertion of either additional alanine or other residues (Kitamura et al., 2009). A common mutation consists of an in-frame 24bp duplication (Szczaluba et al., 2006), whilst longer mutations, 27bp, and 33bp have also be reported (Demos et al., 2009; Reish et al., 2009). The longest known mutation exhibits the addition of eleven alanine residues, resulting in Ohtahara syndrome (Kato et al., 2007). Other intellectual disability, seizures related disorders have also been observed (Turner et al., 2002). In summary, these *ARX* mutations disrupt DNA and protein binding ability, perturbing the transcriptional activity of ARX, thereby affecting cortical development (Nasrallah et al., 2012; Siehr et al., 2020).

Cut-Like Homeobox Genes

The Cut-like homeobox genes encode a transcription factor family [*Cux homeobox 1/2* (*Cux1*/2)], previously called *CCAAT-displacement protein (CDP) or Cut-like homeobox 1/2 (Cut1/2)*, that are the mammalian homologs of the *Drosophila* gene *cut* locus (*ct*) (Blochlinger et al., 1988). *Ct* is responsible for controlling the fate of neuronal progenitor cells in the peripheral nervous system and external sensory organs in *Drosophila* (Bodmer et al., 1987; Blochlinger et al., 1988) and plays a crucial role in dendritic arborization of specific sensory neurons (Grueber et al., 2003). *CUX1* is located on human chromosome 7q22 and is frequently rearranged in cancers (Scherer et al., 1993), while *CUX2* is on chromosome band 12q24.11-q24.12 (Craddock et al., 1993). CUX transcription factors contain up to four DNA binding regions, comprised of one HD, including a histidine residue at the 9th amino acid of the third helix (Blochlinger et al., 1988), and one, two, or three highly homologous Cut repeats of approximately 70 amino acids (CR1, CR2, CR3) (**Figure 2A**; Nepveu, 2001). However, individual Cut repeats are unable to bind to DNA on their own but interact with other Cut repeats or with the Cut HD to bind DNA (Moon et al., 2000). CR1/CR2

mediate transient binding to DNA (Moon et al., 2000) and the CR3 repeat and the HD have been reported to form bipartite high affinity DNA binding interactions (Harada et al., 1994, 1995). *Cux1* and *Cux2* splice variants encode for protein isoforms with different combinations of DNA binding domains (Weiss and Nieto, 2019). Proteolytic cleavage of the full length p200 CUX1 protein generates a p110 protein which contains CR2, CR3 and the HD (Goulet et al., 2004). While the full-length p200 protein acts as a transcriptional repressor, p110 can act as repressor or activator depending on the type of promoter it interacts with (Yoon and Chikaraishi, 1994; Truscott et al., 2004, 2007; Harada et al., 2007). CUX proteins can act as transcriptional repressors either indirectly by competing with transcriptional activators for binding to target sites, or actively suppressing transcription *via* a mechanism that involves recruiting histone deacetylases through the Ala, Pro-enriched carboxyl domain (Cowell and Hurst, 1994; Mailly et al., 1996; Nepveu, 2001). CUX transcriptional activity is regulated by post-translational modifications at the Cut repeats which include acetylation, proteolysis (Sansregret et al., 2010), and phosphorylation by PKC (Coqueret et al., 1996), CKII (Coqueret et al., 1998), cAMP-dependent protein kinase (Michl and Downward, 2006), and cyclin A/Cdk1 (Santaguida et al., 2001), which repress transcriptional activity.

Cux1 expression is detected widely in embryonic and adult tissues (Nieto et al., 2004), while *Cux2* is more specifically expressed in the nervous system (Quaggin et al., 1996) as well as the limb buds and urogenital system (Iulianella et al., 2003). *Cux1 and Cux2* are expressed early during brain development in neural progenitor cells in the ventral and dorsal telencephalon, as early as E14 for *Cux1* and E10.5 for *Cux2*, specifically *Cux1* is expressed in the VZ and SVZ of whole GE (Nieto et al., 2004; Zimmer et al., 2004; **Figure 1**). In contrast, *Cux2* is solely expressed in the SVZ of the MGE, and is enriched in tangentially migrating cortical interneurons (Nieto et al., 2004; Zimmer et al., 2004). Indeed, *Cux2* is mostly expressed in SVZ/IZ early during development while it is later expressed across most of the cortex (Zimmer et al., 2004). Furthermore, *Cux2* expression distinguishes two cortical neuronal subpopulations with different origins, migration models, and phenotypic characteristics: a population of tangentially migrating GABAergic cortical interneurons and another DLX-negative neuronal population produced in the pallium, which migrates radially, divides in the SVZ and accumulates in the IZ (Zimmer et al., 2004).

In addition to controlling neural specification and differentiation in upper cortical layers, CUX proteins can act as repressors for developmental processes such as dendritic arborization (Grueber et al., 2003; Cubelos et al., 2010; Li et al., 2010). Overexpression of *Cux1*, but not *Cux2*, results in decreased dendritic arborization in cultured cortical pyramidal neurons, whereas dendritic complexity increases upon reduction of *Cux1* (Li et al., 2010). A mechanism whereby *Cux1* transcriptionally represses dendritic arborization is through suppression of the cyclin-dependent kinase inhibitor $p27^{Kip7}$ and further plays a role in proliferating cells by repressing the p21 cyclin kinase inhibitor (Coqueret et al., 1998).

Cux2 is regulated by PAX6 and contributes to determining the upper layers (II-IV) of the cortex (Zimmer et al., 2004). Deletion of either *Cux1* or *Cux2* in mice does not alter overall cortical and brain organization (Cubelos et al., 2008a), whereas most *Cux1* and *Cux2* double homozygous mutants die prior to birth (Cubelos et al., 2008b). Although, the few pups that survive P0 do not display defects in neuronal migration or in layer specific protein expression (Cubelos et al., 2008b), *Cux1*/*Cux2* double knockout (DKO) mice display abnormal dendrites and synapses indicating a critical role for *Cux* genes in dendritogenesis (Cubelos et al., 2010). The formation of cortical interneurons in *Cux* single and double mutants is impaired while loss of Reelin expression is only observed in upper cortical layers II-IV in double mutants (Cubelos et al., 2008b).

Cux2 deficient mice display increased brain volume, cell density and thickness of the upper cortical layers (II-IV), caused by an increase in the number of neuronal progenitor cells (Cubelos et al., 2008a). CUX1 target genes include *Nfib*, *Fezf2*, *Pou6f2* and *Sox5* which are all transcriptional regulators highly expressed in lower layers of the cortex (Gray et al., 2017). In addition to regulating upper cortical layer formation, *Cux2* has also been shown to control cell cycle exit (Cubelos et al., 2008a). Therefore, *Cux1* and *Cux*2 regulate neuronal proliferation of intermediate neuron precursors in SVZ, as well as the proliferation rate of neuronal precursor cells fated to form pyramidal cortical neurons in the upper layers of the cortex (Cubelos et al., 2008a,b) and in the spinal cord (Iulianella et al., 2008).

Mutations in *CUX1* have been associated with global developmental delay with or without impaired intellectual development (GDI) (Platzer et al., 2018) while *CUX2* is associated with intellectual disorders, seizures, autism spectrum disorder and bipolar affective disorder (Glaser et al., 2005; Barington et al., 2018). *CUX1* has also been shown to undergo inactivating mutations and loss of heterozygosity (LOH) in a number of human cancers (Ramdzan and Nepveu, 2014; Wong et al., 2014). Loss of *CUX1* activates the phosphoinositide-3-kinase (PI3K) signaling pathway as a result of transcriptional downregulation of the PI3K inhibitor, PIK3Ip1 (Wong et al., 2014). This mutation in *CUX1* results in increased tumor growth and increased susceptibility to PI3K-Akt inhibition (Wong et al., 2014). CUX1 has also been implicated in the regulation of proteosome-mediated degradation of the Src tyrosine kinase resulting in altered tumor cell migration and invasion (Aleksic et al., 2007).

Distalless Genes

Distalless (*dll)* was discovered in *Drosophila* for its essential role in limb development (Cohen et al., 1989). *Dlx* genes are the vertebrate orthologs of *dll*; six members of this gene family can be found in humans and mice, occurring as bigenic clusters (*Dlx1/2, Dlx3/4*, *and Dlx5/6*); however, only *Dlx1*, *Dlx2*, *Dlx5,* and *Dlx6* are expressed in the forebrain (**Figure 1B**). *Dlx1/2* and *Dlx5/6* are located on mouse chromosomes 2 and 6, and on human chromosomes 2q31.1 and 7q21.3, respectively (Stock et al., 1996). These bigenic clusters are organized from tail-to-tail, with highly conserved intergenic enhancers located between the two genes. *Dlx1/2* and *Dlx5/6* each contain two intergenic enhancers: i12a and i12b for *Dlx1/2,* and i56a and i56b for *Dlx5/6* (Ghanem et al., 2003; Ruest et al., 2003). These *cis*-regulatory

elements, although dissimilar in sequence, have overlapping activity and are essential for the expression of these genes (Fazel Darbandi et al., 2016). *Dlx5/6* expression is regulated by *Dlx1/2,* where the absence of *Dlx1/2* reduces *Dlx5/6* expression through the intergenic enhancer, revealed using gene reporter systems (Zerucha et al., 2000; Zhou et al., 2004). Likewise, removing the intergenic enhancers with a targeted mutation attenuates *Dlx5/6* expression in the forebrain, suggesting these intergenic enhancers are necessary for *Dlx* expression (Robledo et al., 2002; Ghanem et al., 2003). *Dlx* transcription factors are expressed in the developing GE and are essential for forebrain development (Pleasure et al., 2000). From embryonic day 9.5 (E9.5), expression is induced in the order of *Dlx2*, *Dlx1*, *Dlx5*, and *Dlx6* (Eisenstat et al., 1999). In mice, *Dlx1/2* are expressed in the VZ in the GE, and are clearly separated at the pallio-subpallial boundary (**Figure 1B**). *Dlx5/6* are expressed in the MZ of the ventral telencephalon, and additionally, *Dlx1, Dlx2, and Dlx5* are expressed in the SVZ in an overlapping manner, coinciding with regions where GABAergic interneurons are produced (Liu et al., 1997; Acampora et al., 1999; Depew et al., 1999; Robledo et al., 2002; Cobos et al., 2005b; Weinschutz Mendes et al., 2020).

Dlx1 and *Dlx2* single gene homozygous knockout (KO) mice die prematurely at postnatal day 0 (P0) with minor abnormalities in GABAergic neuron formation, demonstrating DLX1 and DLX2 are somewhat functionally redundant (Qiu et al., 1997). Cortical neurons are reduced in postnatal *Dlx1* KO mice which can lead to seizures (Cobos et al., 2005b). *Dlx1/2* and *Dlx5/6* double homozygous mutants also die at P0 with a more significant forebrain defect compared to the single KO mice. Tangential interneuron migration from the MGE to the neocortex is blocked in *Dlx1/2* double homozygous mutants both in mice and zebrafish, hindering GABAergic interneuron development (Anderson et al., 1997a,b; MacDonald et al., 2013). *Dlx1/2* double homozygous mutants also have reduced *Dlx5/6* expression, which results in altered progenitor cell fate in the dorsal and ventral telencephalon (Pla et al., 2017). *Dlx5/6* double homozygous mutant mice also exhibit tangential migration defects, with poor specification of parvalbumin GABAergic interneuron subtypes (Wang et al., 2010). Therefore, *Dlx* genes are essential for the differentiation of GABAergic neurons and their subsequent tangential migration (Anderson et al., 1997a,b; Marin et al., 2000).

Distalless genes transcription factors promote interneuron production by regulating transcription of various downstream targets in the ventral telencephalon, binding to the core HD DNA binding motif ATTA/TAAT (Zhou et al., 2004; **Table 2**). A recent report has found that DLX2 binds preferentially to transcription factors to mediate both its' repression and activation functions (Lindtner et al., 2019). GABA is synthesized by glutamic acid decarboxylases 1 and 2 (GAD1; GAD2) which are co-expressed with *Dlx1/2* in the VZ and SVZ of the GE in mouse, zebrafish, and humans (Erlander et al., 1991; Liu et al., 1997; Martin et al., 2000; MacDonald et al., 2010; Al-Jaberi et al., 2015). *Gad1* and *Gad2* expression is dependent on the DLX factors, where DLX1/2 bind directly to the promoters of *Gad1/2 in vivo* and induce *Gad1/2* expression; *Gad* expression is reduced in *Dlx1/2* double homozygous mutant mice (Stühmer et al., 2002; MacDonald et al., 2010; Li et al., 2012a; Le et al., 2017). However, *Gad* expression is not completely ablated in these mutants, which could be due to the compensatory function of residually expressed DLX5/6. Additionally, DLX proteins promote the differentiation of GABAergic and cholinergic interneuron subtypes through regulation of *Lhx6* and *Lhx8*, where both *Lhx* genes have reduced expression in *Dlx1/2* double homozygous mutants (Petryniak et al., 2007; Long et al., 2009). DLX2 also downregulates *Olig2* expression to repress oligodendrocyte development in early neurogenesis, and hence may control the balance between oligodendrocyte and neuron production (Petryniak et al., 2007; Jiang et al., 2020). ASCL1, in turn, represses *Dlx2* expression in later developmental stages, to allow the expression of *Olig2* and promote oligodendrocyte production (Petryniak et al., 2007; Poitras et al., 2007). The repression of *Olig2* by DLX2 also represses the promotion of the progenitor cell states, and likewise DLX2 downregulates several other transcription factors with similar functions, such as *Gsx2*, *Otx2*, and *Pax6* (Yun et al., 2001; Hoch et al., 2015; Lindtner et al., 2019). SMAD transcription factors, which are part of the transforming growth factor-β (TGF-β) signaling pathway, interact with DLX2 in binding to the promoter regions of DLX2 target genes in the telencephalon. Although expression of TGF-β signaling components is unaffected in the *Dlx1/2* double mutants, the interaction between DLX2 and SMAD factors indicate TGF-β could play a role in GABAergic interneuron differentiation (Shi and Massagué, 2003; Maira et al., 2010).

In addition to cell differentiation, DLX transcription factors also regulate interneuron tangential migration. DLX1/2 regulates this process by repressing terminal differentiation of interneurons (Cobos et al., 2007). Interneurons develop axons and dendrites post migration, promoted by proteins that regulate cytoskeleton and cell motility such as MAP2 and PAK3 (Anderson et al., 1997b; Bokoch, 2003; Dehmelt and Halpain, 2004). In *Dlx1/2* DKO mice, interneurons have significantly reduced migration, increased neurite length, and upregulated expression of genes which are normally expressed post-migration. Hence, DLX1/2 represses these genes to enable tangential migration of interneurons to the cortex (Cobos et al., 2007). *Nrp2*, encoding for a Semaphorin-3A and 3F receptor, is also repressed by DLX1/2, as evident in the marked increase of NRP2 expression in the forebrains of *Dlx1/2* DKO mice (Le et al., 2017). In *Dlx5/6* double homozygous mutant mice, a receptor for tangential migration *Cxcr4* is downregulated in the SVZ, which likely contributes to the impaired tangential migration observed in these mutants (Wang et al., 2010, 2011b).

Although DLX transcription factors have not been directly linked to any neurological diseases, many associations have been made between DLX mutations and neurodevelopmental defects (**Table 3**). Epilepsy and Rett syndrome had been linked to *Dlx* mutations in mouse models. Furthermore, *DLX1/2* and *DLX5/6* are found on chromosomes 2q and 7q, which are autism susceptibility loci (Cobos et al., 2005b; Hamilton et al., 2005; Horike et al., 2005). By site-directed mutagenesis of the *Dlx1/2* intergenic enhancer regions, transgenic mice with autism-like phenotypes were generated, showing the possible role of disrupted *Dlx1/2* in autism development (Poitras et al., 2007). Several neurodevelopmental disorders have been related

to *Dlx* genes due to the importance of this gene family in regulating GABAergic interneuron production and migration (Kato and Dobyns, 2004; Verret et al., 2012). A DLX2 direct target *Grin2b* is linked to schizophrenia, epilepsy, intellectual disability, and autism, which provides evidence that DLX2 may contribute to neural diseases (Endele et al., 2010; Pan et al., 2019). DLX2 regulation of transcription factors such as *Arx* and *Olig2* also support that DLX factors may potentially contribute to neurological disease (Lindtner et al., 2019).

Empty Spiracle Genes

Empty spiracles homeobox (*Emx)* genes are the mammalian homologues of the *Drosophila* gene *empty spiracle* (*ems*), which is responsible for head structure development (Younossi-Hartenstein et al., 1997). *Emx1* and *Emx2* are homeobox genes important for dorsal patterning in the forebrain. From mouse studies, *Emx2* is shown to be expressed earlier, from E8.5, whereas *Emx1* is expressed from E9.5 (Simeone et al., 1992; Gulisano et al., 1996; Medina and Abellán, 2009). The expression of *Emx2* is regulated by two sets of enhancers, one at the 5′ region and the other at the 3′ region (Theil et al., 2002; Suda et al., 2010; García-Moreno and Molnár, 2015). *Emx2* expression is directly promoted by DMRT5 and downregulated by the *Emx2* antisense transcript *Emx2OS* (Spigoni et al., 2010; Saulnier et al., 2013). Both *Emx* genes are expressed in the dorsal telencephalon, with the highest level of expression rostrolaterally, and decreased expression in a gradient caudomedially (Mallamaci et al., 1998). *Emx1* expression is nested within *Emx2* expression, and only *Emx2* is expressed in the caudomedial part of dorsal telencephalon (Simeone et al., 1992; Yoshida et al., 1997). While *Emx2* expression is restricted to progenitor cells, *Emx1* is expressed in both progenitor and differentiated cells (Gulisano et al., 1996).

Both *Emx1* and *Emx2* are necessary for the development of the archipallium in the dorsal telencephalon, and especially the development of the hippocampus and cortex in later stages (Simeone et al., 1993; Pellegrini et al., 1996; Yoshida et al., 1997; Hamasaki et al., 2004). *Emx1* and *Emx2* double homozygous mutants do not develop the dorsomedial telencephalon, whilst this phenotype is not observed in *Emx1* or *Emx2* single homozygous mutants (Bishop et al., 2003; Shinozaki et al., 2004). The impaired development of the neocortex could also be due to impaired tangential migration, as interneurons in *Emx1*/*Emx2* double mutant cannot migrate out of the GE into the cortex (Shinozaki et al., 2002). Homozygous *Emx1* mutants do not develop significant defects in the embryonic neocortex (Yoshida et al., 1997; Bishop et al., 2002). However, postnatal studies have shown that *Emx1* could play a role in cortical patterning, as rostral areas were expanded and caudal areas were reduced in the *Emx1* null mice (Stocker and O'Leary, 2016). *Emx1* homozygous mutants also lack development of the corpus callosum, and heterozygous *Emx1* mutants exhibit partial penetrance (Qiu et al., 1996). However, *Emx2* homozygous mutants have reduced neocortex size by E11.5, with defective dorsal telencephalon development, including aberrant hippocampus formation and impaired radial migration of neurons (Pellegrini et al., 1996; Yoshida et al., 1997; Mallamaci et al., 2000). In these mutants, there is ventralization of the dorsal telencephalon with reduced dorsal gene marker expression (*Ngn1*, *Ngn2*, and *Emx1*) and increased ventral marker gene expression (*Gsx2*, *Ascl1*, and *Dlx1/2*). *Emx2*/*Pax6* double homozygous mutants demonstrate a stronger phenotype with a lack of dorsal identity, showing these two homeobox factors function cooperatively to specify dorsal telencephalic identity (Muzio et al., 2002a,b). Furthermore, reciprocal inhibition is observed between *Emx2* and *Pax6*, where the cKO of one factor results in the upregulation of the other (Muzio et al., 2002b).

EMX1 and EMX2 regulate a number of factors required to specify dorsal telencephalic identity (**Table 2**). An important aspect of EMX2 function is its' regulation of the formation of the PSB, along with PAX6 and GSX2 (Yun et al., 2001; Muzio et al., 2002a,b). EMX2 cooperates with DMRT5 and DMRT3 to repress *Gsx2* expression, with all three proteins binding directly to the ventral telencephalon specific *Gsx2* enhancer, thereby contributing to the development of the PSB (Desmaris et al., 2018). A mutual repressive relationship between EMX2 and FGF8 also promotes the differentiation of neural progenitor identity, where EMX2 downregulates FGF8 to promote differentiation, whilst FGF8 represses EMX2 to promote anterior-posterior patterning (Fukuchi-Shimogori and Grove, 2001, 2003; Cholfin and Rubenstein, 2008). In addition, EMX2 represses *Sox2* by inhibiting positive regulators from binding to *Sox2* enhancers. *Sox2* cKO mutants have a defective hippocampal phenotype, rescued when one *Emx2* allele is lost (Mariani et al., 2012). This demonstrates that EMX2 regulates hippocampal development, consistent with the observed *Emx2* homozygous phenotype (Pellegrini et al., 1996). Wnt signaling promotes *Emx2* expression through activation of an *Emx2* telencephalic enhancer, through the Wnt downstream factor GLI3 (Theil et al., 1999, 2002; Muzio et al., 2005). Additionally, EMX2 restricts *Wnt-1* expression in the forebrain, which is essential for maintaining normal neuronal radial migration (Iler et al., 1995; Ligon et al., 2003). EMX1 regulates *Nrp1*, an axonal guidance receptor that regulates cortical connectivity (Wright et al., 2007; Lim et al., 2015). Furthermore, EMX2 regulates Teneurin-1, a transmembrane protein that also functions in axonal guidance, through binding to an alternative promoter (**Table 3**) and promoting the transcription of an alternative transcript (Drabikowski et al., 2005; Li et al., 2006; Beckmann et al., 2011).

While EMX2 functions in promoting cell differentiation in the developing brain, it is considered as a possible tumor suppressor in different cancers, such as sarcoma, colorectal cancer, gastric tumors, and glioblastoma (**Table 3**; Li et al., 2012b; Aykut et al., 2017; Jimenez-García et al., 2021a,b). In many tumors, *EMX2* expression is downregulated due to methylation of the *EMX2* promoter (Okamoto et al., 2010; Qiu et al., 2013). *EMX2* over-expression blocks cell proliferation through inhibiting the canonical Wnt pathway, and also leads to cell cycle arrest with increased cell death of glioblastoma cells (Falcone et al., 2016; Monnier et al., 2018; Jimenez-García et al., 2021a).

Genomic Screened Homeobox Genes

Genomic screened homeobox (*Gsx*, formerly *Gsh*) genes encode a family of transcription factors important for patterning of the

ventral telencephalon. *Gsx* genes are the mammalian orthologues of the *Drosophila intermediate neuroblasts defective* (*ind*) genes; mutation in *Drosophila* induces a loss of intermediate neuroblasts (Weiss et al., 1998). GSX proteins bind to DNA *via* the homeobox domain (**Figure 2**), and GSX2 activity may depend on its dimerization state, where homodimers promote gene activation, and monomers enhance gene repression (Salomone et al., 2021).

Gsx1 and *Gsx2* are widely expressed in the neural progenitors found in the VZ of the LGE (Toresson et al., 2000). *Gsx2* is mostly expressed in the dLGE with lower levels in the vLGE with complementary patterns for *Gsx1*, localizing to the vLGE and MGE (Toresson and Campbell, 2001; Yun et al., 2001). *Gsx1* and *Gsx2* are partially functionally redundant, due to similarities in their consensus DNA binding sites (Hsieh-Li et al., 1995; Valerius et al., 1995; Toresson and Campbell, 2001; Pei et al., 2011). *Gsx2* and *Gsx1/Gsx2* DKO mice have a reduced LGE size, with decreased number of olfactory bulb neurons as well as GABAergic interneurons (Yun et al., 2003). Dorsal markers including *Pax6* and *Ngn2* also expanded ventrally in *Gsx2* homozygous mutants (Szucsik et al., 1997; Corbin et al., 2000; Toresson et al., 2000; Yun et al., 2001). The expression patterns of *Pax6* and *Gsx2* are complementary, separated by the PSB, and these genes function cooperatively to define the dorsoventral identity of the developing forebrain (**Figure 1B**; Yun et al., 2001; Carney et al., 2009). *Gsx2* expression is repressed by a number of genes in the dorsal telencephalon, including *Pax6, Emx2, Dmrt3* and *Dmrt5* (Muzio et al., 2002a,b; Desmaris et al., 2018).

Gsx2 regulates specification of neurons, oligodendrocytes and glial cells in the LGE (Kessaris et al., 2006; Fogarty et al., 2007; Chapman et al., 2018). Neurogenesis and oligodendrogenesis take place in the dLGE and vLGE, respectively, and are tightly controlled by *Gsx2* in a time-dependent manner. Conditional knockout (cKO) of *Gsx2* upregulates the oligodendrocyte precursor cell (OPC) marker *Pdgfrα,* and promotes premature oligodendrocyte differentiation (Corbin et al., 2003; Chapman et al., 2013, 2018). *Ascl1*, a bHLH transcription factor crucial for neurogenesis, has reduced expression levels in *Gsx2* homozygous mutant mice (Chapman et al., 2013). Studies have shown that *Gsx2* upregulates *Ascl1* in earlier embryo stages, promoting neuronal differentiation in early embryonic stages (Méndez-Gómez and Vicario-Abejón, 2012; Chapman et al., 2013; Wang et al., 2013). Since *Ascl1* promotes the NPCs to differentiate into interneurons, GSX2 inhibits ASCL1 activity to regulate the balance between progenitor cell proliferation and differentiation (Roychoudhury et al., 2020). Whilst *GSX2* upregulates *Ascl1*, it inhibits the homo- and heterodimer formation of ASCL1 essential for its DNA binding ability (Johnson et al., 1992; Nakada et al., 2004; Roychoudhury et al., 2020). In earlier embryonic stages (E9-11), *Gsx2* promotes striatal projection neuron specification from the vLGE, and in later embryonic stages (E12.5–E15) olfactory bulb interneurons are specified in the dLGE (Waclaw et al., 2010).

Overexpression of *Gsx2* from E13.5 promotes the specification of dLGE over vLGE, and subsequently favors neurogenesis over oligodendrogenesis (Waclaw et al., 2010; Pei et al., 2011; Chapman et al., 2013). GSX1 in *Ascl1* expressing progenitor cells represses *Gsx2* and promotes the maturation of NPCs by transitioning these cells from the VZ to the SVZ and induces differentiation (Pei et al., 2011). *Gsx1/Gsx2* DKO mice have expanded OPCs comparable to *Gsx2* homozygous mutants (Chapman et al., 2018). However, the reduced proliferation of OPCs in the *Gsx1/2* DKO compared to *Gsx2* homozygous mutants suggests GSX1 functions in promoting OPC proliferation in the ventral telencephalon (Chapman et al., 2018). Hence, *Gsx2* regulates neurogenesis through repressing *Gsx1*, and blocks oligodendrogenesis in early embryonic stages. Furthermore, downregulation of *Gsx2* in late embryonic stages is essential for oligodendrogenesis to proceed, which could be a result of negative autoregulation (Salomone et al., 2021).

Along with promoting *Ascl1* expression, GSX2 regulates neural differentiation *via* increasing *Dlx1* and *Dlx2* expression in the LGE (**Table 2**; Corbin et al., 2000; Toresson et al., 2000; Wang et al., 2013). *Dlx1/Dlx2* are part of the gene regulatory network downstream of *Ascl1* and in turn negatively regulate *Gsx1* and *Gsx2* expression (Yun et al., 2002; Long et al., 2009; Wang et al., 2013). The activation of *Gsx1* and *Gsx2* regulates the patterning of LGE, and later silencing of these two genes by *Dlx1/2* promotes subcortical neural differentiation (Anderson et al., 1997b; Cobos et al., 2005b). Furthermore, GSX2 also represses *Dbx1*, a homeobox transcription factor expressed in the hindbrain and spinal cord that regulates dorsoventral brain patterning and specification of Cajal-Retzius cells (Yun et al., 2001; Bielle et al., 2005; Winterbottom et al., 2010). DBX1 has also been suggested to repress *Gsx1* in the ventral telencephalon; however, further studies are necessary to validate this relationship (Poiana et al., 2020).

Congenital brain malformations may result from mutations in the *GSX2* gene (**Table 3**). Whole exome sequencing of patients with basal ganglia malformations reveals a homozygous missense mutation in *GSX2* HD that impair its transcriptional activity (De Mori et al., 2019). These patients have similar phenotypes to homozygous mutant mice models, with malformations or defective structures derived from the LGE and MGE (putamen, globus pallidus, caudate nucleus and olfactory bulb), as well as maldevelopment of the forebrain midbrain junction (De Mori et al., 2019). These anatomical defects are also associated with a range of neurological disorders, such as Parkinson's and Huntington's Diseases (Zuccoli et al., 2015; **Table 2**).

Iroquois-Related Homeobox 3 Gene

The iroquois-related homeobox 3 (*Irx3*) is a TALE HD containing transcription factor (**Figure 2**), orthologous to the *Iroquois*-complex genes in *Drosophila*, which are responsible for the development of sensory organ, body-wall and wing identity (Gómez-Skarmeta et al., 1996; Bürglin, 1997; Diez del Corral et al., 1999). *Irx* genes in vertebrates are organized into two clusters, *IrxA* and *IrxB*, each containing 3 genes from the family. The *IrxA* cluster consists of *Irx1, Irx2,* and *Irx4*, whereas *IrxB* contains *Irx3, Irx5,* and *Irx6*, located on mouse chromosome 8 and human chromosome 16 (Peters et al., 2000).

Iroquois-related homeobox 3 is important for thalamic patterning in the diencephalon (Robertshaw et al., 2013). *Irx3* is predominantly expressed in the midbrain, hindbrain, and spinal cord in early neurogenesis (E7.5–E9.5), and expression

shifts rostrally to the diencephalon from E10.5 (Bosse et al., 1997). Notably, the expression patterns of *Irx3* and *Ascl1* during early neurogenesis are similar, which may suggest a regulatory relationship between the two transcription factors (Cohen et al., 2000). Similar to *Dlx1/2/5* and *Nkx2.1/2.2*, *Irx3* expression is posterior to the *zona limitans intrathalamica* (ZLI), a region in the diencephalon that releases SHH signaling molecules for the patterning of prethalamus and thalamus (Kitamura et al., 1997; Eisenstat et al., 1999; Robertshaw et al., 2013; Murcia-Ramón et al., 2020a). High levels of SHH signaling induces rostral thalamus, and subsequently the production of GABAergic interneurons, while a low level of SHH promotes caudal thalamus specification and glutamatergic interneurons production (Kiecker and Lumsden, 2005). Consistent with this, ectopic expression of *Irx3* promotes the expression of thalamus differentiation markers *Sox14* and Gbx2, both in the prethalamus and the dorsal telencephalon in response to SHH signaling (Kiecker and Lumsden, 2004; Robertshaw et al., 2013). However, such markers were not expressed upon *Irx3* ectopic expression in the ventral telencephalon, which may be due to SIX3 repression of *Irx3*, which restricts its activity to specify thalamus identity (Kobayashi et al., 2002; Robertshaw et al., 2013). In *Xenopus* models, knockdown of *Irx3* reduces midbrain size, and caudally shifts the forebrain-midbrain boundary, illustrating its function in ensuring the normal patterning of the diencephalon (Rodríguez-Seguel et al., 2009). A key co-regulator of thalamus patterning is PAX6, which is expressed anterior to the forebrain-midbrain boundary and specifies the caudal thalamus. The overlapping expression patterns of *Irx3* and Pax6 (see **Figure 2A**) mark the region of thalamus patterning, while caudal and rostral thalamus identity is determined by levels of SHH signaling (Robertshaw et al., 2013).

Iroquois-related homeobox 3 is considered to be a determinant for obesity, in relation to the fat mass and obesity associated (FTO) genes, due to the role of *Irx3* in neurogenesis at the paraventricular nucleus of the hypothalamus, developed from the anterior hypothalamus (Smemo et al., 2014). *Single-minded 1* (*Sim1*), a bHLH transcription factor in the hypothalamus represses *Irx3* expression, as *Sim1* KO mice exhibit ectopic expression of *Irx3* in the anterior hypothalamus (Caqueret et al., 2006; Son et al., 2021b). *Sim1* homozygous mutant mice are perinatal lethal, whereas *Sim1* heterozygous mutant mice exhibit neurodevelopmental defects and hyperphagia, as *Sim1* is important for neurogenesis in the hypothalamus (Michaud et al., 1998; Holder et al., 2004). The neurogenesis defects in these mutant mice are due to the ectopic expression of *Irx3* and *Irx5* in the anterior hypothalamus (Son et al., 2021a,b). In *Sim1/Irx3/Irx5* triple heterozygous KO mice, the neuronal population at the anterior hypothalamus is restored. Similarly, cKO of *Irx3* at the paraventricular nucleus of the hypothalamus partially rescues the neuronal disruption observed in *Sim1* heterozygous mutant mice, with no observable differences in body weight or hyperphagic phenotype (Son et al., 2021b).

Lhx (LIM-HD Family) Genes

The *Lhx* transcription factors belong to the LIM-HD family of homeobox genes that have both a LIM zinc finger domains and a HD (**Figure 2**; Dawid et al., 1998; Bach, 2000). The LIM zinc finger domain is named after the first three genes discovered in the family, *Lin-11, Isl1* and *Mec-3*, and participates in protein-protein binding (Way and Chalfie, 1988; Freyd et al., 1990; Karlsson et al., 1990). Of the various members of the *Lhx* gene family found in both mouse and humans, *Lhx1, Lhx2, Lhx5, Lhx6,* and L*hx8 (i.e., L3/Lhx7)* are important for differentiation and migration of interneuron in the developing telencephalon (Alifragis et al., 2004; Abellán et al., 2010; Godbole et al., 2018). Mutant mice studies had provided insights into the importance of these *Lhx* genes for forebrain development (Wanaka et al., 1997).

Lhx1 homozygous mutant mice have an increased number of PoA-derived interneurons and glia cells, suggesting *Lhx1* regulates the survival of these cells by regulating the balance between apoptosis and proliferation. Also the PoA-derived interneurons in *Lhx1* null mice migrate through the ventral telencephalon, compared to a more controlled migration in the wild-type mice through the developing neocortex (Symmank et al., 2019). *Lim5* is expressed in the forebrain of zebrafish and *Xenopus*, and *Lhx5,* the *Lhx1* paralog, is the murine ortholog. *Lhx5* is expressed predominantly in the hindbrain at E8, and the developing forebrain starting at E9.5. After E11.5, *Lhx5* is exclusively expressed in the ventral telencephalon, hypothalamus and diencephalon, which is complementary to *Dlx5* expression (**Figure 1B**; Sheng et al., 1997). Both *Lhx1/5* are expressed in the rostral area of the ZLI in the diencephalon, but only *Lhx1* is expressed in the caudal ZLI (Nakagawa and Leary, 2001). *Lhx5* homozygous mutant mice are defective in hippocampus development, where progenitor cells can proliferate but fail to exit the cell cycle to migrate or differentiate (Zhao et al., 1999). Cajal-Retzius neurons are responsible for the organization of the neocortex through the secretion of reelin (Soriano and Del Río, 2005). In mice, *Lhx5* regulates the development and migration of Cajal-Retzius cells, which could be critical to the malformation of the hippocampus in *Lhx5* null mutants (Abellán et al., 2010). *Lhx1* likewise is expressed in some Cajal-Retzius cells, but limited to the septal area, and lateral olfactory to caudomedial zones (Miquelajáuregui et al., 2010). Additionally, *Lhx5* can regulate forebrain development by suppressing Wnt signaling in zebrafish embryos, *via* promoting the expression of Wnt inhibitors *Sfrp1a* and *Sfrp5*, supported by the increase of Wnt signaling in zebrafish embryos lacking *Lhx5* expression (Peng and Westerfield, 2006). There is some evidence of *Lhx5* inhibiting Wnt5a in murine hypothalamus, promoting the growth of the mamillary body; however, more studies are required to confirm this regulatory effect and mechanism. Another possible target of *Lhx5* is *Lmo1* (LIM-only1), where *Lmo1* competes with *Lhx5* to bind with the *Lhx* binding partner LDB, thereby inhibiting *Lhx* function (Bach, 2000; Heide et al., 2015).

Lhx2 is the mammalian ortholog of the *Drosophila apterous* gene, first described in 1913, as an essential gene for *Drosophila* wing development (Metz, 1914; Butterworth and King, 1965). *Lhx2* homozygous mutants have reduced forebrain volume, but expanded neocortex and PSB composing the entire forebrain (Porter et al., 1997; Bulchand et al., 2001; Monuki et al., 2001). *Lhx2* plays a role in suppressing hippocampus (hem) and PSB (antihem) development up to E9.5 and E10.5,

respectively (Roy et al., 2014; Godbole et al., 2018). Suppression of hippocampal development is regulated by interactions between *Lhx2* and other transcription factors, namely *Foxg1* and *Pax6*. *Foxg1* has been shown to directly regulate *Lhx2* expression, where the loss of *Foxg1* also results in a loss of *Lhx2* at E9.5. cKO of *Lhx2* after E9.5 did not alter hippocampus development unless *Foxg1* was also knocked out (Godbole et al., 2018). *Pax6* is expressed in a lateral medial gradient in the neocortex, which is opposite to that of *Lhx2*. In *Pax6*/*Lhx2* DKO, the hippocampus expands more so in the forebrain compared to *Lhx2* null mice, suggesting *Pax6* also suppresses the formation of hippocampus (Godbole et al., 2017).

Lhx6 and *Lhx8* are structurally related and have synergistic functions. *Lhx6* shares 75% homology with *Lhx8*, which is also known as *L3* or *Lhx7* (Matsumoto et al., 1996; Grigoriou et al., 1998). Both these genes are expressed overlappingly in the MGE but are not expressed in the LGE (**Figure 1B**). *Lhx6* is expressed predominantly in the SVZ and the MZ, whilst *Lhx8* is expressed in the MZ (Matsumoto et al., 1996). The expression of both these genes is regulated by *Nkx2.1*, another homeobox transcription factor that specifies ventral telencephalon development (Sandberg et al., 2016).

Lhx6 has similar functions to *Lhx1*. *Lhx6* promotes expression of receptors that regulate cortical interneuron migration and transcription factors that control interneuron production, thereby regulating these events (Alifragis et al., 2004; Zhao et al., 2008; Neves et al., 2013). Tangential migration of GABAergic interneurons from the MGE into the neocortex are blocked in embryonic mice lacking *Lhx6;* normally these interneurons express *Lhx6* in wildtype mice (Lavdas et al., 1999; Alifragis et al., 2004; Liodis et al., 2007). Such migration defects prevent the formation of functional connections between these neurons and their post-synaptic targets. Since *Lhx6* has restricted expression in MGE progenitor cells, it does not regulate the migration of all cortical interneurons during development, especially at later stages where tangentially migrating neurons are born in the LGE (Marin et al., 2000; Nery et al., 2002). Production of GABAergic interneurons and their migration within the MGE are not affected in *Lhx6* mutants, but interneuron subtype specification is dependent on the expression of *Lhx6* (Neves et al., 2013). MGE-derived cortical interneurons are unable to differentiate into sst^+ and pva^+ subtypes, shown by a drastic reduction in the number of these neurons in *Lhx6* null mutants. *Lhx6* KOs had a greater effect on sst^+ interneuron differentiation than pva^+ interneuron differentiation, where pva^+ interneuron differentiation was affected restrictively in the hippocampus (Liodis et al., 2007; Zhao et al., 2008; Yuan et al., 2018).

Lhx8, unlike *Lhx6,* is expressed in cholinergic neurons instead of GABAergic neurons (Lopes et al., 2012). *Lhx8* is essential for the differentiation and specification of cholinergic interneurons, shown by the reduction of cholinergic neurons in *Lhx8* homozygous mutant mice (Zhao et al., 2003; Fragkouli et al., 2005). Progenitor cells proliferate in *Lhx8* homozygous mutant mice; however, they are unable to differentiate into cholinergic interneurons or glutamatergic neurons (Manabe et al., 2007, 2008). Cholinergic neurons are derived from progenitor cells in the MGE, where LHX8 promotes the expression of *Isl1* upon cholinergic commitment, which in turn represses *Lhx6* expression (Zhao et al., 2003). *Lhx8* forms a hexamer with *Isl1* and promotes cholinergic neuron expression by binding to specific motifs in the cholinergic enhancer sequence (Park et al., 2012). The formation of hexamers is necessary for DNA binding and subsequently cholinergic gene expression, whilst LHX8 or ISL1 alone does not bind to cholinergic enhancer sequences and are unable to promote cholinergic interneuron differentiation (Cho et al., 2014). NPCs in the striatum differentiate into GABAergic interneurons instead of cholinergic neurons in *Lhx8* homozygous mutants. This is due to an upregulation of *Lhx6* as a result of a lack of *Isl1*, suggesting the necessity of *Lhx8* in cholinergic neuron specification (Manabe et al., 2005; Bachy and Rétaux, 2006). Additionally, *Lhx6* acts cooperatively with *Lhx8* to promote *shh* expression in the MGE, regulating the production of interneuron progenitors, as well as inhibiting *Nkx2.1* expression in cortical neurons (Flandin et al., 2011). The *Lhx6* and *Lhx8*/*Isl1* regulatory network is therefore essential for regulating the differentiation of GABAergic and cholinergic neurons in the ventral telencephalon.

The LIM-domain transcription factor family is functionally important for the specification, differentiation and migration of neurons in the developing forebrain, and mutations in these genes can result in genetic diseases (**Table 3**). *LHX2* mutations can result in pituitary hormone deficiency, although it is uncommon that a mutation in *LHX2* alone can cause pituitary deficiency and developmental ocular abnormalities (Prez et al., 2012). The importance of *Lhx6* on the differentiation of interneurons into sst^+ and pva^+ subtypes have a pathological link to schizophrenia (Volk et al., 2014; Donegan et al., 2020). There is reduced *LHX6* expression in schizophrenic subjects who also have reduced expression of *GAD1* (otherwise known as GAD67, a GABA synthesizing enzyme), *sst,* and *pva* expression. Reduction in *GAD1* does not downregulate *LHX6* and *vice versa*; hence, upstream factors likely contribute to the regulation of these genes (Volk et al., 2012). Moreover, a decrease in both GABAergic and cholinergic interneurons in the ventral telencephalon has been reported in Tourette Syndrome, suggesting *LHX6* and *LHX8* correlation with Tourette Syndrome due to their role in GABAergic and cholinergic interneuron specification in the striatum (Pagliaroli et al., 2020).

Myeloid Ectopic Viral Integration Site 2 Gene

The *myeloid ectopic viral integration site* (*Meis)* gene family belongs to the TALE class of homeobox proteins, a homolog of the *Drosophila homothorax* gene, which is essential for directing the localization of *Pbx Drosophila* homologue *extradenticle* (Rieckhof et al., 1997). There are three mammalian MEIS transcription factors (*Meis1, Meis2, and Meis3*), and all contain a conserved homothorax domain (**Figure 2A**), which promotes the interaction between MEIS and pre-B cell leukemia homeobox proteins (PBX), a transcription factor known for its regulatory role in organogenesis (Nakamura et al., 1996; Chang et al., 1997; Golonzhka et al., 2015). MEIS proteins are characterized by a three residue loop insertion between helices 1 and 2 of the HD, an important feature for protein-protein interactions (Bürglin, 1997). Out of the three *Meis* genes, only *Meis1* and

Mei2 are expressed in the developing telencephalon (**Figure 1B**). *Meis2* in particular is an important player for striatal progenitors and neuron differentiation, as well as postnatal neuronal differentiation in the olfactory bulb (Toresson et al., 1999; Agoston et al., 2014).

Myeloid ectopic viral integration site 2 is expressed in the VZ of the entire telencephalon from E10.5, and is enriched in the LGE compared to the MGE from E12.5 to E18.5 (**Figure 1B**). From E14.5, MEIS2 is also expressed in the ventral thalamus and the anterior hypothalamus (Cecconi et al., 1997; Toresson et al., 1999, 2000). Additionally, the expression pattern of MEIS2 is similar in the human fetal forebrain, where MEIS2 is expressed in the proliferative zones (Larsen et al., 2010a). In the telencephalon, MEIS2 was initially considered as an LGE-specific marker due to its predominant expression in the LGE; however, MEIS2 is also widely expressed in the CGE progenitors (Toresson et al., 1999; Frazer et al., 2017). Postnatally, interneurons born and derived from the olfactory bulb express MEIS2, as it plays a crucial role, along with other transcription factors, in neuronal differentiation and specification in early postnatal stages (Allen et al., 2007; Agoston et al., 2014).

Myeloid ectopic viral integration site 2 forms complexes with various other transcription factors to cooperatively facilitate the expression of genes required for neurogenesis. As mentioned, MEIS2 interacts with PBX1 proteins and forms heteromeric complexes, which regulate the DNA binding ability of the two transcription factors (Liu et al., 2001; Longobardi et al., 2014). The MEIS2-PBX1 complex further recruits other transcription factors, such as the Kruppel-like factor 4 (*Klf4*) to modulate MEIS2 transcriptional activities (Bjerke et al., 2011). Other than PBX1, MEIS2 also functions synergistically with HOX and PAX homeobox factors, regulating the gene expression of other targets in the midbrain and hindbrain (Agoston et al., 2012). Mechanisms for the interactions between MEIS2 and other factors have been extensively reviewed; notably, MEIS2 recognizes and binds to a specific DNA motif TGACAG (**Table 2**; Chang et al., 1997; Longobardi et al., 2014; Schulte, 2014).

Myeloid ectopic viral integration site 2 controls gene expression and promotes neuronal migration and differentiation during forebrain development. There are three types of serotonin receptor 3a expressing (*Htr3a*+) GABAergic interneurons, which populate different regions of the brain. Type I *Htr3a*+ are enriched in transcription factors expressed in the LGE, including MEIS2, and these interneurons populate the deep cortical layers (von Engelhardt et al., 2011; Frazer et al., 2017). These interneurons originate from the PSB and migrate through to the cortex, contrasting with other types of *Htr3a*+ interneurons which are born from the CGE and populate the superficial cortical layers. Ectopic expression of *Meis2* in CGE born interneurons resulted in a shift of differentiated *Htr3a*+ interneurons to the deep cortical layers, indicating that MEIS2 induces the migration of the LGE-derived interneurons (Frazer et al., 2017). Alternatively, MEIS2 can regulate expression of the *Dlx* family, by interacting with the intergenic enhancers in the *Dlx* bigenic clusters (Ghanem et al., 2003). MEIS2 binds to the I12b intergenic enhancer of *Dlx1/2* and the I56ii intergenic enhancer of *Dlx5/6*. MEIS2 can activate reporter gene transcription with a I56ii promoter sequence *in vitro* (Yang et al., 2000; Poitras et al., 2007; Ghanem et al., 2008). Subsequently, the removal of I56ii sequence reduced *Meis2 and Dlx5/6* expression, suggesting that there may be a positive feedback loop between MEIS2 and DLX5/6, further regulating interneuron migration (Fazel Darbandi et al., 2016). Furthermore, dopamine receptor expressing (D1/D2) MSNs are promoted by MEIS2 in the LGE, where deletion of *Meis2* blocked differentiation of neural progenitors and reduced the medium-spiny neuron population (Su et al., 2022). MEIS2 regulates specification of these striatal projection neurons through the promotion of *Zfp503* and *Six3* expression, while *Meis2* expression itself is regulated by DLX1/2 (Su et al., 2022). Likewise, in the prethalamus, DLX2 drives GABAergic interneuron determination through promoting *Meis2* expression, and SOX14 represses *Meis2* expression to maintain rostral thalamus identity (Sellers et al., 2014). In postnatal stages, interneurons continue to arise from the olfactory bulb SVZ generated neuroblasts where these differentiation events are dependent on the activity of MEIS2 and its interaction with PAX6 and DLX2 (Ming and Song, 2011; Agoston et al., 2014). Indeed, cKO of *Meis2* in the olfactory bulb blocks dopaminergic neuron differentiation, as MEIS2 promotes expression of *Dcx* and *Th*, both crucial genes for dopaminergic neuron subtype specification (Agoston et al., 2014; Kim et al., 2020).

Nkx2.1/2.2 Genes

Nkx2.1 and *Nkx2.2*, homeobox transcription factors of the vertebrate *Nkx* family, are important for the regulation of embryonic telencephalon and diencephalon patterning (Price et al., 1992; Sussel et al., 1999). *Nkx2.1* is the mammalian homolog of the *Drosophila scarecrow* (*scro*), and is also known as the thyroid transcription factor 1 and thyroid specific enhancer binding protein, since it also plays a role in thyroid, lung and pituitary development (Guazzi et al., 1990; Mizuno et al., 1991; Kimura et al., 1996; Maurel-Zaffran and Treisman, 2000). *Nkx2.2* is homologous to the *Drosophila ventral nervous system defective* (*vnd*) gene (Kim and Nirenberg, 1989; Price et al., 1992; Jimenez et al., 1995). *Nkx2.1* and *Nkx2.2* encode both a HD and a NK2 box domain (**Figure 2**). In embryonic forebrain, *Nkx2.1* is expressed in progenitor and post-mitotic cells in the MGE and PoA, and is essential for the patterning of these areas (Xu et al., 2005). *Nkx2.2* is localized to the MGE, the VZ of the thalamus and MZ of the diencephalon; however, *Nkx2.2* expression can vary in different mammalian species (Ericson et al., 1997; Flames et al., 2007; Vue et al., 2007; Bardet et al., 2010; Domínguez et al., 2015). The dorsoventral expression pattern of *Nkx2.1* (ventral) is complementary to that of *Pax6* (dorsal) (**Figure 1B**). Within the thalamus, *Nkx2.2* expression is induced by SHH signaling in the rostral thalamus along with *Ascl1*, resulting in the specification of GABAergic neurons that populate the thalamus; as a result, *Nkx2.2* is often co-expressed with SHH (Briscoe et al., 1999; Vue et al., 2007; Robertshaw et al., 2013).

Nkx2.1 expressing progenitor cells give rise to GABAergic and cholinergic neurons, which populate the neocortex and striatum, respectively (Anderson et al., 2001; Magno et al., 2017). *Nkx2.1* expression in the GABAergic interneurons then diminishes after

they tangentially migrate toward the neocortex, but is sustained in the cholinergic neurons (Marin et al., 2000). In the MGE, *Nkx2.1* silencing is necessary for interneurons to tangentially migrate. *Nkx2.1* silencing promotes the expression of *Nrp1* and *Nrp2*, which then initiates neural migration (Nóbrega-Pereira et al., 2008; Kanatani et al., 2015). *Nkx2.1* expressing neurons in the hypothalamus tangentially migrate into the diencephalon, and develop into GABAergic interneurons (Murcia-Ramón et al., 2020b). Additionally, NKX2.1 regulates astrocyte differentiation in the MGE and PoA from E14.5 to E16.5 in mice, and oligodendrocyte differentiation from E12.5 (Kessaris et al., 2006; Minocha et al., 2015, 2017; Orduz et al., 2019). Transcriptional activity is dependent on epigenetic states (Attanasio et al., 2014; Sandberg et al., 2016).

Nkx2.1 homozygous mutant mice die at birth with lung, thyroid, pituitary and ventral telencephalon defects (Kimura et al., 1996; Takuma et al., 1998; Sussel et al., 1999). In these mutant mice, the MGE is respecified into LGE, and exhibits reduced numbers of GABAergic and cholinergic neurons (Sussel et al., 1999; Fragkouli et al., 2009). However, ∼50% of GABAergic interneurons remain, suggesting that NKX2.1 is not the sole factor required for GABAergic interneuron specification (Sussel et al., 1999). cKO of *Nkx2.1* at E10.5 and E12.5 results in altered identity of the MGE-derived interneurons subtypes. The MGE progenitor cells of these mutants were respecified into calretinin and vasointestinal peptide (VIP) expressing interneuron subtypes, resembling interneuron populations derived from the caudal GE (Xu et al., 2004; Butt et al., 2005), as opposed to pva^+ or sst^+ subtypes (Butt et al., 2008). GABAergic interneuron differentiation, especially pva^+ and sst^+ subtypes, is tightly regulated by *Lhx6* and *Lhx8* in the MGE, and both genes are downstream targets of NKX2.1 (Du et al., 2008; Flandin et al., 2011; Sandberg et al., 2016; Kim et al., 2021). *Lhx6* and *Lhx8* are activated by NKX2.1 expression in the SVZ through the recognition of epigenetic markers, and are essential for the specification of pva^+ and sst^+ interneuron subtypes (Du et al., 2008; Kim et al., 2021). Furthermore, NKX2.1 regulates MGE identity through repression of genes in the SHH, Wnt, and BMP signaling pathways required for cell differentiation and patterning. This repression is likely achieved by recruitment of Gro/TLE, a complex that reduces epigenetic-mediated repression, and induces activation (Patel et al., 2012; Sandberg et al., 2016). Conversely, SHH can induce the expression of *Nkx2.1* in the MGE to specify ventral identity (Ericson et al., 1995). To establish ventral identity in the telencephalon, NKX2.1 also represses *Pax6* expression in the GE, as the *Nkx2.1* cKO showed a dorsal to ventral expansion and ectopic expression of *Pax6* ventrally (Manoli and Driever, 2014). *Pax6*, a dorsal telencephalon specifying gene, in turn represses the expression of *Nkx2.1* in the neocortex. The existence of this mechanism of mutual repression is supported by the complementary expression patterns of these two transcription factors (Sussel et al., 1999; Stoykova et al., 2000).

As *NKX2.1* is essential for formation of various organs, mutations in this gene are linked to multiple phenotypes and diseases, including neurological disease, lung defects and thyroid dysfunction (**Table 3**; Thorwarth et al., 2014). *NKX2.1* may play a role in Hirschsprung disease, a disorder of the developing enteric nervous system, through its interaction with *SOX10* and *PAX3*. Sex-determining factor SRY is reported to displace SOX10's interaction with NKX2.1 and PAX3, thereby promoting a Hirschsprung disease phenotype (Li et al., 2015). Furthermore, hereditary chorea, also known as brain-lung-thyroid disease, is linked to mutations in *NKX2.1* with symptoms such as impaired coordination or speech development (Krude et al., 2002; Monti et al., 2015). Subsequently, *NKX2.1* has also been related to the development of schizophrenia, as *Nkx2.1* regulates GABAergic and cholinergic neuron specification (Sussel et al., 1999; Fragkouli et al., 2009; Malt et al., 2016). The cholinergic specification function of *Nkx2.1* correlates with learning and memory, where the absence of *Nkx2.1* in the septal area results in cognitive impairments (Magno et al., 2017).

Orthodenticle Homeobox Genes

Orthodenticle homeobox (*Otx*) is an ortholog of the *Drosophila orthodenticle* transcription factor, with *OTX1* located in the human chromosome region 2p13 and *OTX2* located in the human chromosome region 14q21-22 (Kastury et al., 1994). *Crx* is another member of the *Otx* family, but its expression is restricted to the retina, and all three *Otx* genes share a common OTX tail domain at the C-terminal (**Figure 2A**; Furukawa et al., 1997). *Otx1* plays an important role in cortical neurogenesis, and along with *Otx2*, both genes are important for forebrain patterning and specification, as well as retinal development (Larsen et al., 2010b). *Otx2* is essential during gastrulation for forebrain specification, and *Otx2* expression continues in both dorsal and ventral telencephalon, diencephalon, and mesencephalon (Acampora et al., 1995; Rhinn et al., 1998; Tian et al., 2002; Kurokawa et al., 2004; Sakurai et al., 2010). The midbrain/hindbrain boundary marks the caudal limit of *Otx2* expression. *Otx2* expression is repressed in the hindbrain and spinal cord (Frantz et al., 1994). This pattern is regulated by fibroblast-growth-factor (*Fgf*)-8 and *Gbx2*, another homeobox gene that is required for caudal brain patterning and formation (Garda et al., 2001). GBX2 recognizes a conserved enhancer sequence in *Otx2*, thereby downregulating *Otx2* in the hindbrain (Kurokawa et al., 2006; Inoue et al., 2012). *Otx1* and *Otx2* exhibit a similar expression pattern early in embryogenesis, and *Otx1* expression is nested within the *Otx2* expressing regions (Simeone et al., 1993). From E8.5, *Otx2* expression starts to diminish in the rostral forebrain. At E11.5, *Otx2* is expressed in the VZ in the GE, and promotes the ventral identity of the MGE. *Otx1* expression patterns change to become complementary to *Otx2*; it is predominantly expressed in the VZ of the dorsal telencephalon and is expressed at lower levels in the dLGE (Hoch et al., 2015; Huang et al., 2018).

Mice lacking *Otx1* survive to birth but develop spontaneous epilepsy and seizures (Acampora et al., 1996). cKO of *Otx1* in the developing neocortex reduces the size of the neocortex as well as the overall cellular population (Pantò et al., 2004). Deletion of *Otx1* reduced the generation of neurons by repressing neural differentiation from cortical NPCs, while NPC proliferation was promoted, subsequently increasing the population of neurons (**Table 3**). This suggests *Otx1* promotes cell cycle exit in cortical

NPCs, thereby maintaining the balance between differentiation and proliferation (Huang et al., 2018).

Deletion of both *Otx1* and *Otx2* in mice results in a gastrulation defect and is embryonic lethal (Acampora et al., 1995). Heterozygous double mutants exhibit a range of phenotypes, including different degrees of craniofacial malformations, ocular defects, abnormalities in central nervous system, pituitary glands dysfunction, and developmental delay (Matsuo et al., 1995; Ang et al., 1996; Ragge et al., 2005; Tajima et al., 2009; Dateki et al., 2010; Mortensen et al., 2015). cKO of *Otx2* at different embryonic developmental timepoints and locations has shown a range of phenotypes indicating the essential role for *Otx2* in processes such as septum formation, specification of the neocortex, neurogenesis and early oligodendrogenesis and NPC fate (Acampora et al., 1997; Puelles and Rubenstein, 2003; Puelles et al., 2006; Silbereis et al., 2014; Hoch et al., 2015). The disruption of septum formation and cortex specification following cKO of *Otx2* after gastrulation suggests that *Otx2* could be regulating specification through FGF signaling (Acampora et al., 1997; Puelles and Rubenstein, 2003; Hoch et al., 2015). MGE interneuron markers such as *Dlx1*, *Arx*, and *Gbx* were downregulated in MGE-deleted *Otx2*, as well as the expression of oligodendrogenesis promoting genes *Olig1* and *Olig2* demonstrating a requirement for *Otx2* in neurogenesis and oligodendrogenesis (Silbereis et al., 2014; Hoch et al., 2015). Furthermore, *Lhx6* and *Lhx8* expression were reduced, suggesting *Otx2* plays a role in regulating cholinergic neurons. *Otx2* deletion in the thalamus resulted in a switch in NPC fate from glutamatergic neurons to GABAergic interneurons, demonstrating a requirement for *Otx2* in glutamatergic neuron specification (Puelles et al., 2006).

OTX1 and *OTX2* are overexpressed in medulloblastoma (Boon et al., 2005; Zakrzewska et al., 2013), which is a malignant pediatric brain tumor located in the posterior fossa that is divided into four molecular groups based on genomic and transcriptomic alterations: Wnt, SHH, Group 3, and Group 4 (Rudin et al., 2009; Northcott et al., 2011; Taylor et al., 2012). *OTX2* is overexpressed in over 60% of medulloblastoma, usually in Groups 3 and 4 (Boon et al., 2005; Bunt et al., 2010). It has been postulated that the cellular context dependent nature of *OTX2* expression could attribute to its overexpression in some groups of medulloblastoma (Kaur et al., 2015). As *MYC*, another oncogene is also overexpressed in Group 3 medulloblastoma, OTX2 may promote tumorigenesis by cooperatively binding with MYC to target genes (Bunt et al., 2011). Furthermore, OTX2 promotes the proliferation of tumors in Groups 3 and 4 (Lu et al., 2017; Zagozewski et al., 2020). The overexpression pattern observed may be a result of autoregulation. Chromatin accessibility is altered in medulloblastoma, where histone modifications may allow for increased *OTX2* expression, and hence a positive feedback loop for *OTX2* (Wortham et al., 2014). Recent studies have also shown *OTX2* is potentially required for tumor proliferation in the SHH group, although not necessarily for tumor formation (El Nagar et al., 2018). In Group 3 medulloblastoma, OTX2 represses *PAX3* and PAX6. Overexpression of *PAX3* and *PAX6* is associated with increased patient survival (Zagozewski et al., 2020). Additionally, OTX1 and OTX2 have been shown to act as oncogenes, promoting tumorigenesis and proliferation for cancers such as hepatocellular carcinoma, breast cancer, and Hodgkin or non-Hodgkin lymphomas (Omodei et al., 2009; Terrinoni et al., 2011; Nagel et al., 2015; Li et al., 2016; Tu et al., 2020).

Paired Box 6 Gene

The highly conserved *Pax6* transcription factor was first identified as a member of the *Paired box* (*Pax*) gene family based on its homology to the *Drosophila* gene *eyeless*. There are nine PAX transcription factors identified in mammals; all contain a paired domain and can be further categorized according to the presence or absence of additional domains, usually a HD. *Pax6* contains two DNA binding domains, a paired domain and a HD, as well as a proline-serine/threonine rich domain in the carboxyl-terminal (**Figure 2A**; Glaser et al., 1992; Duan et al., 2013). Hence, PAX6 binds to paired-HD and HD consensus DNA binding motifs (Sun et al., 2015). The *Pax6* homologue *eyeless* was first described in *Drosophila* as a gene essential for segmentation and eye development (Walther et al., 1991; Gehring, 1996), and in mammals it is important for the development of the CNS, eyes, pancreas, and pituitary gland (Dohrmann et al., 2000; Jones et al., 2002). In mice, *Pax6* expression begins from E8, and is then expressed in the forebrain, hindbrain, and spinal cord by E10 (Stoykova and Gruss, 1994; Inoue et al., 2000). Mice with homozygous mutation of *Pax6* die upon birth with malformation in the cerebral cortex (Tyas et al., 2003), whereas heterozygous mutation results in development of a thinner cortex, and have small or reduced eyes (Hill et al., 1991; Schmahl et al., 1993; Fukuda et al., 2000; Haubst et al., 2004; Quinn et al., 2007). Mice with cortex-specific KO of *Pax6* have reduced cortical size and an increased volume of the caudal cortex but without affecting thalamocortical identity (Piñon et al., 2008).

Within the telencephalon, *Pax6* is expressed in the VZ of the dorsal telencephalon, as well as the PSB, with a rostral-caudal gradient (**Figure 1B**; Bishop et al., 2000; Hirata et al., 2002). *Pax6* expression is repressed in the ventral telencephalon by OLIG2, which ensures the ventral identity of the forebrain (Lim et al., 2019). PAX6 promotes the expression of *Ngn2* in the dorsal telencephalon, together specifying dorsal identity (Scardigli et al., 2003). This pattern of expression is largely complimentary to that of genes specifying ventral identity such as *Ascl1, Dlx1/2*, and *Gsx2*. The only overlapping areas in which these genes are co-expressed are the PSB and the VZ of the LGE (Puelles et al., 1999; Cocas et al., 2011). *Pax6* is necessary for specification of the PSB, maintaining the physical boundary as well as a genetic boundary separating the dorsal and ventral telencephalon (Stenman et al., 2003). *Pax6* homozygous mutants fail to develop such a boundary, with upregulation of ventral genes such as *Gsx2*, and downregulation of dorsal genes like *Ngn2* in the dorsal telencephalon (Stoykova et al., 1996, 2000; Toresson et al., 2000; Yun et al., 2001; Quinn et al., 2007). As a result, the ventral telencephalon, in particular the dLGE, is expanded into the dorsal telencephalon, crossing over the PSB. During the course of development, *Gsx2* expressing progenitor cells in the dorsal LGE can change fate by expressing *Pax6*, distinguishing either *Pax6*-expressing (dorsal) and *Gsx2*-expressing (ventral)

progenitor cells at the PSB (Cocas et al., 2011). *Pax6* may regulate the formation of this boundary *via* regulation of cell adhesion molecules (Tyas et al., 2003). Progenitor cells in dorsal and ventral telencephalon expresses R-cadherin and cadherin-6, respectively (Matsunami and Takeichi, 1995; Inoue et al., 1997). Absence of PAX6 in the dorsal telencephalon reduces the expression of R-cadherin, allowing the dorsal and ventral cells to aggregate more readily and consequently disrupts the PSB (Stoykova et al., 1997).

Thalamic patterning is also partly regulated by PAX6 within the dorsal thalamus, where *Pax6* homozygous mutants display altered expression of factors that dictate dorsal identity patterning (Pratt et al., 2000). The patterning function of PAX6 is *via* regulation of *neurogenin2* (*Ngn2*), a bHLH transcription factor (Wang et al., 2011a). NGN1/2 are required for maintaining the normal population of basal progenitor cells, and in *Pax6* homozygous mutants, there is a reduction in *Ngn2* expression, and subsequently reduced a number of basal progenitor cells (Wang et al., 2011a).

Disruption of the PSB also affects the tangential migration of GABAergic interneurons. The PSB functions to limit the number of interneurons arriving at the neocortex, with increased interneurons observed in the neocortex of *Pax6* homozygous mutants along with a loss of PSB (Neyt et al., 1997; Chapouton et al., 1999). Although this outcome may be due to impaired tangential migration, it could also result from ventralization of progenitor cells (Kroll and O'Leary, 2005; Quinn et al., 2007). In these mutants, ventral GABAergic interneuron markers *Dlx1/2, Ascl1, Gsx2,* and *Gad1* are expressed dorsally, and promote differentiation of GABAergic interneurons instead of glutamatergic neurons in cortical progenitors (Kroll and O'Leary, 2005; Long et al., 2009; Wang et al., 2013). This also suggests that *Pax6* represses these dorsal specifying transcription factors, in order to promote the generation of cortical glutamatergic neurons. Unlike early corticogenesis (E12.5), in late corticogenesis (E15.5) there is an addition of differentiated neurons acquiring a GABAergic interneuron phenotype (Schuurmans et al., 2004). These results suggest that *Pax6* controls the differentiation of glutamatergic neurons, whilst suppressing GABAergic interneuron production in late corticogenesis. However, in the diencephalon, *Pax6* also promotes the development of GABAergic interneurons (Robertshaw et al., 2013).

Furthermore, *Pax6* controls the balance between NPCs proliferation and differentiation through regulation of the cell cycle (Manuel and Price, 2005; Georgala et al., 2011). *Pax6* directly regulates various genes that promote neurogenesis, and represses genes essential for non-neuronal fates depending on the histone modifications at the target promoters (**Table 3**; Sun et al., 2015; Thakurela et al., 2016). *Pax6* homozygous mutants have shortened cell cycles at the start of corticogenesis, but as corticogenesis progresses cell cycle length increases (Estivill-Torrus et al., 2002; Mi et al., 2013). This phenomenon was observed in the cells with the longest cell cycles in the wildtype; in these cells in the *Pax6* mutant mice the shortening of the cell cycle was associated with increased neuronal differentiation (Sansom et al., 2009; Mi et al., 2013; Walcher et al., 2013). Overexpressing *Pax6* increased differentiation of cortical neural stem cells into basal progenitor cells (Sansom et al., 2009). As a result, neural stem cell proliferation is disrupted, and the quantity of neurons is also reduced (Heins et al., 2002; Jones et al., 2002; Hack et al., 2004; Georgala et al., 2011). Hence, an optimal level of *Pax6* expression is necessary for the normal growth and development of the cortex. Further evidence indicates that the balance between proliferation and differentiation is *Foxg1* dependent; *Foxg1* determines whether *Pax6* promotes proliferation or differentiation (Quintana-Urzainqui et al., 2018). *Pax6* itself is regulated by the lncRNA *PAUPAR* in human embryonic stem cells, and such regulation is necessary for cortical differentiation (Xu et al., 2021).

Paired box 6 mutation in humans can result in neurological diseases, more commonly as a result of heterozygous mutations, including intellectual disability, autism, and impaired audition (Malandrini et al., 2001; Davis et al., 2008), likely to be related to reduced cerebral cortex size (Sisodiya et al., 2001; Ellison-Wright et al., 2004). Conversely, only four patients were reported to have mutations in both *PAX6* alleles, of which two survived postnatally (Glaser et al., 1994; Schmidt-Sidor et al., 2009; Solomon et al., 2009). All cases exhibited cerebral cortical malformation, and in the two cases that died before birth, the cerebral cortex was only one-third the size of a normally developed cerebral cortex (Schmidt-Sidor et al., 2009; **Table 3**).

Pit-Oct-Unc Class 3 Homeobox 2 Gene

The POU (*Pit-1*, *Oct-1/2*, and *Unc-86*) gene family encodes a transcription factor family (Pou1f-Pou6f) of which *Pou3f2* (*Brn2*) encodes a neural transcription factor that is necessary for mammalian CNS development and also for the production of corticotropin-releasing hormone (McEvilly et al., 2002; Castro et al., 2006). *Pou3f2* regulates neuronal differentiation, migration, and upper cortical layer formation during mammalian embryogenesis (McEvilly et al., 2002; Sugitani et al., 2002; Castro et al., 2006; Dominguez et al., 2012; Chen et al., 2018). The protein contains a conserved POU domain composed of 150–160 amino acids, shared by the mammalian transcription factors *P*ituitary-specific PIT1, *O*ctamer transcription factor proteins OCT1/2, and the nematode neural transcription factor *U*NC-86 (He et al., 1989; Ryan et al., 1997; **Figure 2A**). The DNA binding region of the POU protein is composed of two elements, a POU domain of approximately 75 amino acids present near the N-terminal and a classical HD of 60 amino acids located near the C-terminal separated by a short linker sequence (**Figure 2A**; Sumiyama et al., 1996; Ryan et al., 1997). Both domains are comprised of a helix-turn-helix structure (4 alpha helices in the POU domain and 3 alpha helices in the HD), which enables DNA recognition and confers DNA-binding specificity at the third helix (Klemm et al., 1994; Cook et al., 2008).

Interactions between the POU domain and its target sequence occur by recognition followed by specific binding to the canonical ATGCAAAT octameric sequence (**Figure 2B**). However, the linker region between the POU domain and HD is flexible (Herr and Cleary, 1995). The POU3F linker can fold as an alpha-helix which allows homo- or heterodimerization with the target DNA sequence (Blaud et al., 2004). POU3F2 has been reported to

form homodimers on an octamer-like sequence of the L-amino acid decarboxylase (AADC), corticotropin (CRRH) and aldose C gene promoters in a non-cooperative fashion (Blaud et al., 2004). *Pou3f2* is located on human chromosome 6q16.1 and dysregulation of this gene has been reported in disorders such as schizophrenia and bipolar disorder, as well as in melanoma (**Table 3**; Goodall et al., 2004a; Simmons et al., 2017; Chen et al., 2018; Ding et al., 2021).

The onset expression of *Pou3f2* occurs in the VZ of the whole cortical lateral-to-medial axis during early brain development and in the paraventricular nuclei (PVN) of the hypothalamus (**Figure 1B**; He et al., 1989; Nakai et al., 1995; Dominguez et al., 2012). Using an antibody that detects both BRN1 and BRN2, POU3F2 expression was detected in radial migrating cells from the VZ up to the superficial cortical layers at P0 in mouse brain (Dominguez et al., 2012). Embryonic mice with homozygous *Pou3f2* mutations exhibit hypothalamic and pituitary deficiencies, such as hypoplastic posterior lobe of the pituitary gland and failure to express corticotropin-releasing hormone in the PVN, and die soon after birth (Nakai et al., 1995; Schonemann et al., 1995). *Pou3f2/3* (*Brn1/2*) DKO mice display an abnormal brain phenotype with decreased neocortical thickness and significant reduction of upper layer cells (Sugitani et al., 2002). The olfactory bulb is hypoplastic, the cerebellum is less foliated, accompanied by loosely packed Purkinje cells. Therefore, failure of radial migration results in cortical laminar inversion in the mutant mice (McEvilly et al., 2002; Sugitani et al., 2002). Therefore, *Pou3f2/3* transcription factors redundantly regulate cortical neuron migration and therefore layer production, in addition to neuronal differentiation (Castro et al., 2006).

Two potential mechanisms have been suggested to explain the disruption in the cortical layering defect: *via Pou3f2/3* regulation of CDK5 regulatory subunits *p35* and *p39* in migrating neurons (McEvilly et al., 2002) or through *Pou3f2/3* regulation of *Dab1* (Sugitani et al., 2002). The *Pou3f2/3* double mutant display similar phenotypic abnormalities (McEvilly et al., 2002) to Cdk5-null mutants and *p35/p39*-null mutants (Ko et al., 2001). However, *Pou3f2/3* expression is observed in both a late pool of neural precursor cells as well as in postmitotic neurons, including *Tbr1*+ cells in the cortical plate (Dominguez et al., 2012). Interestingly, when *Pou3f2* is downregulated, there is an excessive number *of Tbr1*+/*NeuroD1*+ cells accumulating within the IZ (Dominguez et al., 2012). POUF3F2 may also regulate *Dab1* as the loss of *Dab1*+ cells in neurons was observed at a later phase (McEvilly et al., 2002; Sugitani et al., 2002; Dominguez et al., 2012). Furthermore, *Dab1* expression was markedly reduced in the *Pou3f2/3* DKO mice at a late stage during which neurons fail to reach the marginal zone and remain beneath the cortical subplate (Sugitani et al., 2002).

Pit-Oct-Unc class 3 homeobox 2 interacts co-operatively with other transcription factors to regulate a number of neurodevelopmental genes, including *Ascl1* in the regulation of Notch signaling, thereby controlling cell cycle exit of progenitors in addition to neuronal differentiation and radial migration in the embryonic telencephalon (Artavanis-Tsakonas et al., 1999). Disruption of POU3F2 binding was shown to prevent transcription of Notch pathway target genes, *Delta1* and *Hes5-1* (Castro et al., 2006). In contrast, overexpression of *Pou3f2* and *Ascl1* in chick neural tube resulted in excessive migration of electroporated cells in the marginal zone of the neural tube and disrupted neuronal differentiation (Castro et al., 2006).

Pit-Oct-Unc class 3 homeobox 2 dysregulation can have severe neurodevelopmental impacts, contributing to brain malformation, neurodevelopmental delays, and neuropsychiatric disorders (**Table 3**; Castro et al., 2006; Chen et al., 2018; Hashizume et al., 2018; Westphal et al., 2018; Ding et al., 2021). *POU3f2* has been found to be associated with schizophrenia and bipolar disorder, as a hub for a gene regulatory network related to these disorders (Potkin et al., 2009; Mühleisen et al., 2014; Chen et al., 2018; Pearl et al., 2019; Ding et al., 2021). When POU3F2 is overexpressed in NSCs, several genes which are differentially expressed in the prefrontal cortex of people suffering from schizophrenia and bipolar disorder, are dysregulated. This confirms the role of POU3F2 as a key regulator of gene expression in these disorder (Pearl et al., 2019). POU3F2 and PAX6 were found to regulate the transcription of *TRIM8* (Ding et al., 2021) as well as the *VRK2* (Pearl et al., 2019), other genes associated with schizophrenia and bipolar disorder (Gandal et al., 2018; Li et al., 2018; Ding et al., 2021). While POU3F2 is crucial in regulating genes involved in CNS development, it is also a lineage-determining transcription factor crucial for the regulation of melanocytic lineage. It is overexpressed in many cancer types including carcinomas, neuroblastomas, and melanomas (Schreiber et al., 1990, 1992; Thomson et al., 1995; Leonard and Bell, 1997). Upregulation of *POU3F2* represses Microphthalmia-associated transcription factor (MITF) expression in some melanomas by binding to its promoter region, which drives the cells to adopt a more stem-like and aggressive phenotype (Goodall et al., 2004a; Bonvin et al., 2012). This upregulation is due to the activation of BRAF, a key component of the mitogen-activated protein (MAP) kinase signaling pathway (Goodall et al., 2004b).

Non-cell Autonomous and Combinatorial Roles of Homeodomain-Containing Transcription Factors

Other than the regulatory functions discussed above, homeodomain-containing transcription factors can regulate forebrain development through *non-cell autonomous* roles as well as by combinatorial modes of action. As an example, PAX6 exhibits non-cell autonomous activity in the development of other organs such as the eye and spinal cord (Collinson et al., 2004; Lesaffre et al., 2007; Di Lullo et al., 2011). This activity is due to two short sequences found within the HD, which are considered essential for secretion and internalization (Prochiantz and Joliot, 2003; Joliot and Prochiantz, 2004). This suggests that some homeobox genes encode transcription factors that have the ability to act as signaling molecules, and are capable of intercellular transfer. Disruption of extracellular PAX6 has functional consequences, leading to defective eye development, with reduction in eye size (Lesaffre et al., 2007). PAX6 extracellular activities can affect cell migration in the

embryonic chick spinal cord (Di Lullo et al., 2011). OPCs, a highly migratory cell population, were studied, also due to their delayed specification and dorsal shift in *Pax6* mutants (Sun et al., 1998). OPCs were observed to be in close proximity to PAX6+ cells, and the ablation of extracellular PAX6 resulted in reduced migration of the OPC population (Di Lullo et al., 2011). Another transcription factor shown to have non-cell autonomous activity is OTX2 in the visual cortex, during the regulation of the timing of heightened plasticity, an important timepoint for proper visual development (Lee et al., 2017; Apulei et al., 2018). Extracellular OTX2 has been shown to regulate expression of *Gadd45b*, a gene that may play a role in epigenetic gene activation, as a downstream target for modulating visual cortex plasticity (Ma et al., 2009; Apulei et al., 2018). These examples demonstrate that homeobox genes, through their encoded transcription factors can also function non-cell autonomously. However, this role is yet to be fully understood for the majority of homeobox genes.

In addition, although transcription factors display highly specific expression patterns, many are co-expressed at early stages of development, and work in a combinatorial manner. This concept of a *combinatorial code* has been well documented and studied in the spinal cord (Sugimori et al., 2007; Sagner and Briscoe, 2019). The organization and patterning of the spinal cord is initiated during the development of the neural tube, in accordance with the activities of various morphogens that induce different transcription factor families (Briscoe et al., 2000; Jessell, 2000). A similar code is under active study for forebrain development, where cortical regionalization and patterning are tightly regulated by a transcriptional network consisting of transcription factors and their regulatory elements (Ypsilanti et al., 2021). In particular, cortical expression of transcriptional network members at E11.5 in the mouse forebrain is either in a gradient or in homogenous patterns, with coregulation by transcription factors such as *Pax6*, *Emx2*, and *Nr2f1* (Muzio et al., 2002b; Ypsilanti et al., 2021). Also, transcription factors expressed in the pallium have been implicated in co-binding to regulatory elements through chromatin conformation analysis (Ypsilanti et al., 2021). Future directions employing co-immunoprecipitation, ATACseq, ChIPseq and ChIP-re-ChIP experiments will enable increased understanding of how transcription factors work cooperatively in the regulation of forebrain patterning and regionalization.

CONCLUSION

In this review of vertebrate forebrain development (**Figure 1**), selected transcription factors from the HD, paired, POU and TALE HD gene families necessary for forebrain development have been discussed and summarized (**Table 1** and **Figure 2**). Where known, gene targets of these transcription factors have been specified (**Table 2**) and correlations to human diseases, including neurodevelopmental disorders and brain tumors have been briefly outlined (**Table 3**). Further studies are necessary to delineate protein-protein interactions and to identify and characterize post-translational modifications, such as phosphorylation, sumoylation, and ubiquitination, which regulate transcription factor function and link these modifications to signaling pathways in CNS development and disease.

AUTHOR CONTRIBUTIONS

RL, AG, ER, MF, JW, and DE conceived and wrote the manuscript. RL, MF, and AG designed and generated the figures. All authors contributed to editing and approved the final version of the manuscript.

FUNDING

RL is supported by a Melbourne Research Scholarship, University of Melbourne, Parkville, VIC, Australia. DE is supported by an establishment grant from The Royal Children's Hospital Foundation, Parkville, VIC, Australia.

REFERENCES

Abellán, A., Vernier, B., Rétaux, S., and Medina, L. (2010). Similarities and differences in the forebrain expression of Lhx1 and Lhx5 between chicken and mouse: insights for understanding telencephalic development and evolution. *J. Comp. Neurol.* 518, 3512–3528. doi: 10.1002/cne.22410

Acampora, D., Avantaggiato, V., Tuorto, F., and Simeone, A. (1997). Genetic control of brain morphogenesis through Otx gene dosage requirement. *Development* 124, 3639–3650. doi: 10.1242/dev.124.18.3639

Acampora, D., Mazan, S., Lallemand, Y., Avantaggiato, V., Maury, M., Simeone, A., et al. (1995). Forebrain and midbrain regions are deleted in Otx2−/− mutants due to a defective anterior neuroectoderm specification during gastrulation. *Development* 121, 3279–3290. doi: 10.1242/dev.121.10.3279

Acampora, D., Mazan, S., Avantaggiato, V., Barone, P., Tuorto, F., Lallemand, Y., et al. (1996). Epilepsy and brain abnormalities in mice lacking the Otx1 gene. *Nat. Genet.* 14, 218–222. doi: 10.1038/ng1096-218

Acampora, D., Merlo, G. R., Paleari, L., Zerega, B., Postiglione, M. P., Mantero, S., et al. (1999). Craniofacial, vestibular and bone defects in mice lacking the Distal-less-related gene Dlx5. *Development* 126, 3795–3809. doi: 10.1242/dev.126.17.3795

Agoston, Z., Heine, P., Brill, M. S., Grebbin, B. M., Hau, A.-C., Kallenborn-Gerhardt, W., et al. (2014). Meis2 is a Pax6 co-factor in neurogenesis and dopaminergic periglomerular fate specification in the adult olfactory bulb. *Development* 141, 28–38. doi: 10.1242/dev.097295

Agoston, Z., Li, N., Haslinger, A., Wizenmann, A., and Schulte, D. (2012). Genetic and physical interaction of Meis2, Pax3 and Pax7 during dorsal midbrain development. *BMC Dev. Biol.* 12:12. doi: 10.1186/1471-213X-12-10

Allen Institute for Brain Science. (2019). *Atlas of the Developing Human Brain*. Seattle, WA, Allen Institute for Brain Science.

Aleksic, T., Bechtel, M., Krndija, D., von Wichert, G., Knobel, B., Giehl, K., et al. (2007). CUTL1 promotes tumor cell migration by decreasing proteasome-mediated Src degradation. *Oncogene* 26, 5939–5949. doi: 10.1038/sj.onc.1210398

Alifragis, P., Liapi, A., and Parnavelas, J. G. (2004). Lhx6 regulates the migration of cortical interneurons from the ventral telencephalon but does not specify their GABA phenotype. *J. Neurosci.* 24:5643. doi: 10.1523/JNEUROSCI.1245-04.2004

Al-Jaberi, N., Lindsay, S., Sarma, S., Bayatti, N., and Clowry, G. J. (2015). The early fetal development of human neocortical GABAergic interneurons. *Cereb. Cortex* 25, 631–645. doi: 10.1093/cercor/bht254

Allen, Z. J., Waclaw, R. R., Colbert, M. C., and Campbell, K. (2007). Molecular identity of olfactory bulb interneurons: transcriptional codes of periglomerular neuron subtypes. *J. Mol. Histol.* 38, 517–525. doi: 10.1007/s10735-007-9115-4

Alvarez-Buylla, A., and Lim, D. A. (2004). For the long run: maintaining germinal niches in the adult brain. *Neuron* 41, 683–686. doi: 10.1016/s0896-6273(04)00111-4

Amin, S., Donaldson, I. J., Zannino, D. A., Hensman, J., Rattray, M., Losa, M., et al. (2015). Hoxa2 selectively enhances Meis binding to change a branchial arch ground state. *Dev. Cell* 32, 265–277. doi: 10.1016/j.devcel.2014.12.024

Anderson, D. J., Choi, G., and Zhou, Q. (2002a). Olig genes and the genetic logic of CNS neural cell fate determination. *Clin. Neurosci. Res.* 2, 17–28. doi: 10.1016/s1566-2772(02)00014-2

Anderson, S., Eisenstat, D., Shi, L., and Rubenstein, J. (1997a). Interneuron migration from basal forebrain to neocortex: dependence on Dlx genes. *Science* 278, 474–476. doi: 10.1126/science.278.5337.474

Anderson, S., Marín, O., Horn, C., Jennings, K., and Rubenstein, J. (2001). Distinct cortical migrations from the medial and lateral ganglionic eminences. *Development* 128, 353–363. doi: 10.1242/dev.128.3.353

Anderson, S. A., Kaznowski, C. E., Horn, C., Rubenstein, J. L., and McConnell, S. K. (2002b). Distinct origins of neocortical projection neurons and interneurons in vivo. *Cereb. Cortex* 12, 702–709. doi: 10.1093/cercor/12.7.702

Anderson, S. A., Qiu, M., Bulfone, A., Eisenstat, D. D., Meneses, J., Pedersen, R., et al. (1997b). Mutations of the homeobox genes Dlx-1 and Dlx-2 disrupt the striatal subventricular zone and differentiation of late born striatal neurons. *Neuron* 19, 27–37. doi: 10.1016/s0896-6273(00)80345-1

Andrews, W., Liapi, A., Plachez, C., Camurri, L., Zhang, J., Mori, S., et al. (2006). Robo1 regulates the development of major axon tracts and interneuron migration in the forebrain. *Development* 133, 2243–2252. doi: 10.1242/dev.02379

Ang, S.-L., Jin, O., Rhinn, M., Daigle, N., Stevenson, L., and Rossant, J. (1996). A targeted mouse Otx2 mutation leads to severe defects in gastrulation and formation of axial mesoderm and to deletion of rostral brain. *Development* 122, 243–252. doi: 10.1242/dev.122.1.243

Apulei, J., Kim, N., Testa, D., Ribot, J., Morizet, D., Bernard, C., et al. (2018). Non-cell autonomous otx2 homeoprotein regulates visual cortex plasticity through Gadd45b/g. *Cereb. Cortex* 29, 2384–2395. doi: 10.1093/cercor/bhy108

Artavanis-Tsakonas, S., Rand, M. D., and Lake, R. J. (1999). Notch signaling: cell fate control and signal integration in development. *Science* 284, 770–776. doi: 10.1126/science.284.5415.770

Attanasio, C., Nord, A. S., Zhu, Y., Blow, M. J., Biddie, S. C., Mendenhall, E. M., et al. (2014). Tissue-specific SMARCA4 binding at active and repressed regulatory elements during embryogenesis. *Genome Res.* 24, 920–929. doi: 10.1101/gr.168930.113

Ayala, R., Shu, T., and Tsai, L.-H. (2007). Trekking across the brain: the journey of neuronal migration. *Cell* 128, 29–43. doi: 10.1016/j.cell.2006.12.021

Aykut, B., Ochs, M., Radhakrishnan, P., Brill, A., Höcker, H., Schwarz, S., et al. (2017). EMX2 gene expression predicts liver metastasis and survival in colorectal cancer. *BMC Cancer* 17:555. doi: 10.1186/s12885-017-3556-2

Bach, I. (2000). The LIM domain: regulation by association. *Mech. Dev.* 91, 5–17. doi: 10.1016/s0925-4773(99)00314-7

Bachy, I., and Rétaux, S. (2006). GABAergic specification in the basal forebrain is controlled by the LIM-hd factor Lhx7. *Dev. Biol.* 291, 218–226. doi: 10.1016/j.ydbio.2005.10.023

Bardet, S., Ferran, J., Sanchez-Arrones, L., and Puelles, L. (2010). Ontogenetic expression of sonic hedgehog in the chicken subpallium. *Front. Neuroanat.* 4:28. doi: 10.3389/fnana.2010.00028

Barington, M., Risom, L., Ek, J., Uldall, P., and Ostergaard, E. (2018). A recurrent de novo CUX2 missense variant associated with intellectual disability, seizures, and autism spectrum disorder. *Eur. J. Hum. Genet.* 26, 1388–1391. doi: 10.1038/s41431-018-0184-5

Batista-Brito, R., Close, J., Machold, R., and Fishell, G. (2008). The distinct temporal origins of olfactory bulb interneuron subtypes. *J. Neurosci.* 28:3966. doi: 10.1523/JNEUROSCI.5625-07.2008

Bayer, S. A. (1983). 3H-thymidine-radiographic studies of neurogenesis in the rat olfactory bulb. *Exp. Brain Res.* 50, 329–340. doi: 10.1007/BF00239197

Beckmann, J., Vitobello, A., Ferralli, J., Brož, D. K., Rijli, F. M., and Chiquet-Ehrismann, R. (2011). Human teneurin-1 is a direct target of the homeobox transcription factor EMX2 at a novel alternate promoter. *BMC Dev. Biol.* 11:35. doi: 10.1186/1471-213X-11-35

Bertrand, N., Castro, D. S., and Guillemot, F. (2002). Proneural genes and the specification of neural cell types. *Nat. Rev. Neurosci.* 3, 517–530. doi: 10.1038/nrn874

Bienvenu, T., Poirier, K., Friocourt, G., Bahi, N., Beaumont, D., Fauchereau, F., et al. (2002). ARX. *Hum. Mol. Genet.* 11, 981–991.

Bielle, F., Griveau, A., Narboux-Nême, N., Vigneau, S., Sigrist, M., Arber, S., et al. (2005). Multiple origins of Cajal-Retzius cells at the borders of the developing pallium. *Nat. Neurosci.* 8, 1002–1012. doi: 10.1038/nn1511

Bilioni, A., Craig, G., Hill, C., and McNeill, H. (2005). Iroquois transcription factors recognize a unique motif to mediate transcriptional repression in vivo. *Proc. Natl. Acad. Sci.* 102, 14671–14676. doi: 10.1073/pnas.0502480102

Billeter, M., Güntert, P., Luginbühl, P., and Wüthrich, K. (1996). Hydration and DNA recognition by homeodomains. *Cell* 85, 1057–1065. doi: 10.1016/s0092-8674(00)81306-9

Bishop, K. M., Garel, S., Nakagawa, Y., Rubenstein, J. L., and O'Leary, D. D. (2003). Emx1 and Emx2 cooperate to regulate cortical size, lamination, neuronal differentiation, development of cortical efferents, and thalamocortical pathfinding. *J. Comp. Neurol.* 457, 345–360. doi: 10.1002/cne.10549

Bishop, K. M., Goudreau, G., and O'Leary, D. D. M. (2000). Regulation of area identity in the mammalian neocortex by Emx2 and Pax6. *Science* 288, 344–349. doi: 10.1126/science.288.5464.344

Bishop, K. M., Rubenstein, J. L., and O'Leary, D. D. (2002). Distinct actions of Emx1, Emx2, andPax6 in regulating the specification of areas in the developing neocortex. *J. Neurosci.* 22, 7627–7638. doi: 10.1523/JNEUROSCI.22-17-07627.2002

Bjerke, G. A., Hyman-Walsh, C., and Wotton, D. (2011). Cooperative transcriptional activation by Klf4, Meis2, and Pbx1. *Mol. Cell. Biol.* 31, 3723–3733. doi: 10.1128/MCB.01456-10

Blaud, M., Vossen, C., Joseph, G., Alazard, R., Erard, M., and Nieto, L. (2004). Characteristic patterns of N Oct-3 binding to a set of neuronal promoters. *J. Mol. Biol.* 339, 1049–1058. doi: 10.1016/j.jmb.2004.04.035

Blochlinger, K., Bodmer, R., Jack, J., Jan, L. Y., and Jan, Y. N. (1988). Primary structure and expression of a product from cut, a locus involved in specifying sensory organ identity in Drosophila. *Nature* 333, 629–635. doi: 10.1038/333629a0

Bodmer, R., Barbel, S., Sheperd, S., Jack, J. W., Jan, L. Y., and Jan, Y. N. (1987). Transformation of sensory organs by mutations of the cut locus of *D. melanogaster*. *Cell* 51, 293–307. doi: 10.1016/0092-8674(87)90156-5

Bokoch, G. M. (2003). Biology of the p21-activated kinases. *Annu. Rev. Biochem.* 72, 743–781. doi: 10.1146/annurev.biochem.72.121801.161742

Bosse, A., Zülch, A., Becker, M. B., Torres, M., ómez-Skarmeta, J. L. G., Modolell, J., et al. (1997). Identification of the vertebrate Iroquois homeobox gene family with overlapping expression during early development of the nervous system. *Mech. Dev.* 69, 169–181. doi: 10.1016/s0925-4773(97)00165-2

Bonvin, E., Falletta, P., Shaw, H., Delmas, V., and Goding, C. R. (2012). A phosphatidylinositol 3-kinase-Pax3 axis regulates Brn-2 expression in melanoma. *Mol. Cell. Biol.* 32, 4674–4683. doi: 10.1128/MCB.01067-12

Boon, K., Eberhart, C. G., and Riggins, G. J. (2005). Genomic amplification of orthodenticle homologue 2 in medulloblastomas. *Cancer Res.* 65, 703–707.

Briata, P., Ilengo, C., Bobola, N., and Corte, G. (1999). Binding properties of the human homeodomain protein OTX2 to a DNA target sequence. *FEBS Lett.* 445, 160–164. doi: 10.1016/s0014-5793(99)00113-1

Briscoe, J., Pierani, A., Jessell, T. M., and Ericson, J. (2000). A homeodomain protein code specifies progenitor cell identity and neuronal fate in the ventral neural tube. *Cell* 101, 435–445. doi: 10.1016/s0092-8674(00)80853-3

Briscoe, J., Sussel, L., Serup, P., Hartigan-O'Connor, D., Jessell, T. M., Rubenstein, J. L., et al. (1999). Homeobox gene Nkx2.2 and specification of neuronal identity by graded Sonic hedgehog signalling. *Nature* 398, 622–627. doi: 10.1038/19315

Britz, O., Mattar, P., Nguyen, L., Langevin, L.-M., Zimmer, C., Alam, S., et al. (2006). A role for proneural genes in the maturation of cortical progenitor cells. *Cereb. Cortex* 16, i138–i151. doi: 10.1093/cercor/bhj168

Bulstrode, H., Johnstone, E., Marques-Torrejon, M. A., Ferguson, K. M., Bressan, R. B., Blin, C., et al. (2017). Elevated FOXG1 and SOX2 in glioblastoma enforces neural stem cell identity through transcriptional control of cell cycle and epigenetic regulators. *Genes Dev.* 31, 757–773. doi: 10.1101/gad.293027.116

Bunt, J., de Haas, T. G., Hasselt, N. E., Zwijnenburg, D. A., Koster, J., Versteeg, R., et al. (2010). Regulation of cell cycle genes and induction of senescence by overexpression of OTX2 in medulloblastoma cell lines. *Mol. Cancer Res.* 8, 1344–1357. doi: 10.1158/1541-7786.MCR-09-0546

Bunt, J., Hasselt, N. E., Zwijnenburg, D. A., Koster, J., Versteeg, R., and Kool, M. (2011). Joint binding of OTX2 and MYC in promotor regions is associated with high gene expression in medulloblastoma. *PLoS One* 6:e26058. doi: 10.1371/journal.pone.0026058

Butterworth, F., and King, R. (1965). The developmental genetics of apterous mutants of *Drosophila melanogaster*. *Genetics* 52, 1153. doi: 10.1093/genetics/52.6.1153

Butt, S. J., Fuccillo, M., Nery, S., Noctor, S., Kriegstein, A., Corbin, J. G., et al. (2005). The temporal and spatial origins of cortical interneurons predict their physiological subtype. *Neuron* 48, 591–604. doi: 10.1016/j.neuron.2005.09.034

Butt, S. J., Sousa, V. H., Fuccillo, M. V., Hjerling-Leffler, J., Miyoshi, G., Kimura, S., et al. (2008). The requirement of Nkx2-1 in the temporal specification of cortical interneuron subtypes. *Neuron* 59, 722–732. doi: 10.1016/j.neuron.2008.07.031

Bulchand, S., Grove, E., Porter, F., and Tole, S. (2001). LIM-homeodomain gene Lhx2 regulates the formation of the cortical hem. *Mech. Dev.* 100, 165–175. doi: 10.1016/s0925-4773(00)00515-3

Bürglin, T. R. (1997). Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. *Nucleic Acids Res.* 25, 4173–4180. doi: 10.1093/nar/25.21.4173

Caqueret, A., Boucher, F., and Michaud, J. L. (2006). Laminar organization of the early developing anterior hypothalamus. *Dev. Biol.* 298, 95–106. doi: 10.1016/j.ydbio.2006.06.019

Carney, R. S., Cocas, L. A., Hirata, T., Mansfield, K., and Corbin, J. G. (2009). Differential regulation of telencephalic pallial–subpallial boundary patterning by Pax6 and Gsh2. *Cereb. Cortex* 19, 745–759. doi: 10.1093/cercor/bhn123

Carri, A. D., Onorati, M., Lelos, M. J., Castiglioni, V., Faedo, A., Menon, R., et al. (2013). Developmentally coordinated extrinsic signals drive human pluripotent stem cell differentiation toward authentic DARPP-32+ medium-sized spiny neurons. *Development* 140, 301–312. doi: 10.1242/dev.084608

Castro, D. S., Skowronska-Krawczyk, D., Armant, O., Donaldson, I. J., Parras, C., Hunt, C., et al. (2006). Proneural bHLH and Brn proteins coregulate a neurogenic program through cooperative binding to a conserved DNA motif. *Dev. Cell* 11, 831–844. doi: 10.1016/j.devcel.2006.10.006

Cau, E., Casarosa, S., and Guillemot, F. (2002). Mash1 and Ngn1 control distinct steps of determination and differentiation in the olfactory sensory neuron lineage. *Development* 129, 1871–1880. doi: 10.1242/dev.129.8.1871

Cecconi, F., Proetzel, G., Alvarez-Bolado, G., Jay, D., and Gruss, P. (1997). Expression of Meis2, a Knotted-related murine homeobox gene, indicates a role in the differentiation of the forebrain and the somitic mesoderm. *Dev. Dyn.* 210, 184–190. doi: 10.1002/(SICI)1097-0177(199710)210:2<184::AID-AJA10>3.0.CO;2-E

Chang, C. P., Jacobs, Y., Nakamura, T., Jenkins, N. A., Copeland, N. G., and Cleary, M. L. (1997). Meis proteins are major in vivo DNA binding partners for wild-type but not chimeric Pbx proteins. *Mol. Cell Biol.* 17, 5679–5687. doi: 10.1128/MCB.17.10.5679

Chapman, H., Riesenberg, A., Ehrman, L. A., Kohli, V., Nardini, D., Nakafuku, M., et al. (2018). Gsx transcription factors control neuronal versus glial specification in ventricular zone progenitors of the mouse lateral ganglionic eminence. *Dev. Biol.* 442, 115–126. doi: 10.1016/j.ydbio.2018.07.005

Chapman, H., Waclaw, R. R., Pei, Z., Nakafuku, M., and Campbell, K. (2013). The homeobox gene Gsx2 controls the timing of oligodendroglial fate specification in mouse lateral ganglionic eminence progenitors. *Development* 140, 2289–2298. doi: 10.1242/dev.091090

Chapouton, P., Gartner, A., and Gotz, M. (1999). The role of Pax6 in restricting cell migration between developing cortex and basal ganglia. *Development* 126, 5569–5579. doi: 10.1242/dev.126.24.5569

Chen, C., Meng, Q., Xia, Y., Ding, C., Wang, L., Dai, R., et al. (2018). The transcription factor POU3F2 regulates a gene coexpression network in brain tissue from patients with psychiatric disorders. *Sci. Transl. Med.* 10:eaat8178. doi: 10.1126/scitranslmed.aat8178

Chen, G., Sima, J., Jin, M., Wang, K.-y., Xue, X.-j., Zheng, W., et al. (2008). Semaphorin-3A guides radial migration of cortical neurons during development. *Nat. Neurosci.* 11, 36–44. doi: 10.1038/nn2018

Cho, G., Nasrallah, M. P., Lim, Y., and Golden, J. A. (2012). Distinct DNA binding and transcriptional repression characteristics related to different ARX mutations. *Neurogenetics* 13, 23–29. doi: 10.1007/s10048-011-0304-7

Cho, M. H., McDonald, M.-L. N., Zhou, X., Mattheisen, M., Castaldi, P. J., Hersh, C. P., et al. (2014). Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir. Med.* 2, 214–225. doi: 10.1016/S2213-2600(14)70002-5

Cholfin, J. A., and Rubenstein, J. L. (2008). Frontal cortex subdivision patterning is coordinately regulated by Fgf8, Fgf17, and Emx2. *J. Comp. Neurol.* 509, 144–155. doi: 10.1002/cne.21709

Cobos, I., Borello, U., and Rubenstein, J. L. R. (2007). Dlx transcription factors promote migration through repression of axon and dendrite growth. *Neuron* 54, 873–888. doi: 10.1016/j.neuron.2007.05.024

Cobos, I., Broccoli, V., and Rubenstein, J. L. (2005a). The vertebrate ortholog of Aristaless is regulated by Dlx genes in the developing forebrain. *J. Comp. Neurol.* 483, 292–230. doi: 10.1002/cne.20405

Cobos, I., Calcagnotto, M. E., Vilaythong, A. J., Thwin, M. T., Noebels, J. L., Baraban, S. C., et al. (2005b). Mice lacking Dlx1 show subtype-specific loss of interneurons, reduced inhibition and epilepsy. *Nat. Neurosci.* 8, 1059–1068. doi: 10.1038/nn1499

Cocas, L. A., Georgala, P. A., Mangin, J.-M., Clegg, J. M., Kessaris, N., Haydar, T. F., et al. (2011). Pax6 is required at the telencephalic pallial–subpallial boundary for the generation of neuronal diversity in the postnatal limbic system. *J. Neurosci.* 31, 5313–5324. doi: 10.1523/JNEUROSCI.3867-10.2011

Cohen, D. R., Cheng, C. W., Cheng, S. H., and Hui, C.-c (2000). Expression of two novel mouse Iroquois homeobox genes during neurogenesis. *Mech. Dev.* 91, 317–321. doi: 10.1016/s0925-4773(99)00263-4

Cohen, S. M., Brönner, G., Küttner, F., Jürgens, G., and Jäckle, H. (1989). Distal-less encodes a homoeodomain protein required for limb development in Drosophila. *Nature* 338, 432–434. doi: 10.1038/338432a0

Colasante, G., Collombat, P., Raimondi, V., Bonanomi, D., Ferrai, C., Maira, M., et al. (2008). Arx is a direct target of Dlx2 and thereby contributes to the tangential migration of GABAergic interneurons. *J. Neurosci.* 28, 10674–10678. doi: 10.1523/JNEUROSCI.1283-08.2008

Colasante, G., Sessa, A., Crispi, S., Calogero, R., Mansouri, A., Collombat, P., et al. (2009). Arx acts as a regional key selector gene in the ventral telencephalon mainly through its transcriptional repression activity. *Dev. Biol.* 334, 59–67. doi: 10.1016/j.ydbio.2009.07.014

Colasante, G., Simonet, J. C., Calogero, R., Crispi, S., Sessa, A., Cho, G., et al. (2015). ARX regulates cortical intermediate progenitor cell expansion and upper layer neuron formation through repression of Cdkn1c. *Cereb. Cortex* 25, 322–333. doi: 10.1093/cercor/bht222

Collinson, J. M., Chanas, S. A., Hill, R. E., and West, J. D. (2004). Corneal development, limbal stem cell function, and corneal epithelial cell migration in the pax6 +/- mouse. *Invest. Ophthalmol. Vis. Sci.* 45, 1101–1108. doi: 10.1167/iovs.03-1118

Colombo, E., Collombat, P., Colasante, G., Bianchi, M., Long, J., Mansouri, A., et al. (2007). Inactivation of Arx, the murine ortholog of the X-linked lissencephaly with ambiguous genitalia gene, leads to severe disorganization of the ventral telencephalon with impaired neuronal migration and differentiation. *J. Neurosci.* 27, 4786–4798. doi: 10.1523/JNEUROSCI.0417-07.2007

Colombo, E., Galli, R., Cossu, G., Gécz, J., and Broccoli, V. (2004). Mouse orthologue of ARX, a gene mutated in several X-linked forms of mental retardation and epilepsy, is a marker of adult neural stem cells and forebrain GABAergic neurons. *Dev. Dyn.* 231, 631–639. doi: 10.1002/dvdy.20164

Conforto, T. L., Zhang, Y., Sherman, J., and Waxman, D. J. (2012). Impact of CUX2 on the female mouse liver transcriptome: activation of female-biased genes and repression of male-biased genes. *Mol. Cell. Biol.* 32, 4611–4627. doi: 10.1128/MCB.00886-12

Cook, A. L., Sturm, R. A., and domain, P. O. U. (2008). transcription factors: BRN2 as a regulator of melanocytic growth and tumourigenesis. *Pigment Cell Melanoma Research* 21, 611–626. doi: 10.1111/j.1755-148X.2008.00510.x

Coqueret, O., Bérubé, G., and Nepveu, A. (1996). DNA binding by cut homeodomain proteins is down-modulated by protein kinase C*. *J. Biol. Chem.* 271, 24862–24868. doi: 10.1074/jbc.271.40.24862

Coqueret, O., Martin, N., Bérubé, G., Rabbat, M., Litchfield, D. W., and Nepveu, A. (1998). DNA binding by cut homeodomain proteins is down-modulated by casein kinase II*. *J. Biol. Chem.* 273, 2561–2566. doi: 10.1074/jbc.273.5.2561

Corbin, J. G., Gaiano, N., Machold, R. P., Langston, A., and Fishell, G. (2000). The Gsh2 homeodomain gene controls multiple aspects of telencephalic development. *Development* 127, 5007–5020. doi: 10.1242/dev.127.23.5007

Corbin, J. G., Rutlin, M., Gaiano, N., and Fishell, G. (2003). Combinatorial function of the homeodomain proteins Nkx2. 1 and Gsh2 in ventral telencephalic patterning. *Development* 130, 4895–4906. doi: 10.1242/dev.00717

Cowell, I. G., and Hurst, H. C. (1994). Transcriptional repression by the human bZIP factor E4BP4: definition of a minimal repression domain. *Nucleic Acids Res.* 22, 59–65. doi: 10.1093/nar/22.1.59

Craddock, N., Dawson, E., Burge, S., Parfitt, L., Mant, B., Roberts, Q., et al. (1993). The gene for Darier's disease maps to chromosome 12q23-q24. *Hum. Mol. Genet.* 2, 1941–1943. doi: 10.1093/hmg/2.11.1941

Cubelos, B., Sebastián-Serrano, A., Beccari, L., Calcagnotto, M. E., Cisneros, E., Kim, S., et al. (2010). Cux1 and Cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. *Neuron* 66, 523–535. doi: 10.1016/j.neuron.2010.04.038

Cubelos, B., Sebastián-Serrano, A, Kim, S., Moreno-Ortiz, C., Redondo, J. M., Walsh, C. A., et al. (2008a). Cux-2 controls the proliferation of neuronal intermediate precursors of the cortical subventricular zone. *Cereb. Cortex* 18, 1758–1770. doi: 10.1093/cercor/bhm199

Cubelos, B., Sebastián-Serrano, A., Kim, S., Redondo, J. M., Walsh, C., and Nieto, M. (2008b). Cux-1 and Cux-2 control the development of Reelin expressing cortical interneurons. *Dev. Neurobiol.* 68, 917–925. doi: 10.1002/dneu.20626

Dai, S., Li, J., Zhang, H., Chen, X., Guo, M., Chen, Z., et al. (2020). Structural basis for DNA recognition by FOXG1 and the characterization of disease-causing FOXG1 mutations. *J. Mol. Biol.* 432, 6146–6156. doi: 10.1016/j.jmb.2020.10.007

Dateki, S., Kosaka, K., Hasegawa, K., Tanaka, H., Azuma, N., Yokoya, S., et al. (2010). Heterozygous orthodenticle homeobox 2 mutations are associated with variable pituitary phenotype. *J. Clin. Endocrinol. Metab.* 95, 756–764. doi: 10.1210/jc.2009-1334

Davis, L., Meyer, K., Rudd, D., Librant, A., Epping, E., Sheffield, V., et al. (2008). Pax6 3' deletion results in aniridia, autism and mental retardation. *Hum. Genet.* 123, 371–378. doi: 10.1007/s00439-008-0484-x

Dawid, I. B., Breen, J. J., and Toyama, R. (1998). LIM domains: multiple roles as adapters and functional modifiers in protein interactions. *Trends Genet.* 14, 156–162. doi: 10.1016/s0168-9525(98)01424-3

DeDiego, I., Smith-Fernández, A., and Fairén, A. (1994). Cortical cells that migrate beyond area boundaries: characterization of an early neuronal population in the lower intermediate zone of prenatal rats. *Eur. J. Neurosci.* 6, 983–997. doi: 10.1111/j.1460-9568.1994.tb00593.x

Dehmelt, L., and Halpain, S. (2004). The MAP2/Tau family of microtubule-associated proteins. *Genome Biol.* 6:204.

Demos, M. K., Fullston, T., Partington, M. W., Gécz, J., and Gibson, W. T. (2009). Clinical study of two brothers with a novel 33 bp duplication in the ARX gene. *Am. J. Med. Genet. Part A* 149, 1482–1486. doi: 10.1002/ajmg.a.32851

De Marchis, S., Bovetti, S., Carletti, B., Hsieh, Y.-C., Garzotto, D., Peretto, P., et al. (2007). Generation of distinct types of periglomerular olfactory bulb interneurons during development and in adult mice: implication for intrinsic properties of the subventricular zone progenitor population. *J. Neurosci.* 27, 657–664. doi: 10.1523/JNEUROSCI.2870-06.2007

De Mori, R., Severino, M., Mancardi, M. M., Anello, D., Tardivo, S., Biagini, T., et al. (2019). Agenesis of the putamen and globus pallidus caused by recessive mutations in the homeobox gene GSX2. *Brain* 142, 2965–2978. doi: 10.1093/brain/awz247

Depew, M. J., Liu, J. K., Long, J. E., Presley, R., Meneses, J. J., Pedersen, R. A., et al. (1999). Dlx5 regulates regional development of the branchial arches and sensory capsules. *Development* 126, 3831–3846. doi: 10.1242/dev.126.17.3831

Desmaris, E., Keruzore, M., Saulnier, A., Ratié, L., Assimacopoulos, S., De Clercq, S., et al. (2018). DMRT5, DMRT3, and EMX2 cooperatively repress GSX2 at the pallium–subpallium boundary to maintain cortical identity in dorsal telencephalic progenitors. *J. Neurosci.* 38, 9105–9121. doi: 10.1523/JNEUROSCI.0375-18.2018

Diez del Corral, R., Aroca, P., mez-Skarmeta, J. L. G., Cavodeassi, F., and Modolell, J. (1999). The Iroquois homeodomain proteins are required to specify body wall identity in Drosophila. *Genes Dev.* 13, 1754–1761. doi: 10.1101/gad.13.13.1754

Ding, C., Zhang, C., Kopp, R., Kuney, L., Meng, Q., Wang, L., et al. (2021). Transcription factor POU3F2 regulates TRIM8 expression contributing to cellular functions implicated in schizophrenia. *Mol. Psychiatry* 26, 3444–3460. doi: 10.1038/s41380-020-00877-2

Di Lullo, E., Haton, C., Le Poupon, C., Volovitch, M., Joliot, A., Thomas, J.-L., et al. (2011). Paracrine Pax6 activity regulates oligodendrocyte precursor cell migration in the chick embryonic neural tube. *Development* 138, 4991–5001. doi: 10.1242/dev.066282

Dohrmann, C., Gruss, P., and Lemaire, L. (2000). Pax genes and the differentiation of hormone-producing endocrine cells in the pancreas. *Mech. Dev.* 92, 47–54. doi: 10.1016/s0925-4773(99)00324-x

Dominguez, M. H., Ayoub, A. E., and Rakic, P. (2012). POU-III transcription factors (Brn1, Brn2, and Oct6) influence neurogenesis, molecular identity, and migratory destination of upper-layer cells of the cerebral cortex. *Cereb. Cortex* 23, 2632–2643. doi: 10.1093/cercor/bhs252

Domínguez, L., González, A., and Moreno, N. (2015). Patterns of hypothalamic regionalization in amphibians and reptiles: common traits revealed by a genoarchitectonic approach. *Front. Neuroanat.* 9:3. doi: 10.3389/fnana.2015.00003

Donegan, J. J., Boley, A. M., Glenn, J. P., Carless, M. A., and Lodge, D. J. (2020). Developmental alterations in the transcriptome of three distinct rodent models of schizophrenia. *PLoS One* 15:e0232200. doi: 10.1371/journal.pone.0232200

Drabikowski, K., Trzebiatowska, A., and Chiquet-Ehrismann, R. (2005). ten-1, an essential gene for germ cell development, epidermal morphogenesis, gonad migration, and neuronal pathfinding in *Caenorhabditis elegans*. *Dev. Biol.* 282, 27–38. doi: 10.1016/j.ydbio.2005.02.017

Du, T., Xu, Q., Ocbina, P. J., and Anderson, S. A. (2008). NKX2. 1 specifies cortical interneuron fate by activating Lhx6. *Development* 135, 1559–1567. doi: 10.1242/dev.015123

Duan, D., Fu, Y., Paxinos, G., and Watson, C. (2013). Spatiotemporal expression patterns of Pax6 in the brain of embryonic, newborn, and adult mice. *Brain Struct. Funct.* 218, 353–372. doi: 10.1007/s00429-012-0397-2

Eisenstat, D. D., Liu, J. K., Mione, M., Zhong, W., Yu, G., Anderson, S. A., et al. (1999). DLX-1, DLX-2, and DLX-5 expression define distinct stages of basal forebrain differentiation. *J. Comp. Neurol.* 414, 217–237. doi: 10.1002/(sici)1096-9861(19991115)414:2<217::aid-cne6>3.0.co;2-i

Ellenberger, T., Fass, D., Arnaud, M., and Harrison, S. C. (1994). Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. *Genes Dev.* 8, 970–980. doi: 10.1101/gad.8.8.970

Ellison-Wright, Z., Heyman, I., Frampton, I., Rubia, K., Chitnis, X., Ellison-Wright, I., et al. (2004). Heterozygous PAX6 mutation, adult brain structure and fronto-striato-thalamic function in a human family. *Eur. J. Neurosci.* 19, 1505–1512. doi: 10.1111/j.1460-9568.2004.03236.x

El Nagar, S., Chakroun, A., Le Greneur, C., Figarella-Branger, D., Di Meglio, T., Lamonerie, T., et al. (2018). Otx2 promotes granule cell precursor proliferation and shh-dependent medulloblastoma maintenance in vivo. *Oncogenesis* 7, 1–12. doi: 10.1038/s41389-018-0070-6

Emery, B., and Lu, Q. R. (2015). Transcriptional and epigenetic regulation of oligodendrocyte development and myelination in the central nervous system. *Cold Spring Harb. Perspect. Biol.* 7:a020461. doi: 10.1101/cshperspect.a020461

Endele, S., Rosenberger, G., Geider, K., Popp, B., Tamer, C., Stefanova, I., et al. (2010). Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat. Genet.* 42, 1021–1026. doi: 10.1038/ng.677

Ericson, J., Muhr, J., Placzek, M., Lints, T., Jessel, T. á, and Edlund, T. (1995). Sonic hedgehog induces the differentiation of ventral forebrain neurons: a common signal for ventral patterning within the neural tube. *Cell* 81, 747–756. doi: 10.1016/0092-8674(95)90536-7

Ericson, J., Rashbass, P., Schedl, A., Brenner-Morton, S., Kawakami, A., Van Heyningen, V., et al. (1997). Pax6 controls progenitor cell identity and neuronal fate in response to graded Shh signaling. *Cell* 90, 169–180. doi: 10.1016/s0092-8674(00)80323-2

Erlander, M. G., Tillakaratne, N. J., Feldblum, S., Patel, N., and Tobin, A. J. (1991). Two genes encode distinct glutamate decarboxylases. *Neuron* 7, 91–100. doi: 10.1016/0896-6273(91)90077-d

Estivill-Torrus, G., Pearson, H., van Heyningen, V., Price, D. J., and Rashbass, P. (2002). Pax6 is required to regulate the cell cycle and the rate of progression

from symmetrical to asymmetrical division in mammalian cortical progenitors. *Development* 129, 455–466. doi: 10.1242/dev.129.2.455

Falcone, C., Daga, A., Leanza, G., and Mallamaci, A. (2016). Emx2 as a novel tool to suppress glioblastoma. *Oncotarget* 7:41005. doi: 10.18632/oncotarget.9322

Fazel Darbandi, S., Poitras, L., Monis, S., Lindtner, S., Yu, M., Hatch, G., et al. (2016). Functional consequences of I56ii Dlx enhancer deletion in the developing mouse forebrain. *Dev. Biol.* Online ahead of print, doi: 10.1016/j.ydbio.2016.10.015,

Figueres-Oñate, M., and López-Mascaraque, L. L. (2016). Adult olfactory bulb interneuron phenotypes identified by targeting embryonic and postnatal neural progenitors. *Front. Neurosci.* 10:194. doi: 10.3389/fnins.2016.00194

Filbin, M. G., Tirosh, I., Hovestadt, V., Shaw, M. L., Escalante, L. E., Mathewson, N. D., et al. (2018). Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* 360, 331–335. doi: 10.1126/science.aao4750

Flames, N., Pla, R., Gelman, D. M., Rubenstein, J. L., Puelles, L., and Marín, O. (2007). Delineation of multiple subpallial progenitor domains by the combinatorial expression of transcriptional codes. *J. Neurosci.* 27, 9682–9695. doi: 10.1523/JNEUROSCI.2750-07.2007

Flandin, P., Zhao, Y., Vogt, D., Jeong, J., Long, J., Potter, G., et al. (2011). Lhx6 and Lhx8 coordinately induce neuronal expression of Shh that controls the generation of interneuron progenitors. *Neuron* 70, 939–950. doi: 10.1016/j.neuron.2011.04.020

Florian, C., Bahi-Buisson, N., and Bienvenu, T. (2011). FOXG1-related disorders: from clinical description to molecular genetics. *Mol. Syndromol.* 2, 153–163. doi: 10.1159/000327329

Feng, W., Simoes-de-Souza, F., Finger, T. E., Restrepo, D., and Williams, T. (2009). Disorganized olfactory bulb lamination in mice deficient for transcription factor AP-2epsilon. *Mol. Cell Neurosci.* 42, 161–171. doi: 10.1016/j.mcn.2009.06.010

Fode, C., Ma, Q., Casarosa, S., Ang, S.-L., Anderson, D. J., and Guillemot, F. (2000). A role for neural determination genes in specifying the dorsoventral identity of telencephalic neurons. *Genes Dev.* 14, 67–80. doi: 10.1101/gad.14.1.67

Fogarty, M., Grist, M., Gelman, D., Marín, O., Pachnis, V., and Kessaris, N. (2007). Spatial genetic patterning of the embryonic neuroepithelium generates GABAergic interneuron diversity in the adult cortex. *J. Neurosci.* 27, 10935–10946. doi: 10.1523/JNEUROSCI.1629-07.2007

Fragkouli, A., Hearn, C., Errington, M., Cooke, S., Grigoriou, M., Bliss, T., et al. (2005). Loss of forebrain cholinergic neurons and impairment in spatial learning and memory in LHX7-deficient mice. *Eur. J. Neurosci.* 21, 2923–2938. doi: 10.1111/j.1460-9568.2005.04141.x

Franco, S. J., Martinez-Garay, I., Gil-Sanz, C., Harkins-Perry, S. R., and Müller, U. (2011). Reelin regulates cadherin function via Dab1/Rap1 to control neuronal migration and lamination in the neocortex. *Neuron* 69, 482–497. doi: 10.1016/j.neuron.2011.01.003

Frantz, G. D., Weimann, J. M., Levin, M. E., and McConnell, S. K. (1994). Otx1 and Otx2 define layers and regions in developing cerebral cortex and cerebellum. *J. Neurosci.* 14, 5725–5740. doi: 10.1523/JNEUROSCI.14-10-05725.1994

Frazer, S., Prados, J., Niquille, M., Cadilhac, C., Markopoulos, F., Gomez, L., et al. (2017). Transcriptomic and anatomic parcellation of 5-HT(3A)R expressing cortical interneuron subtypes revealed by single-cell RNA sequencing. *Nat. Commun.* 8, 14219–14219. doi: 10.1038/ncomms14219

Freyd, G., Kim, S. K., and Horvitz, H. R. (1990). Novel cysteine-rich motif and homeodomain in the product of the Caenorhabditis elegans cell lineage gene lin-II. *Nature* 344, 876–879. doi: 10.1038/344876a0

Fragkouli, A., van Wijk, N. V., Lopes, R., Kessaris, N., and Pachnis, V. (2009). LIM homeodomain transcription factor-dependent specification of bipotential MGE progenitors into cholinergic and GABAergic striatal interneurons. *Development* 136, 3841–3851. doi: 10.1242/dev.038083

Friocourt, G., Kanatani, S., Tabata, H., Yozu, M., Takahashi, T., Antypa, M., et al. (2008). Cell-autonomous roles of ARX in cell proliferation and neuronal migration during corticogenesis. *J. Neurosci.* 28, 5794–5780. doi: 10.1523/JNEUROSCI.1067-08.2008

Friocourt, G., Poirier, K., Rakić, S., Parnavelas, J. G., and Chelly, J. (2006). The role of ARX in cortical development. *Eur. J. Neurosci.* 23, 869–876. doi: 10.1111/j.1460-9568.2006.04629.x

Friocourt, G., and Parnavelas, J. G. (2010). Mutations in ARX result in several defects involving GABAergic neurons. *Front. Cell. Neurosci.* 4:4. doi: 10.3389/fncel.2010.00004

Fukuchi-Shimogori, T., and Grove, E. A. (2003). Emx2 patterns the neocortex by regulating FGF positional signaling. *Nat. Neurosci.* 6, 825–831. doi: 10.1038/nn1093

Fukuchi-Shimogori, T., and Grove, E. A. (2001). Neocortex patterning by the secreted signaling molecule FGF8. *Science* 294, 1071–1074. doi: 10.1126/science.1064252

Fukuda, T., Kawano, H., Osumi, N., Eto, K., and Kawamura, K. (2000). Histogenesis of the cerebral cortex in rat fetuses with a mutation in the Pax-6 gene. *Dev. Brain Res.* 120, 65–75. doi: 10.1016/s0165-3806(99)00187-x

Furukawa, T., Morrow, E. M., and Cepko, C. L. (1997). Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiatiosn. *Cell* 91, 531–541. doi: 10.1016/s0092-8674(00)80439-0

Furusho, M., Ono, K., Takebayashi, H., Masahira, N., Kagawa, T., Ikeda, K., et al. (2006). Involvement of the Olig2 transcription factor in cholinergic neuron development of the basal forebrain. *Dev. Biol.* 293, 348–357. doi: 10.1016/j.ydbio.2006.01.031

Fulp, C. T., Cho, G., Marsh, E. D., Nasrallah, I. M., Labosky, P. A., and Golden, J. A. (2008). Identification of Arx transcriptional targets in the developing basal forebrain. *Hum. Mol. Genet.* 17, 3740–3760. doi: 10.1093/hmg/ddn271

Gandal, M. J., Haney, J. R., Parikshak, N. N., Leppa, V., Ramaswami, G., Hartl, C., et al. (2018). Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* 359, 693–697.

García-Moreno, F., and Molnár, Z. (2015). Subset of early radial glial progenitors that contribute to the development of callosal neurons is absent from avian brain. *Proc. Natl. Acad. Sci.* 112, E5058–E5067. doi: 10.1073/pnas.1506377112

Garda, A.-L., Echevarría, D., and Martínez, S. (2001). Neuroepithelial co-expression of Gbx2 and Otx2 precedes Fgf8 expression in the isthmic organizer. *Mech. Dev.* 101, 111–118. doi: 10.1016/s0925-4773(00)00567-0

Garel, S., Marín, F., Grosschedl, R., and Charnay, P. (1999). Ebf1 controls early cell differentiation in the embryonic striatum. *Development* 126, 5285–5294. doi: 10.1242/dev.126.23.5285

Gehring, W. J., Affolter, M., and Bürglin, T. (1994). Homeodomain proteins. *Annu. Rev. Biochem.* 63, 487–526.

Gelman, D. M., Martini, F. J., óbrega-Pereira, S. N., Pierani, A., Kessaris, N., and Marín, O. (2009). The embryonic preoptic area is a novel source of cortical GABAergic interneurons. *J. Neurosci.* 29, 9380–9389. doi: 10.1523/JNEUROSCI.0604-09.2009

Georgala, P. A., Carr, C. B., and Price, D. J. (2011). The role of Pax6 in forebrain development. *Dev. Neurobiol.* 71, 690–709. doi: 10.1002/dneu.20895

Ghanem, N., Jarinova, O., Amores, A., Long, Q., Hatch, G., Park, B. K., et al. (2003). Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. *Genome Res.* 13, 533–543. doi: 10.1101/gr.716103

Ghanem, N., Yu, M., Poitras, L., Rubenstein, J. L. R., and Ekker, M. (2008). Characterization of a distinct subpopulation of striatal projection neurons expressing the Dlx genes in the basal ganglia through the activity of the I56ii enhancer. *Dev. Biol.* 322, 415–424. doi: 10.1016/j.ydbio.2008.07.029

Gherzi, R., Briata, P., Boncinelli, E., Ponassi, M., Querzè, G., Viti, F., et al. (1997). The human homeodomain protein OTX2 binds to the human tenascin-C promoter and trans-represses its activity in transfected cells. *DNA Cell Biol.* 16, 559–567. doi: 10.1089/dna.1997.16.559

Glaser, T., Jepeal, L., Edwards, J. G., Young, S. R., Favor, J., and Maas, R. L. (1994). PAX6 gene dosage effect in a family with congenital cataracts, aniridia, anophthalmia and central nervous system defects. *Nat. Genet.* 7, 463–471. doi: 10.1038/ng0894-463

Glaser, B., Kirov, G., Green, E., Craddock, N., and Owen, M. J. (2005). Linkage disequilibrium mapping of bipolar affective disorder at 12q23-q24 provides evidence for association at CUX2 and FLJ3235. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 132b, 38–45. doi: 10.1002/ajmg.b.30081

Glaser, T., Walton, D. S., and Maas, R. L. (1992). Genomic structure, evolutionary conservation and aniridia mutations in the human PAX6 gene. *Nat. Genet.* 2, 232–239. doi: 10.1038/ng1192-232

Godbole, G., Roy, A., Shetty, A. S., and Tole, S. (2017). Novel functions of LHX2 and PAX6 in the developing telencephalon revealed upon combined loss of both genes. *Neural Dev.* 12, 1–8. doi: 10.1186/s13064-017-0097-y

Godbole, G., Shetty, A. S., Roy, A., D'Souza, L., Chen, B., Miyoshi, G., et al. (2018). Hierarchical genetic interactions between FOXG1 and LHX2 regulate the formation of the cortical hem in the developing telencephalon. *Development* 145:dev154583. doi: 10.1242/dev.154583

Goodall, J., Martinozzi, S., Dexter, T. J., Champeval, D., Carreira, S., Larue, L., et al. (2004a). Brn-2 expression controls melanoma proliferation and is directly regulated by β-catenin. *Mol. Cell. Biol.* 24, 915–2922.

Goodall, J., Wellbrock, C., Dexter, T. J., Roberts, K., Marais, R., and Goding, C.R. (2004b). The brn-2 transcription factor links activated braf to melanoma proliferation. *Mol. Cell. Biol.* 24, 2923–2931. doi: 10.1128/MCB.24.7.2923-2931.2004

Golonzhka, O., Nord, A., Tang, P. L. F., Lindtner, S., Ypsilanti, A. R., Ferretti, E., et al. (2015). Pbx regulates patterning of the cerebral cortex in progenitors and postmitotic neurons. *Neuron* 88, 1192–1207. doi: 10.1016/j.neuron.2015.10.045

Gómez-Skarmeta, J.-L., del Corral, R. D., de la Calle-Mustienes, E., Ferrés-Marcó, D., and Modolell, J. (1996). Araucan and caupolican, two members of the novel iroquois complex, encode homeoproteins that control proneural and vein-forming genes. *Cell* 85, 95–105. doi: 10.1016/s0092-8674(00)81085-5

Goulet, B., Baruch, A., Moon, N.-S., Poirier, M., Sansregret, L. L., Erickson, A., et al. (2004). A cathepsin L isoform that is devoid of a signal peptide localizes to the nucleus in S phase and processes the CDP/Cux transcription factor. *Mol. Cell* 14, 207–219. doi: 10.1016/s1097-2765(04)00209-6

Graveland, G., and DiFiglia, M. (1985). The frequency and distribution of medium-sized neurons with indented nuclei in the primate and rodent neostriatum. *Brain Res.* 327, 307–311. doi: 10.1016/0006-8993(85)91524-0

Gray, L. T., Yao, Z., Nguyen, T. N., Kim, T. K., Zeng, H., and Tasic, B. (2017). Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. *Elife* 6:e2188. doi: 10.7554/eLife.21883

Grigoriou, M., Tucker, A. S., Sharpe, P. T., and Pachnis, V. (1998). Expression and regulation of Lhx6 and Lhx7, a novel subfamily of LIM homeodomain encoding genes, suggests a role in mammalian head development. *Development* 125, 2063–2074. doi: 10.1242/dev.125.11.2063

Grueber, W. B., Jan, L. Y., and Jan, Y. N. (2003). Different levels of the homeodomain protein cut regulate distinct dendrite branching patterns of Drosophila multidendritic neurons. *Cell* 112, 805–818. doi: 10.1016/s0092-8674(03)00160-0

Guazzi, S., Price, M., De Felice, M., Damante, G., Mattei, M.-G., and Di Lauro, R. (1990). Thyroid nuclear factor 1 (TTF-1) contains a homeodomain and displays a novel DNA binding specificity. *EMBO J.* 9, 3631–3639. doi: 10.1002/j.1460-2075.1990.tb07574.x

Gulisano, M., Broccoli, V., Pardini, C., and Boncinelli, E. (1996). Emx1 and Emx2 show different patterns of expression during proliferation and differentiation of the developing cerebral cortex in the mouse. *Eur. J. Neurosci.* 8, 1037–1050. doi: 10.1111/j.1460-9568.1996.tb01590.x

Gehring, W. J. (1996). The master control gene for morphogenesis and evolution of the eye. *Genes Cells* 1, 11–15. doi: 10.1046/j.1365-2443.1996.11011.x

Hack, M., Sugimori, M., Lundberg, C., Nakafuku, M., and Götz, M. (2004). Regionalization and fate specification in neurospheres: the role of Olig2 and Pax6. *Mol. Cell. Neurosci.* 25, 664–678. doi: 10.1016/j.mcn.2003.12.012

Hamasaki, T., Leingärtner, A., Ringstedt, T., and O'Leary, D. D. (2004). EMX2 regulates sizes and positioning of the primary sensory and motor areas in neocortex by direct specification of cortical progenitors. *Neuron* 43, 359–372. doi: 10.1016/j.neuron.2004.07.016

Hamilton, S. P., Woo, J. M., Carlson, E. J., Ghanem, N., Ekker, M., and Rubenstein, J. L. R. (2005). Analysis of four DLX homeobox genes in autistic probands. *BMC Genet.* 6:52. doi: 10.1186/1471-2156-6-52

Hanashima, C., Shen, L., Li, S. C., and Lai, E. (2002). Brain factor-1 controls the proliferation and differentiation of neocortical progenitor cells through independent mechanisms. *J. Neurosci.* 22, 6526–6536. doi: 10.1523/JNEUROSCI.22-15-06526.2002

Hansen, D. V., Lui, J. H., Flandin, P., Yoshikawa, K., Rubenstein, J. L., Alvarez-Buylla, A., et al. (2013). Non-epithelial stem cells and cortical interneuron production in the human ganglionic eminences. *Nat. Neurosci.* 16, 1576–1587. doi: 10.1038/nn.3541

Harada, R., Bérubé, G., Tamplin, O. J., Denis-Larose, C., and Nepveu, A. (1995). DNA-binding specificity of the cut repeats from the human cut-like protein. *Mol. Cell. Biol.* 15, 129–140. doi: 10.1128/MCB.15.1.129

Harada, R., Dufort, D., Denis-Larose, C., and Nepveu, A. (1994). Conserved cut repeats in the human cut homeodomain protein function as DNA binding domains. *J. Biol. Chem.* 269, 2062–2067. doi: 10.1016/s0021-9258(17)42135-1

Harada, R., Vadnais, C., Sansregret, L., Leduy, L., Bérubé, G., Robert, F., et al. (2007). Genome-wide location analysis and expression studies reveal a role for p110 CUX1 in the activation of DNA replication genes. *Nucleic Acids Res.* 36, 189–202. doi: 10.1093/nar/gkm970

Hashizume, K., Yamanaka, M., and Ueda, S. (2018). POU3F2 participates in cognitive function and adult hippocampal neurogenesis via mammalian-characteristic amino acid repeats. *Genes Brain Behav.* 17, 118–125. doi: 10.1111/gbb.12408

Haubensak, W., Attardo, A., Denk, W., and Huttner, W. B. (2004). Neurons arise in the basal neuroepithelium of the early mammalian telencephalon: a major site of neurogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 101, 3196–3201. doi: 10.1073/pnas.0308600100

Haubst, N., Berger, J., Radjendirane, V., Graw, J., Favor, J., Saunders, G. F., et al. (2004). Molecular dissection of Pax6 function: the specific roles of the paired domain and homeodomain in brain development. *Development* 131, 6131–6140. doi: 10.1242/dev.01524

Haydar, T. F., and Reeves, R. H. (2012). Trisomy 21 and early brain development. *Trends Neurosci.* 35, 81–91. doi: 10.1016/j.tins.2011.11.001

He, X., Treacy, M. N., Simmons, D. M., Ingraham, H. A., Swanson, L. W., and Rosenfeld, M. G. (1989). Expression of a large family of POU-domain regulatory genes in mammalian brain development. *Nature* 340, 35–42. doi: 10.1038/340035a0

Heide, M., Zhang, Y., Zhou, X., Zhao, T., Miquelajáuregui, A., Varela-Echavarría, A., et al. (2015). Lhx5 controls mamillary differentiation in the developing hypothalamus of the mouse. *Front. Neuroanat.* 9:113. doi: 10.3389/fnana.2015.00113

Heins, N., Malatesta, P., Cecconi, F., Nakafuku, M., Tucker, K. L., Hack, M. A., et al. (2002). Glial cells generate neurons: the role of the transcription factor Pax6. *Nat. Neurosci.* 5, 308–315. doi: 10.1038/nn828

Hendry, S. H., Schwark, H., Jones, E., and Yan, J. (1987). Numbers and proportions of GABA-immunoreactive neurons in different areas of monkey cerebral cortex. *J. Neurosci.* 7, 1503–1519. doi: 10.1523/JNEUROSCI.07-05-01503.1987

Herr, W., and Cleary, M. A. (1995). The POU domain: versatility in transcriptional regulation by a flexible two-in-one DNA-binding domain. *Genes Dev.* 9, 1679–1693. doi: 10.1101/gad.9.14.1679

Hill, R. E., Favor, J., Hogan, B. L., Ton, C. C., Saunders, G. F., Hanson, I. M., et al. (1991). Mouse small eye results from mutations in a paired-like homeobox-containing gene. *Nature* 354, 522–525. doi: 10.1038/354522a0

Hirata, T., Nomura, T., Takagi, Y., Sato, Y., Tomioka, N., Fujisawa, H., et al. (2002). Mosaic development of the olfactory cortex with Pax6-dependent and-independent components. *Dev. Brain Res.* 136, 17–26. doi: 10.1016/s0165-3806(02)00304-8

Hoch, R. V., Lindtner, S., Price, J. D., and Rubenstein, J. L. R. (2015). OTX2 transcription factor controls regional patterning within the medial ganglionic eminence and regional identity of the septum. *Cell Rep.* 12, 482–494. doi: 10.1016/j.celrep.2015.06.043

Holder, J. L. Jr., Zhang, L., Kublaoui, B. M., DiLeone, R. J., Oz, O. K., Bair, C. H., et al. (2004). Sim1 gene dosage modulates the homeostatic feeding response to increased dietary fat in mice. *Am. J. Physiol. Endocrinol. Metab.* 287, E105–E113. doi: 10.1152/ajpendo.00446.2003

Holland, P. W., Booth, H. A. F., and Bruford, E. A. (2007). Classification and nomenclature of all human homeobox genes. *BMC Biol.* 5:47. doi: 10.1186/1741-7007-5-47

Horike, S.-i, Cai, S., Miyano, M., Cheng, J.-F., and Kohwi-Shigematsu, T. (2005). Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat. Genet.* 37, 31–40. doi: 10.1038/ng1491

Hou, P.-S., Miyoshi, G., and Hanashima, C. (2019). Sensory cortex wiring requires preselection of short-and long-range projection neurons through an Egr-Foxg1-COUP-TFI network. *Nat. Commun.* 10, 1–18. doi: 10.1038/s41467-019-11043-w

Hsieh-Li, H. M., Witte, D. P., Szucsik, J. C., Weinstein, M., Li, H., and Potter, S. S. (1995). Gsh-2, a murine homeobox gene expressed in the developing brain. *Mech. Dev.* 50, 177–186. doi: 10.1016/0925-4773(94)00334-j

Huang, B., Li, X., Tu, X., Zhao, W., Zhu, D., Feng, Y., et al. (2018). OTX1 regulates cell cycle progression of neural progenitors in the developing cerebral cortex. *J. Biol. Chem.* 293, 2137–2148. doi: 10.1074/jbc.RA117.001249

Ide, M., Yamada, K., Toyota, T., Iwayama, Y., Ishitsuka, Y., Minabe, Y., et al. (2005). Genetic association analyses of PHOX2B and ASCL1 in neuropsychiatric disorders: evidence for association of ASCL1 with Parkinson's disease. *Hum. Genet.* 117, 520–527. doi: 10.1007/s00439-005-1342-8

Iler, N., Rowitch, D. H., Echelard, Y., McMahon, A. P., and Abate-Shen, C. (1995). A single homeodomain binding site restricts spatial expression of Wnt-1 in the developing brain. *Mech. Dev.* 53, 87–96. doi: 10.1016/0925-4773(95)00427-0

Imamura, F., Ayoub, A. E., Rakic, P., and Greer, C. A. (2011). Timing of neurogenesis is a determinant of olfactory circuitry. *Nat. Neurosci.* 14, 331–337. doi: 10.1038/nn.2754

Imamura, F., and Greer, C. A. (2013). Pax6 regulates Tbr1 and Tbr2 expressions in olfactory bulb mitral cells. *Mol. Cell Neurosci.* 54, 58–70. doi: 10.1016/j.mcn.2013.01.002

Inada, H., Watanabe, M., Uchida, T., Ishibashi, H., Wake, H., Nemoto, T., et al. (2011). GABA regulates the multidirectional tangential migration of GABAergic interneurons in living neonatal mice. *PLoS One* 6:e27048. doi: 10.1371/journal.pone.0027048

Inoue, T., Chisaka, O., Matsunami, H., and Takeichi, M. (1997). Cadherin-6 expression transiently delineates specific rhombomeres, other neural tube subdivisions, and neural crest subpopulations in mouse embryos. *Dev. Biol.* 183, 183–194. doi: 10.1006/dbio.1996.8501

Inoue, T., Nakamura, S., and Osumi, N. (2000). Fate mapping of the mouse prosencephalic neural plate. *Dev. Biol.* 219, 373–383. doi: 10.1006/dbio.2000.9616

Inoue, F., Kurokawa, D., Takahashi, M., and Aizawa, S. (2012). Gbx2 directly restricts Otx2 expression to forebrain and midbrain, competing with class III POU factors. *Mol. Cell. Biol.* 32, 2618–2627. doi: 10.1128/MCB.00083-12

Iulianella, A., Sharma, M., Durnin, M., Vanden Heuvel, G. B., and Trainor, P. A. (2008). Cux2 (Cutl2) integrates neural progenitor development with cell-cycle progression during spinal cord neurogenesis. *Development* 135, 729–741. doi: 10.1242/dev.013276

Iulianella, A., Vanden Heuvel, G., and Trainor, P. (2003). Dynamic expression of murine Cux2 in craniofacial, limb, urogenital and neuronal primordia. *Gene Expr. Patterns* 3:571–577. doi: 10.1016/s1567-133x(03)00123-6

Jennings, B. H., Pickles, L. M., Wainwright, S. M., Roe, S. M., Pearl, L. H., and Ish-Horowicz, D. (2006). Molecular recognition of transcriptional repressor motifs by the WD domain of the Groucho/TLE corepressor. *Mol. Cell* 22, 645–655. doi: 10.1016/j.molcel.2006.04.024

Jessell, T.M. (2000). Neuronal specification in the spinal cord: inductive signals and transcriptional codes. *Nat. Rev. Genet.* 1, 20–29. doi: 10.1038/35049541

Jiang, Q., Zagozewski, J., Godbout, R., and Eisenstat, D. D. (2020). Distal-less genes Dlx1/Dlx2 repress oligodendrocyte genesis through transcriptional inhibition of Olig2 expression in the developing vertebrate forebrain. *bioRxiv* [Preprint].

Jimenez, F., Martin-Morris, L. E., Velasco, L., Chu, H., Sierra, J., Rosen, D., et al. (1995). vnd, a gene required for early neurogenesis of Drosophila, encodes a homeodomain protein. *EMBO J.* 14, 3487–3495. doi: 10.1002/j.1460-2075.1995.tb07355.x

Jimenez-García, M. P., Lucena-Cacace, A., Otero-Albiol, D., and Carnero, A. (2021a). Empty spiracles homeobox genes EMX1 and EMX2 regulate WNT pathway activation in sarcomagenesis. *J. Exp. Clin. Cancer Res.* 40, 1–18. doi: 10.1186/s13046-021-02048-9

Jimenez-García, M. P., Lucena-Cacace, A., Otero-Albiol, D., and Carnero, A. (2021b). Regulation of sarcomagenesis by the empty spiracles homeobox genes EMX1 and EMX2. *Cell Death Dis.* 12, 1–17. doi: 10.1038/s41419-021-03801-w

Johnson, J. E., Birren, S. J., Saito, T., and Anderson, D. J. (1992). DNA binding and transcriptional regulatory activity of mammalian achaete-scute homologous (MASH) proteins revealed by interaction with a muscle-specific enhancer. *Proc. Natl. Acad. Sci.* 89:3596. doi: 10.1073/pnas.89.8.3596

Joliot, A., and Prochiantz, A. (2004). Transduction peptides: from technology to physiology. *Nat. Cell Biol.* 6, 189–196. doi: 10.1038/ncb0304-189

Jones, L., ópez-Bendito, G. L., Gruss, P., Stoykova, A., and Molnár, Z. (2002). Pax6 is required for the normal development of the forebrain axonal connections. *Development* 129, 5041–5052. doi: 10.1242/dev.129.21.5041

Kanatani, S., Honda, T., Aramaki, M., Hayashi, K., Kubo, K.-i., Ishida, M., et al. (2015). The COUP-TFII/Neuropilin-2 is a molecular switch steering diencephalon-derived GABAergic neurons in the developing mouse brain. *Proc. Natl. Acad. Sci.* 112, E4985–E4994. doi: 10.1073/pnas.1420701112

Karlsson, O., Thor, S., Norberg, T., Ohlsson, H., and Edlund, T. (1990). Insulin gene enhancer binding protein Isl-1 is a member of a novel class of proteins containing both a homeo-and a Cys–His domain. *Nature* 344, 879–882. doi: 10.1038/344879a0

Kappen, C. (2000). The homeodomain: an ancient evolutionary motif in animals and plants. *Comput. Chem.* 24, 95–103. doi: 10.1016/s0097-8485(99)00049-2

Kastury, K., Druck, T., Huebner, K., Barletta, C., Acampora, D., Simeone, A., et al. (1994). Chromosome locations of human EMX and OTX genes. *Genomics* 22, 41–45. doi: 10.1006/geno.1994.1343

Kato, M., and Dobyns, W. B. (2004). X-linked lissencephaly with abnormal genitalia as a tangential migration disorder causing intractable epilepsy: proposal for a new term, "interneuronopathy". *J. Child Neurol.* 19, 392–397. doi: 10.1177/088307380502000420 01

Kato, M., Das, S., Petras, K., Kitamura, K., Morohashi, K. i., Abuelo, D. N., et al. (2004). Mutations of ARX are associated with striking pleiotropy and consistent genotype–phenotype correlation. *Hum. Mutat.* 23, 147–155. doi: 10.1002/humu.10310

Kato, M., Das, S., Petras, K., Sawaishi, Y., and Dobyns, W. B. (2003). Polyalanine expansion of ARX associated with cryptogenic West syndrome. *Neurology* 61, 267–277. doi: 10.1212/01.wnl.0000068012.69928.92

Kato, M., Saitoh, S., Kamei, A., Shiraishi, H., Ueda, Y., Akasaka, M., et al. (2007). A longer polyalanine expansion mutation in the ARX gene causes early infantile epileptic encephalopathy with suppression-burst pattern (Ohtahara syndrome). *Am. J. Hum. Genet.* 81, 361–366. doi: 10.1086/518903

Kaur, R., Aiken, C., Morrison, L. C., Rao, R., Del Bigio, M. R., Rampalli, S., et al. (2015). OTX2 exhibits cell-context-dependent effects on cellular and molecular properties of human embryonic neural precursors and medulloblastoma cells. *Dis. Models Mech.* 8, 1295–1309. doi: 10.1242/dmm.020594

Kessaris, N., Fogarty, M., Iannarelli, P., Grist, M., Wegner, M., and Richardson, W. D. (2006). Competing waves of oligodendrocytes in the forebrain and postnatal elimination of an embryonic lineage. *Nat. Neurosci.* 9, 173–179. doi: 10.1038/nn1620

Kiecker, C., and Lumsden, A. (2004). Hedgehog signaling from the ZLI regulates diencephalic regional identity. *Nat. Neurosci.* 7, 1242–1249. doi: 10.1038/nn1338

Kiecker, C., and Lumsden, A. (2005). Compartments and their boundaries in vertebrate brain development. *Nat. Rev. Neurosci.* 6, 553–564. doi: 10.1038/nrn1702

Kim, Y., and Nirenberg, M. (1989). Drosophila NK-homeobox genes. *Proc. Natl. Acad. Sci.* 86, 7716–7720. doi: 10.1073/pnas.86.20.7716

Kim, D. W., Liu, K., Wang, Z. Q., Zhang, Y. S., Bathini, A., Brown, M. P., et al. (2021). Gene regulatory networks controlling differentiation, survival, and diversification of hypothalamic Lhx6-expressing GABAergic neurons. *Commun. Biol.* 4, 1–16. doi: 10.1038/s42003-020-01616-7

Kim, D. W., Washington, P. W., Wang, Z. Q., Lin, S. H., Sun, C., Ismail, B. T., et al. (2020). The cellular and molecular landscape of hypothalamic patterning and differentiation from embryonic to late postnatal development. *Nat. Commun.* 11:4360.

Kimura, S., Hara, Y., Pineau, T., Fernandez-Salguero, P., Fox, C. H., Ward, J. M., et al. (1996). The T/ebp null mouse: thyroid-specific enhancer-binding protein is essential for the organogenesis of the thyroid, lung, ventral forebrain, and pituitary. *Genes Dev.* 10, 60–69. doi: 10.1101/gad.10.1.60

Kitamura, K., Itou, Y., Yanazawa, M., Ohsawa, M., Suzuki-Migishima, R., Umeki, Y., et al. (2009). Three human ARX mutations cause the lissencephaly-like and mental retardation with epilepsy-like pleiotropic phenotypes in mice. *Hum. Mol. Genet.* 18, 3708–3802. doi: 10.1093/hmg/ddp318

Kitamura, K., Miura, H., Yanazawa, M., Miyashita, T., and Kato, K. (1997). Expression patterns of Brx1 (Rieg gene), Sonic hedgehog, Nkx2. 2, Dlx1 and Arx during zona limitans intrathalamica and embryonic ventral lateral geniculate nuclear formation. *Mech. Dev.* 67, 83–96. doi: 10.1016/s0925-4773(97)00110-x

Kitamura, K., Yanazawa, M., Sugiyama, N., Miura, H., Iizuka-Kogo, A., Kusaka, M., et al. (2002). Mutation of ARX causes abnormal development of forebrain and testes in mice and X-linked lissencephaly with abnormal genitalia in humans. *Nat. Genet.* 32, 359–369. doi: 10.1038/ng1009

Klemm, J. D., Rould, M. A., Aurora, R., Herr, W., and Pabo, C. O. (1994). Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA

recognition with tethered DNA-binding modules. *Cell* 77, 21–32. doi: 10.1016/0092-8674(94)90231-3

Ko, J., Humbert, S., Bronson, R. T., Takahashi, S., Kulkarni, A. B., Li, E., et al. (2001). p35 and p39 are essential for cyclin-dependent kinase 5 function during neurodevelopment. *J. Neurosci.* 21, 6758–71. doi: 10.1523/JNEUROSCI.21-17-06758.2001

Kobayashi, D., Kobayashi, M., Matsumoto, K., Ogura, T., Nakafuku, M., and Shimamura, K. (2002). Early subdivisions in the neural plate define distinct competence for inductive signals. *Development* 129, 83–93. doi: 10.1242/dev.129.1.83

Kriegstein, A., and Alvarez-Buylla, A. (2009). The glial nature of embryonic and adult neural stem cells. *Annu. Rev. Neurosci.* 32, 149–184. doi: 10.1146/annurev.neuro.051508.135600

Kroll, T. T., and O'Leary, D. D. (2005). Ventralized dorsal telencephalic progenitors in Pax6 mutant mice generate GABA interneurons of a lateral ganglionic eminence fate. *Proc. Natl. Acad. Sci.* 102, 7374–7379. doi: 10.1073/pnas.0500819102

Krude, H., Schütz, B., Biebermann, H., Von Moers, A., Schnabel, D., Neitzel, H., et al. (2002). Choreoathetosis, hypothyroidism, and pulmonary alterations due to human NKX2-1 haploinsufficiency. *J. Clin. Invest.* 109, 475–480. doi: 10.1172/JCI14341

Kumamoto, T., and Hanashima, C. (2017). Evolutionary conservation and conversion of Foxg1 function in brain development. *Dev. Growth Differ.* 59, 258–269. doi: 10.1111/dgd.12367

Kurokawa, D., Kiyonari, H., Nakayama, R., Kimura-Yoshida, C., Matsuo, I., and Aizawa, S. (2004). Regulation of Otx2 expression and its functions in mouse forebrain and midbrain. *Development* 131, 3319–3331. doi: 10.1242/dev.01220

Kurokawa, D., Sakurai, Y., Inoue, A., Nakayama, R., Takasaki, N., Suda, Y., et al. (2006). Evolutionary constraint on Otx2 neuroectoderm enhancers-deep conservation from skate to mouse and unique divergence in teleost. *Proc. Natl. Acad. Sci.* 103, 19350–19355. doi: 10.1073/pnas.0604686103

Küspert, M., Hammer, A., Bösl, M. R., and Wegner, M. (2011). Olig2 regulates Sox10 expression in oligodendrocyte precursors through an evolutionary conserved distal enhancer. *Nucleic Acids Res.* 39, 1280–1293. doi: 10.1093/nar/gkq951

Larsen, K. B., Lutterodt, M. C., Laursen, H., Graem, N., Pakkenberg, B., Møllgård, K., et al. (2010a). Spatiotemporal distribution of PAX6 and MEIS2 expression and total cell numbers in the ganglionic eminence in the early developing human forebrain. *Dev. Neurosci.* 32, 149–162. doi: 10.1159/000297602

Larsen, K. B., Lutterodt, M. C., Møllgård, K., and Møller, M. (2010b). Expression of the homeobox genes OTX2 and OTX1 in the early developing human brain. *J. Histochem. Cytochem.* 58, 669–678. doi: 10.1369/jhc.2010.955757

Lavdas, A. A., Grigoriou, M., Pachnis, V., and Parnavelas, J. G. (1999). The medial ganglionic eminence gives rise to a population of early neurons in the developing cerebral cortex. *J. Neurosci.* 19, 7881–7888. doi: 10.1523/JNEUROSCI.19-18-07881.1999

Le, T. N., Zhou, Q.-P., Cobos, I., Zhang, S., Zagozewski, J., Japoni, S., et al. (2017). GABAergic interneuron differentiation in the basal forebrain is mediated through direct regulation of glutamic acid decarboxylase isoforms by dlx homeobox transcription factors. *J. Neurosci.* 37, 8816–8829. doi: 10.1523/JNEUROSCI.2125-16.2017

Lee, J. E. (1997). Basic helix-loop-helix genes in neural development. *Curr. Opin. Neurobiol.* 7, 13–20. doi: 10.1016/s0959-4388(97)80115-8

Lee, H. H. C., Bernard, C., Ye, Z., Acampora, D., Simeone, A., Prochiantz, A., et al. (2017). Genetic Otx2 mis-localization delays critical period plasticity across brain regions. *Mol. Psychiatry* 22, 680–688. doi: 10.1038/mp.2017.1

Lesaffre, B., Joliot, A., Prochiantz, A., and Volovitch, M. (2007). Direct non-cell autonomous Pax6 activity regulates eye development in the zebrafish. *Neural Dev.* 2:2. doi: 10.1186/1749-8104-2-2

Li, H., Bishop, K. M., and O'Leary, D. D. (2006). Potential target genes of EMX2 include Odz/Ten-M and other gene families with implications for cortical patterning. *Mol. Cell. Neurosci.* 33, 136–149. doi: 10.1016/j.mcn.2006.06.012

Li, H., Hader, A. T., Han, Y. R., Wong, J. A., Babiarz, J., Ricupero, C. L., et al. (2012a). Isolation of a novel rat neural progenitor clone that expresses dlx family transcription factors and gives rise to functional gabaergic neurons in culture. *Dev. Neurobiol.* 72, 805–820. doi: 10.1002/dneu.20977

Li, H., Lu, Y., Smith, H. K., and Richardson, W. D. (2007). Olig1 and Sox10 interact synergistically to drive myelin basic protein transcription in oligodendrocytes. *J. Neurosci.* 27, 14375–14382. doi: 10.1523/JNEUROSCI.4456-07.2007

Li, H., Miao, Q., Xu, C.-w., Huang, J.-h., Zhou, Y.-f., and Wu, M.-j. (2016). OTX1 contributes to hepatocellular carcinoma progression by regulation of ERK/MAPK pathway. *J. Korean Med. Sci.* 31, 1215–1223. doi: 10.3346/jkms.2016.31.8.1215

Li, J., Mo, M., Chen, Z., Chen, Z., Sheng, Q., Mu, H., et al. (2012b). Adenoviral delivery of the EMX2 gene suppresses growth in human gastric cancer. *PLoS One* 7:e45970. doi: 10.1371/journal.pone.0045970

Li, J., Wang, C., Zhang, Z., Wen, Y., An, L., Liang, Q., et al. (2017). Transcription factors Sp8 and Sp9 coordinately regulate olfactory bulb interneuron development. *Cereb. Cortex* 28, 3278–3294. doi: 10.1093/cercor/bhx199

Li, M., Santpere, G., Kawasawa, Y. I., Evgrafov, O. V., Gulden, F. O., Pochareddy, S., et al. (2018). Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* 362:eaat7615. doi: 10.1126/science.aat7615

Li, N., Zhao, C.-T., Wang, Y., and Yuan, X.-B. (2010). The transcription factor Cux1 regulates dendritic morphology of cortical pyramidal neurons. *PLoS One* 5:e1059. doi: 10.1371/journal.pone.0010596

Li, T., Stark, M. R., Johnson, A. D., and Wolberger, C. (1995). Crystal structure of the MATa1/MATα2 homeodomain heterodimer bound to DNA. *Science* 270, 262–269. doi: 10.1126/science.270.5234.262

Li, Y., Kido, T., Garcia-Barcelo, M. M., Tam, P. K., Tabatabai, Z. L., and Lau, Y.-F. C. (2015). SRY interference of normal regulation of the RET gene suggests a potential role of the Y-chromosome gene in sexual dimorphism in Hirschsprung disease. *Hum. Mol. Genet.* 24, 685–697. doi: 10.1093/hmg/ddu488

Ligon, K. L., Echelard, Y., Assimacopoulos, S., Danielian, P. S., Kaing, S., Grove, E. A., et al. (2003). Loss of Emx2 function leads to ectopic expression of Wnt1 in the developing telencephalon and cortical dysplasia. *Development* 130, 2275–2287. doi: 10.1242/dev.00421

Lim, J. W., Donahoo, A.-L. S., Bunt, J., Edwards, T. J., Fenlon, L. R., Liu, Y., et al. (2015). EMX1 regulates NRP1-mediated wiring of the mouse anterior cingulate cortex. *Development* 142, 3746–3757. doi: 10.1242/dev.119909

Lim, Y., Cho, I.-T., Shi, X., Grinspan, J. B., Cho, G., and Golden, J. A. (2019). Arx expression suppresses ventralization of the developing dorsal forebrain. *Sci. Rep.* **9**:226. doi: 10.1038/s41598-018-36194-6

Lindtner, S., Catta-Preta, R., Tian, H., Su-Feher, L., Price, J. D., Dickel, D. E., et al. (2019). Genomic resolution of DLX-orchestrated transcriptional circuits driving development of forebrain GABAergic neurons. *Cell Rep.* 28, 2048.e8–2063.e8. doi: 10.1016/j.celrep.2019.07.022

Liodis, P., Denaxa, M., Grigoriou, M., Akufo-Addo, C., Yanagawa, Y., and Pachnis, V. (2007). Lhx6 activity is required for the normal migration and specification of cortical interneuron subtypes. *J. Neurosci.* 27, 3078–3089. doi: 10.1523/JNEUROSCI.3055-06.2007

Liu, J. K., Ghattas, I., Liu, S., Chen, S., and Rubenstein, J. L. R. (1997). Dlx genes encode DNA-binding proteins that are expressed in an overlapping and sequential pattern during basal ganglia differentiation. *Dev. Dyn.* 210, 498–512. doi: 10.1002/(SICI)1097-0177(199712)210:4<498::AID-AJA12>3.0.CO;2-3

Liu, Y., MacDonald, R. J., and Swift, G. H. (2001). DNA binding and transcriptional activation by a PDX1.PBX1b.MEIS2b trimer and cooperation with a pancreas-specific basic helix-loop-helix complex. *J. Biol. Chem.* 276, 17985–17993. doi: 10.1074/jbc.M100678200

Leonard, J. H., and Bell, J. R. (1997). Insights into the Merkel cell phenotype from Merkel cell carcinoma cell lines. *Australas J. Dermatol.* 38 (Suppl 1), S91–S98. doi: 10.1111/j.1440-0960.1997.tb01019.x

Letinic, K., Zoncu, R., and Rakic, P. (2002). Origin of GABAergic neurons in the human neocortex. *Nature* 417, 645–649. doi: 10.1038/nature00779

Lois, C., and Alvarez-Buylla, A. (1994). Long-distance neuronal migration in the adult mammalian brain. *Science* 264, 1145–1148. doi: 10.1126/science.8178174

Long, J. E., Cobos, I., Potter, G. B., and Rubenstein, J. L. R. (2009). Dlx1&2 and Mash1 transcription factors control MGE and CGE patterning and differentiation through parallel and overlapping pathways. *Cereb. Cortex* 19(Suppl. 1), i96–i106.

Long, J. E., Garel, S., Alvarez-Dolado, M., Yoshikawa, K., Osumi, N., Alvarez-Buylla, A., et al. (2007). Dlx-dependent and -independent regulation of olfactory bulb interneuron differentiation. *J. Neurosci.* 27:3230. doi: 10.1523/JNEUROSCI.5265-06.2007

Longobardi, E., Penkov, D., Mateos, D., De Florian, G., Torres, M., and Blasi, F. (2014). Biochemistry of the tale transcription factors PREP, MEIS, and PBX in vertebrates. *Dev. Dyn.* 243, 59–75. doi: 10.1002/dvdy.24016

LoTurco, J. J., and Bai, J. (2006). The multipolar stage and disruptions in neuronal migration. *Trends Neurosci.* 29, 407–413. doi: 10.1016/j.tins.2006.05.006

López-Bendito, G., Sánchez-Alcaniz, J. A., Pla, R., Borrell, V., Picó, E., Valdeolmillos, M., et al. (2008). Chemokine signaling controls intracortical migration and final distribution of GABAergic interneurons. *J. Neurosci.* 28, 1613–1624. doi: 10.1523/JNEUROSCI.4651-07.2008

Lopes, R., van Wijk, N. V., Neves, G., and Pachnis, V. (2012). Transcription factor LIM homeobox 7 (Lhx7) maintains subtype identity of cholinergic interneurons in the mammalian striatum. *Proc. Natl. Acad. Sci.* 109, 3119–3124. doi: 10.1073/pnas.1109251109

Lowenstein, E.D., Rusanova, A., Stelzer, J., Hernaiz-Llorens, M., Schroer, A. E., Epifanova, E., et al. (2021). Olig3 regulates early cerebellar development. *Elife* 10:e64684. doi: 10.7554/eLife.64684

Lu, Q. R., Sun, T., Zhu, Z., Ma, N., Garcia, M., Stiles, C. D., et al. (2002). Common developmental requirement for Olig function indicates a motor neuron/oligodendrocyte connection. *Cell* 109, 75–86. doi: 10.1016/s0092-8674(02)00678-5

Lu, Q. R., Yuk, D.-i., Alberta, J. A., Zhu, Z., Pawlitzky, I., Chan, J., et al. (2000). Sonic hedgehog–regulated oligodendrocyte lineage genes encoding bHLH proteins in the mammalian central nervous system. *Neuron* 25, 317–329. doi: 10.1016/s0896-6273(00)80897-1

Lu, Y., Labak, C. M., Jain, N., Purvis, I. J., Guda, M. R., Bach, S. E., et al. (2017). OTX2 expression contributes to proliferation and progression in Myc-amplified medulloblastoma. *Am. J. Cancer Res.* 7:647.

Ma, D. K., Jang, M.-H., Guo, J. U., Kitabatake, Y., Chang, M.-l., Pow-Anpongkul, N., et al. (2009). Neuronal activity–induced Gadd45b promotes epigenetic DNA demethylation and adult neurogenesis. *Science* 323, 1074–1077. doi: 10.1126/science.1166859

Ma, T., Wang, C., Wang, L., Zhou, X., Tian, M., Zhang, Q., et al. (2013). Subcortical origins of human and monkey neocortical interneurons. *Nat. Neurosci.* 16, 1588–1597. doi: 10.1038/nn.3536

MacDonald, R. B., Debiais-Thibaud, M., Talbot, J. C., and Ekker, M. (2010). The relationship between dlx and gad1 expression indicates highly conserved genetic pathways in the zebrafish forebrain. *Dev. Dyn.* 239, 2298–2306. doi: 10.1002/dvdy.22365

MacDonald, R. B., Pollack, J. N., Debiais-Thibaud, M., Heude, E., Coffin Talbot, J., and Ekker, M. (2013). The ascl1a and dlx genes have a regulatory role in the development of GABAergic interneurons in the zebrafish diencephalon. *Dev. Biol.* 381, 276–285. doi: 10.1016/j.ydbio.2013.05.025

Magno, L., Barry, C., Schmidt-Hieber, C., Theodotou, P., Häusser, M., and Kessaris, N. (2017). NKX2-1 is required in the embryonic septum for cholinergic system development, learning, and memory. *Cell Rep.* 20, 1572–1584. doi: 10.1016/j.celrep.2017.07.053

Mailly, F., Bérubé, G., Harada, R., Mao, P. L., Phillips, S., and Nepveu, A. (1996). The human cut homeodomain protein can repress gene expression by two distinct mechanisms: active repression and competition for binding site occupancy. *Mol. Cell. Biol.* 16, 5346–5357. doi: 10.1128/MCB.16.10.5346

Maira, M., Long, J. E., Lee, A. Y., Rubenstein, J. L. R., and Stifani, S. (2010). Role for TGF-beta superfamily signaling in telencephalic GABAergic neuron development. *J. Neurodev. Disord.* 2, 48–60. doi: 10.1007/s11689-009-9035-6

Malandrini, A., Mari, F., Palmeri, S., Gambelli, S., Berti, G., Bruttini, M., et al. (2001). PAX6 mutation in a family with aniridia, congenital ptosis, and mental retardation. *Clin. Genet.* 60, 151–154. doi: 10.1034/j.1399-0004.2001.600210.x

Mallamaci, A., Iannone, R., Briata, P., Pintonello, L., Mercurio, S., Boncinelli, E., et al. (1998). EMX2 protein in the developing mouse brain and olfactory area. *Mech. Dev.* 77, 165–172. doi: 10.1016/s0925-4773(98)00141-5

Mallamaci, A., Mercurio, S., Muzio, L., Cecchi, C., Pardini, C. L., Gruss, P., et al. (2000). The lack of Emx2 causes impairment of reel in signaling and defects of neuronal migration in the developing cerebral cortex. *J. Neurosci.* 20, 1109–1118. doi: 10.1523/JNEUROSCI.20-03-01109.2000

Malt, E. A., Juhasz, K., Malt, U. F., and Naumann, T. (2016). A role for the transcription factor Nk2 homeobox 1 in schizophrenia: convergent evidence from animal and human studies. *Front. Behav. Neurosci.* 10:59. doi: 10.3389/fnbeh.2016.00059

Mariani, J., Coppola, G., Zhang, P., Abyzov, A., Provini, L., Tomasini, L., et al. (2015). FOXG1-dependent dysregulation of GABA/glutamate neuron differentiation in autism spectrum disorders. Cell 162, 375–390. doi: 10.1016/j.cell.2015.06.034

Mariani, J., Favaro, R., Lancini, C., Vaccari, G., Ferri, A. L., Bertolini, J., et al. (2012). Emx2 is a dose-dependent negative regulator of Sox2 telencephalic enhancers. *Nucleic Acids Res.* 40, 6461–6476. doi: 10.1093/nar/gks295

Marín, O., and Rubenstein, J. L. (2003). Cell migration in the forebrain. *Annu. Rev. Neurosci.* 26, 441–483. doi: 10.1146/annurev.neuro.26.041002.131058

Marín, O., Yaron, A., Bagri, A., Tessier-Lavigne, M., and Rubenstein, J. L. (2001). Sorting of striatal and cortical interneurons regulated by semaphorin-neuropilin interactions. *Science* 293, 872–875. doi: 10.1126/science.1061891

Marin, O., Anderson, S. A., and Rubenstein, J. L. (2000). Origin and molecular specification of striatal interneurons. *J. Neurosci.* 20, 6063–6076. doi: 10.1523/JNEUROSCI.20-16-06063.2000

Martin, D. L., Liu, H., Martin, S. B., and Wu, S. J. (2000). Structural features and regulatory properties of the brain glutamate decarboxylases. *Neurochem. Int.* 37, 111–119. doi: 10.1016/s0197-0186(00)00014-0

Martynoga, B., Morrison, H., Price, D. J., and Mason, J. O. (2005). Foxg1 is required for specification of ventral telencephalon and region-specific regulation of dorsal telencephalic precursor proliferation and apoptosis. *Dev. Biol.* 283, 113–127. doi: 10.1016/j.ydbio.2005.04.005

Marsh, E., Fulp, C., Gomez, E., Nasrallah, I., Minarcik, J., Sudi, J., et al. (2009). Targeted loss of Arx results in a developmental epilepsy mouse model and recapitulates the human phenotype in heterozygous females. *Brain* 132, 1563–1567. doi: 10.1093/brain/awp107

Marsh, E. D., Nasrallah, M. P., Walsh, C., Murray, K. A., Sunnen, C. N., McCoy, A., et al. (2016). Developmental interneuron subtype deficits after targeted loss of Arx. *BMC Neurosci.* 17:3. doi: 10.1186/s12868-016-0265-8

Mattiske, T., Tan, M. H., Dearsley, O., Cloosterman, D., Hii, C. S., Gécz, J., et al. (2018). Regulating transcriptional activity by phosphorylation: a new mechanism for the ARX homeodomain transcription factor. *PLoS One* 13:e0206914. doi: 10.1371/journal.pone.0206914

Matsumoto, K., Tanaka, T., Furuyama, T., Kashihara, Y., Mori, T., Ishii, N., et al. (1996). L3, a novel murine LIM-homeodomain transcription factor expressed in the ventral telencephalon and the mesenchyme surrounding the oral cavity. *Neurosci. Lett.* 204, 113–116. doi: 10.1016/0304-3940(96)12341-7

Matsunami, H., and Takeichi, M. (1995). Fetal brain subdivisions defined by R-and E-cadherin expressions: evidence for the role of cadherin activity in region-specific, cell–cell adhesion. *Dev. Biol.* 172, 466–478. doi: 10.1006/dbio.1995.8029

Matsuo, I., Kuratani, S., Kimura, C., Takeda, N., and Aizawa, S. (1995). Mouse Otx2 functions in the formation and patterning of rostral head. *Genes Dev.* 9, 2646–2658. doi: 10.1101/gad.9.21.2646

Manabe, T., Tatsumi, K., Inoue, M., Matsuyoshi, H., Makinodan, M., Yamauchi, T., et al. (2008). Knockdown of the L3/Lhx8 gene suppresses cholinergic differentiation of murine embryonic stem cell-derived spheres. *Int. J. Dev. Neurosci.* 26, 249–252. doi: 10.1016/j.ijdevneu.2007.11.004

Manabe, T., Tatsumi, K., Inoue, M., Matsuyoshi, H., Makinodan, M., Yokoyama, S., et al. (2005). L3/Lhx8 is involved in the determination of cholinergic or GABAergic cell fate. *J. Neurochem.* 94, 723–730. doi: 10.1111/j.1471-4159.2005.03261.x

Manabe, T., Tatsumi, K., Inoue, M., Makinodan, M., Yamauchi, T., Makinodan, E., et al. (2007). L3/Lhx8 is a pivotal factor for cholinergic differentiation of murine embryonic stem cells. *Cell Death Differ.* 14, 1080–1085. doi: 10.1038/sj.cdd.4402106

Manoli, M., and Driever, W. (2014). nkx2. 1 and nkx2. 4 genes function partially redundant during development of the zebrafish hypothalamus, preoptic region, and pallidum. *Front. Neuroanat.* 8:145. doi: 10.3389/fnana.2014.00145

Manuel, M., and Price, D. J. (2005). Role of Pax6 in forebrain regionalization. *Brain Res. Bull.* 66, 387–393. doi: 10.1016/j.brainresbull.2005.02.006

Maurel-Zaffran, C., and Treisman, J. E. (2000). pannier acts upstream of wingless to direct dorsal eye disc development in Drosophila. *Development* 127, 1007–1016. doi: 10.1242/dev.127.5.1007

Mayer, C., Hafemeister, C., Bandler, R. C., Machold, R., Brito, R. B., Jaglin, X., et al. (2018). Developmental diversification of cortical inhibitory interneurons. *Nature* 555, 457–462. doi: 10.1038/nature25999

McEvilly, R. J., de Diaz, M. O., Schonemann, M. D., Hooshmand, F., and Rosenfeld, M. G. (2002). Transcriptional regulation of cortical neuron migration by POU domain factors. *Science* 295, 1528–1532. doi: 10.1126/science.1067132

McKenzie, O., Ponte, I., Mangelsdorf, M., Finnis, M., Colasante, G., Shoubridge, C., et al. (2007). Aristaless-related homeobox gene, the gene responsible for West syndrome and related disorders, is a Groucho/transducin-like enhancer of split dependent transcriptional repressor. *Neuroscience* 146, 236–247. doi: 10.1016/j.neuroscience.2007.01.038

Medina, L., and Abellán, A. (2009). Development and evolution of the pallium. *Semin. Cell Dev. Biol.* 20, 698–711. doi: 10.1016/j.semcdb.2009.04.008

Méndez-Gómez, H. R. M., and Vicario-Abejón, C. (2012). The homeobox gene Gsx2 regulates the self-renewal and differentiation of neural stem cells and the cell fate of postnatal progenitors. *PLoS One* 7:e29799. doi: 10.1371/journal.pone.0029799

Merabet, S., and Lohmann, I. (2015). Toward a new twist in Hox and TALE DNA-binding specificity. *Dev. Cell* 32, 259–261. doi: 10.1016/j.devcel.2015.01.030

Metz, C. W. (1914). An apterous Drosophila and its genetic behavior. *Am. Nat.* 48, 675–692. doi: 10.1086/279439

Mi, D., Carr, C. B., Georgala, P. A., Huang, Y.-T., Manuel, M. N., Jeanes, E., et al. (2013). Pax6 exerts regional control of cortical progenitor proliferation via direct repression of Cdk6 and hypophosphorylation of pRb. *Neuron* 78, 269–284. doi: 10.1016/j.neuron.2013.02.012

Michaud, J. L., Rosenquist, T., May, N. R., and Fan, C. M. (1998). Development of neuroendocrine lineages requires the bHLH-PAS transcription factor SIM1. *Genes Dev.* 12, 3264–3275. doi: 10.1101/gad.12.20.3264

Michl, P., and Downward, J. (2006). CUTL1: a key mediator of TGFbeta-induced tumor invasion. *Cell Cycle* 5, 132–144. doi: 10.4161/cc.5.2.2311

Ming, G.-L., and Song, H. (2011). Adult neurogenesis in the mammalian brain: significant answers and significant questions. *Neuron* 70, 687–702. doi: 10.1016/j.neuron.2011.05.001

Minocha, S., Valloton, D., Arsenijevic, Y., Cardinaux, J.-R., Guidi, R., Hornung, J.-P., et al. (2017). Nkx2. 1 regulates the generation of telencephalic astrocytes during embryonic development. *Sci. Rep.* 7, 1–20. doi: 10.1038/srep43093

Minocha, S., Valloton, D., Ypsilanti, A. R., Fiumelli, H., Allen, E. A., Yanagawa, Y., et al. (2015). Nkx2. 1-derived astrocytes and neurons together with Slit2 are indispensable for anterior commissure formation. *Nat. Commun.* 6, 1–15. doi: 10.1038/ncomms7887

Miquelajáuregui, A., Varela-Echavarría, A., Ceci, M. L., García-Moreno, F., Ricano, I., Hoang, K., et al. (2010). LIM-homeobox gene Lhx5 is required for normal development of Cajal–Retzius cells. *J. Neurosci.* 30, 10551–10562. doi: 10.1523/JNEUROSCI.5563-09.2010

Miura, H., Yanazawa, M., Kato, K., and Kitamura, K. (1997). Expression of a novel aristaless related homeobox gene 'Arx'in the vertebrate telencephalon, diencephalon and floor plate. *Mech. Dev.* 65, 99–109. doi: 10.1016/s0925-4773(97)00062-2

Miyata, T., Kawaguchi, A., Okano, H., and Ogawa, M. (2001). Asymmetric inheritance of radial glial fibers by cortical neurons. *Neuron* 31, 727–741. doi: 10.1016/s0896-6273(01)00420-2

Miyoshi, G., Hjerling-Leffler, J., Karayannis, T., Sousa, V. H., Butt, S. J., Battiste, J., et al. (2010). Genetic fate mapping reveals that the caudal ganglionic eminence produces a large and diverse population of superficial cortical interneurons. *J. Neurosci.* 30, 1582–1594. doi: 10.1523/JNEUROSCI.4515-09.2010

Mizuno, K., Gonzalez, F. J., and Kimura, S. (1991). Thyroid-specific enhancer-binding protein (T/EBP): cDNA cloning, functional characterization, and structural identity with thyroid transcription factor TTF-1. *Mol. Cell. Biol.* 11, 4927–4933. doi: 10.1128/mcb.11.10.4927-4933.1991

Molotkova, N., Molotkov, A., and Duester, G. (2007). Role of retinoic acid during forebrain development begins late when Raldh3 generates retinoic acid in the ventral subventricular zone. *Dev. Biol.* 303, 601–610. doi: 10.1016/j.ydbio.2006.11.035

Moon, N. S., Bérubé, G., and Nepveu, A. (2000). CCAAT displacement activity involves CUT repeats 1 and 2, not the cut homeodomain*. *J. Biol. Chem.* 275, 31325–31334. doi: 10.1074/jbc.M002912200

Monnier, A., Boniface, R., Bouvet, R., Etcheverry, A., Aubry, M., Avril, T., et al. (2018). The expression of EMX2 lead to cell cycle arrest in glioblastoma cell line. *BMC Cancer* 18:1213. doi: 10.1186/s12885-018-5094-y

Monti, S., Nicoletti, A., Cantasano, A., Krude, H., and Cassio, A. (2015). NKX2. 1-Related Disorders: a novel mutation with mild clinical presentation. *Italian J. Pediatr.* 41, 1–5. doi: 10.1186/s13052-015-0150-6

Monuki, E. S., Porter, F. D., and Walsh, C. A. (2001). Patterning of the dorsal telencephalon and cerebral cortex by a roof plate-Lhx2 pathway. *Neuron* 32, 591–604. doi: 10.1016/s0896-6273(01)00504-9

Mortensen, A. H., Schade, V., Lamonerie, T., and Camper, S. A. (2015). Deletion of OTX2 in neural ectoderm delays anterior pituitary development. *Hum. Mol. Genet.* 24, 939–953. doi: 10.1093/hmg/ddu506

Mühleisen, T.W., Leber, M., Schulze, T.G., Strohmaier, J., Degenhardt, F., Treutlein, J., et al. (2014). Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat. Commun.* 5:3339. doi: 10.1038/ncomms4339

Murcia-Ramón, R., Company, V., Juárez-Leal, I., Andreu-Cervera, A., Almagro-García, F., Martínez, S., et al. (2020a). Neuronal tangential migration from Nkx2.1-positive hypothalamus. *Brain Struct. Funct.* 225, 2857–2869.

Murcia-Ramón, R., Juárez-Leal, I., Andreu-Cervera, A., Almagro-García, F., Martínez, S., Echevarría, D., et al. (2020b). Neuronal tangential migration from Nkx2. 1-positive hypothalamus. *Brain Struct. Funct.* 225, 2857–2869. doi: 10.1007/s00429-020-02163-x

Murre, C., McCaw, P. S., Vaessin, H., Caudy, M., Jan, L., Jan, Y., et al. (1989). Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. *Cell* 58, 537–544. doi: 10.1016/0092-8674(89)90434-0

Muzio, L., Di Benedetto, B., Stoykova, A., Boncinelli, E., Gruss, P., and Mallamaci, A. (2002a). Conversion of cerebral cortex into basal ganglia in Emx2−/− Pax6 Sey/Sey double-mutant mice. *Nat. Neurosci.* 5, 737–745. doi: 10.1038/nn892

Muzio, L., Di Benedetto, B., Stoykova, A., Boncinelli, E., Gruss, P., and Mallamaci, A. (2002b). Emx2 and Pax6 control regionalization of the pre-neuronogenic cortical primordium. *Cereb. Cortex* 12, 129–139. doi: 10.1093/cercor/12.2.12

Muzio, L., Soria, J., Pannese, M., Piccolo, S., and Mallamaci, A. (2005). A mutually stimulating loop involving emx2 and canonical wnt signalling specifically promotes expansion of occipital cortex and hippocampus. *Cereb. Cortex* 15, 2021–2028. doi: 10.1093/cercor/bhi077

Nadarajah, B., Brunstrom, J. E., Grutzendler, J., Wong, R. O., and Pearlman, A. L. (2001). Two modes of radial migration in early development of the cerebral cortex. *Nat. Neurosci.* 4, 143–150. doi: 10.1038/83967

Nagel, S., Ehrentraut, S., Meyer, C., Kaufmann, M., Drexler, H. G., and MacLeod, R. A. (2015). Aberrantly expressed OTX homeobox genes deregulate B-cell differentiation in Hodgkin lymphoma. *PLoS One* 10:e0138416. doi: 10.1371/journal.pone.0138416

Nakada, Y., Hunsaker, T. L., Henke, R. M., and Johnson, J. E. (2004). Distinct domains within Mash1 and Math1 are required for function in neuronal differentiation versus neuronal cell-type specification. *Development* 131, 1319–1330. doi: 10.1242/dev.01008

Nakagawa, Y., and Leary, D. D. M. (2001). Combinatorial expression patterns of LIM-homeodomain and other regulatory genes parcellate developing thalamus. *J. Neurosci.* 21:2711. doi: 10.1523/JNEUROSCI.21-08-02711.2001

Nakai, S., Kawano, H., Yudate, T., Nishi, M., Kuno, J., Nagata, A., et al. (1995). The POU domain transcription factor Brn-2 is required for the determination of specific neuronal lineages in the hypothalamus of the mouse. *Genes Dev.* 9, 3109–3121. doi: 10.1101/gad.9.24.3109

Nakamura, T., Jenkins, N. A., and Copeland, N. G. (1996). Identification of a new family of Pbx-related homeobox genes. *Oncogene* 13, 2235–2242.

Nasrallah, M. P., Cho, G., Simonet, J. C., Putt, M. E., Kitamura, K., and Golden, J. A. (2012). Differential effects of a polyalanine tract expansion in Arx on neural development and gene expression. *Hum. Mol. Genet.* 21, 1090–1098. doi: 10.1093/hmg/ddr538

NCBI Datasets (2021). *NCBI Datasets Gene*, Bethesda, MD: National Center for Biotechnology Information.

Nery, S., Fishell, G., and Corbin, J. G. (2002). The caudal ganglionic eminence is a source of distinct cortical and subcortical cell populations. *Nat. Neurosci.* 5, 1279–1287. doi: 10.1038/nn971

Neul, J. L., Kaufmann, W. E., Glaze, D. G., Christodoulou, J., Clarke, A. J., Bahi-Buisson, N., et al. (2010). Rett syndrome: revised diagnostic criteria and nomenclature. *Ann. Neurol.* 68, 944–950. doi: 10.1002/ana.22124

Nepveu, A. (2001). Role of the multifunctional CDP/Cut/Cux homeodomain transcription factor in regulating differentiation, cell growth and development. *Gene* 270, 1–15. doi: 10.1016/s0378-1119(01)00485-1

Neves, G., Shah, M. M., Liodis, P., Achimastou, A., Denaxa, M., Roalfe, G., et al. (2013). The LIM homeodomain protein Lhx6 regulates maturation of interneurons and network excitability in the mammalian cortex. *Cereb. Cortex* 23, 1811–1823. doi: 10.1093/cercor/bhs159

Neyt, C., Welch, M., Langston, A., Kohtz, J., Fishell, G., and Short-Range Signal, A. (1997). Restricts cell movement between telencephalic proliferative zones. *J. Neurosci.* 17:9194. doi: 10.1523/JNEUROSCI.17-23-09194.1997

Nieto, M., Monuki, E. S., Tang, H., Imitola, J., Haubst, N., Khoury, S. J., et al. (2004). Expression of Cux-1 and Cux-2 in the subventricular zone and upper layers II-IV of the cerebral cortex. *J. Comp. Neurol.* 479, 168–180. doi: 10.1002/cne.20322

Nieto, M., Schuurmans, C., Britz, O., and Guillemot, F. (2001). Neural bHLH genes control the neuronal versus glial fate decision in cortical progenitors. *Neuron* 29, 401–413. doi: 10.1016/s0896-6273(01)00214-8

Nóbrega-Pereira, S., Kessaris, N., Du, T., Kimura, S., Anderson, S. A., and Marín, O. (2008). Postmitotic Nkx2-1 controls the migration of telencephalic interneurons by direct repression of guidance receptors. *Neuron* 59, 733–745. doi: 10.1016/j.neuron.2008.07.024

Noctor, S. C., Flint, A. C., Weissman, T. A., Dammerman, R. S., and Kriegstein, A. R. (2001). Neurons derived from radial glial cells establish radial units in neocortex. *Nature* 409, 714–720. doi: 10.1038/35055553

Noctor, S. C., Martínez-Cerdeño, V., Ivic, L., and Kriegstein, A. R. (2004). Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. *Nat. Neurosci.* 7, 136–144. doi: 10.1038/nn1172

Nomura, T., Haba, H., and Osumi, N. (2007). Role of a transcription factor Pax6 in the developing vertebrate olfactory system. *Dev. Growth Differ.* 49, 683–690. doi: 10.1111/j.1440-169X.2007.00965.x

Northcott, P. A., Korshunov, A., Witt, H., Hielscher, T., Eberhart, C. G., Mack, S., et al. (2011). Medulloblastoma comprises four distinct molecular variants. *J. Clin. Oncol.* 29:1408. doi: 10.1200/jco.2009.27.4324

Noyes, M. B., Christensen, R. G., Wakabayashi, A., Stormo, G. D., Brodsky, M. H., and Wolfe, S. A. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* 133, 1277–1289. doi: 10.1016/j.cell.2008.05.023

Olivetti, P. R., and Noebels, J. L. (2012). Interneuron, interrupted: molecular pathogenesis of ARX mutations and X-linked infantile spasms. *Curr. Opin. Neurobiol.* 22, 859–865. doi: 10.1016/j.conb.2012.04.006

Okamoto, J., Hirata, T., Chen, Z., Zhou, H., Mikami, I., Li, H., et al. (2010). EMX2 is epigenetically silenced and suppresses growth in human lung cancer. *Oncogene* 29, 5969–5975. doi: 10.1038/onc.2010.330

Omodei, D., Acampora, D., Russo, F., De Filippi, R., Severino, V., Di Francia, R., et al. (2009). Expression of the brain transcription factor OTX1 occurs in a subset of normal germinal-center B cells and in aggressive Non-Hodgkin Lymphoma. *Am. J. Pathol.* 175, 2609–2617. doi: 10.2353/ajpath.2009.090542

Ono, K., Takebayashi, H., Ikeda, K., Furusho, M., Nishizawa, T., Watanabe, K., et al. (2008). Regional-and temporal-dependent changes in the differentiation of Olig2 progenitors in the forebrain, and the impact on astrocyte development in the dorsal pallium. *Dev. Biol.* 320, 456–468. doi: 10.1016/j.ydbio.2008.06.001

Orduz, D., Benamer, N., Ortolani, D., Coppola, E., Vigier, L., Pierani, A., et al. (2019). Developmental cell death regulates lineage-related interneuron-oligodendroglia functional clusters and oligodendrocyte homeostasis. *Nat. Commun.* 10, 1–13. doi: 10.1038/s41467-019-11904-4

Pagliaroli, L., Vereczkei, A., Padmanabhuni, S. S., Tarnok, Z., Farkas, L., Nagy, P., et al. (2020). Association of genetic variation in the 3'UTR of LHX6, IMMP2L, and AADAC with tourette syndrome. *Front. Neurol.* 11:803. doi: 10.3389/fneur.2020.00803

Pan, Y.-H., Wu, N., and Yuan, X.-B. (2019). Toward a better understanding of neuronal migration deficits in autism spectrum disorders. *Front. Cell Dev. Biol.* 7:205. doi: 10.3389/fcell.2019.00205

Pantò, M. R., Zappalà, A., Tuorto, F., and Cicirata, F. (2004). Role of the Otx1 gene in cell differentiation of mammalian cortex. *Eur. J. Neurosci.* 19, 2893–2902. doi: 10.1111/j.0953-816X.2004.03326.x

Park, M., Jeon, S., Jeong, J.-H., Park, M., Lee, D.-R., Yoon, T. K., et al. (2012). Identification and characterization of LHX8 DNA binding elements. *Dev. Reprod.* 16, 379–384. doi: 10.12717/DR.2012.16.4.379

Park, N. I., Guilhamon, P., Desai, K., McAdam, R.F., Langille, E., O'Connor, M., et al. (2017). ASCL1 reorganizes chromatin to direct neuronal fate and suppress tumorigenicity of glioblastoma stem cells. *Cell Stem Cell* 21, 209.e7–224.e7.

Patel, S. R., Bhumbra, S. S., Paknikar, R. S., and Dressler, G. R. (2012). Epigenetic mechanisms of Groucho/Grg/TLE mediated transcriptional repression. *Mol. Cell* 45, 185–195. doi: 10.1016/j.molcel.2011.11.007

Pratt, T., Vitalis, T., Warren, N., Edgar, J. M., Mason, J. O., and Price, D. J. (2000). A role for Pax6 in the normal development of dorsal thalamus and its cortical connections. *Development* 127, 5167–5178. doi: 10.1242/dev.127.23.5167

Pearl, J. R., Colantuoni, C., Bergey, D. E., Funk, C. C., Shannon, P., Basu, B. (2019). Genome-scale transcriptional regulatory network models of psychiatric and neurodegenerative disorders. *Cell Systems* 8, 122.e7–135.e7. doi: 10.1016/j.cels.2019.01.002

Pei, Z., Wang, B., Chen, G., Nagao, M., Nakafuku, M., and Campbell, K. (2011). Homeobox genes Gsx1 and Gsx2 differentially regulate telencephalic progenitor maturation. *Proc. Natl. Acad. Sci.* 108:1675. doi: 10.1073/pnas.1008824108

Pellegrini, M., Mansouri, A., Simeone, A., Boncinelli, E., and Gruss, P. (1996). Dentate gyrus formation requires Emx2. *Development* 122, 3893–3898. doi: 10.1242/dev.122.12.3893

Peng, G., and Westerfield, M. (2006). Lhx5 promotes forebrain development and activates transcription of secreted Wnt antagonists. *Development* 133, 191–200. doi: 10.1242/dev.02485

Peters, T., Dildrop, R., Ausmeier, K., and Rüther, U. (2000). Organization of mouse Iroquois homeobox genes in two clusters suggests a conserved regulation and function in vertebrate development. *Genome Res.* 10, 1453–1462. doi: 10.1101/gr.144100

Petryniak, M. A., Potter, G. B., Rowitch, D. H., and Rubenstein, J. L. R. (2007). Dlx1 and Dlx2 control neuronal versus oligodendroglial cell fate acquisition in the developing forebrain. *Neuron* 55, 417–433. doi: 10.1016/j.neuron.2007.06.036

Piñon, M. C., Tuoc, T. C., Ashery-Padan, R., Molnár, Z., and Stoykova, A. (2008). Altered molecular regionalization and normal thalamocortical connections in cortex-specific Pax6 knock-out mice. *J. Neurosci.* 28, 8724–8734. doi: 10.1523/JNEUROSCI.2565-08.2008

Pla, R., Stanco, A., Howard, M. A., Rubin, A. N., Vogt, D., Mortimer, N., et al. (2017). Dlx1 and Dlx2 promote interneuron GABA synthesis, synaptogenesis, and dendritogenesis. *Cereb. Cortex* 28, 3797–3815. doi: 10.1093/cercor/bhx241

Platzer, K., Cogné, B., Hague, J., Marcelis, C. L., Mitter, D., Oberndorff, K., et al. (2018). Haploinsufficiency of CUX1 causes nonsyndromic global developmental delay with possible catch-up development. *Ann. Neurol.* 84, 200–207. doi: 10.1002/ana.25278

Pleasure, S. J., Anderson, S., Hevner, R., Bagri, A., Marin, O., Lowenstein, D. H., et al. (2000). Cell migration from the ganglionic eminences is required for the development of hippocampal GABAergic interneurons. *Neuron* 28, 727–740. doi: 10.1016/s0896-6273(00)00149-5

Poiana, G., Gioia, R., Sineri, S., Cardarelli, S., Lupo, G., and Cacci, E. (2020). Transcriptional regulation of adult neural stem/progenitor cells: tales from the subventricular zone. *Neural Regen. Res.* 15:1773. doi: 10.4103/1673-5374.280301

Poitras, L., Ghanem, N., Hatch, G., and Ekker, M. (2007). The proneural determinant MASH1 regulates forebrain Dlx1/2 expression through the I12b intergenic enhancer. *Development* 134, 1755–1765. doi: 10.1242/dev.02845

Porter, F. D., Drago, J., Xu, Y., Cheema, S. S., Wassif, C., Huang, S.-P., et al. (1997). Lhx2, a LIM homeobox gene, is required for eye, forebrain, and definitive erythrocyte development. *Development* 124, 2935–2944. doi: 10.1242/dev.124.15.2935

Potkin, S. G., Turner, J. A., Guffanti, G., Lakatos, A., Fallon, J. H., Nguyen, D. D., et al. (2009). A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype. *Schizophr. Bull.* 35, 96–108. doi: 10.1093/schbul/sbn155

Prez, C., Dastot-Le Moal, F., Collot, N., Legendre, M., Abadie, I., Bertrand, A.-M., et al. (2012). Screening of LHX2 in patients presenting growth retardation with posterior pituitary and ocular abnormalities. *Eur. J. Endocrinol.* 167:85. doi: 10.1530/EJE-12-0026

Price, M., Lazzaro, D., Pohl, T., Mattei, M., Rüther, U., Olivo, J.-C., et al. (1992). Regional expression of the homeobox gene Nkx-2.2 in the developing mammalian forebrain. *Neuron* 8, 241–255. doi: 10.1016/0896-6273(92)90291-k

Price, M. G., Yoo, J. W., Burgess, D. L., Deng, F., Hrachovy, R. A., Frost, J. D., et al. (2009). A triplet repeat expansion genetic mouse model of infantile spasms syndrome, Arx (GCG) 10+ 7, with interneuronopathy, spasms in infancy, persistent seizures, and adult cognitive and behavioral impairment. *J. Neurosci.* 29, 8752–8756. doi: 10.1523/JNEUROSCI.0915-09.2009

Prochiantz, A., and Joliot, A. (2003). Can transcription factors function as cell–cell signalling molecules? *Nat. Rev. Mol. Cell Biol.* 4, 814–819. doi: 10.1038/nrm1227

Powell, E. M., Mars, W. M., and Levitt, P. (2001). Hepatocyte growth factor/scatter factor is a motogen for interneurons migrating from the ventral to dorsal telencephalon. *Neuron* 30, 79–89. doi: 10.1016/s0896-6273(01)00264-1

Puelles, E., Acampora, D., Gogoi, R., Tuorto, F., Papalia, A., Guillemot, F., et al. (2006). Otx2 controls identity and fate of glutamatergic progenitors of the thalamus by repressing GABAergic differentiation. *J. Neurosci.* 26, 5955–5964. doi: 10.1523/JNEUROSCI.1097-06.2006

Puelles, L., Kuwana, E., Puelles, E., and Rubenstein, J. (1999). Comparison of the mammalian and avian telencephalon from the perspective of gene expression data. *Eur. J. Morphol.* 37, 139–150. doi: 10.1076/ejom.37.2.139.4756

Puelles, L. (2013). "Plan of the developing vertebrate nervous system," in *Pattering and Cell Type Specification in the Developing CNS and PNS*, eds J. Rubenstein and P. Rakic (Amsterdam: Elsevier), 187–209. doi: 10.1016/b978-0-12-397265-1.00118-0

Puelles, L. (2018). Developmental studies of avian brain organization. *Int. J. Dev. Biol.* 62, 207–224. doi: 10.1387/ijdb.170279LP

Puelles, L., and Rubenstein, J. L. (2003). Forebrain gene expression domains and the evolving prosomeric model. *Trends Neurosci.* 26, 469–476. doi: 10.1016/S0166-2236(03)00234-0

Qiu, H., Yan, Q., Luo, X., Zhang, H., Bao, W., and Wan, X. (2013). EMX2 is downregulated in endometrial cancer and correlated with tumor progression. *Int. J. Gynecol. Pathol.* 32, 193–198. doi: 10.1097/PGP.0b013e31825d8049

Qiu, M., Anderson, S., Chen, S., Meneses, J. J., Hevner, R., Kuwana, E., et al. (1996). Mutation of theEmx-1Homeobox gene disrupts the corpus callosum. *Dev. Biol.* 178, 174–178. doi: 10.1006/dbio.1996.0207

Qiu, M., Bulfone, A., Ghattas, I., Meneses, J. J., Christensen, L., Sharpe, P. T., et al. (1997). Role of the Dlx homeobox genes in proximodistal patterning of the branchial arches: mutations of Dlx-1, Dlx-2, and Dlx-1 and −2 alter morphogenesis of proximal skeletal and soft tissue structures derived from the first and second arches. *Dev. Biol.* 185, 165–184. doi: 10.1006/dbio.1997.8556

Quaggin, S. E., Heuvel, G. B. V., Golden, K., Bodmer, R., and Igarashi, P. (1996). Primary structure, neural-specific expression, and chromosomal localization of cux-2, a second murine homeobox gene related to Drosophila cut*. *J. Biol. Chem.* 271, 22624–22634. doi: 10.1074/jbc.271.37.22624

Quille, M.-L., Carat, S., Quéméner-Redon, S., Hirchaud, E., Baron, D., Benech, C., et al. (2011). High-throughput analysis of promoter occupancy reveals new targets for Arx, a gene mutated in mental retardation and interneuronopathies. *PLoS One* 6:e2518. doi: 10.1371/journal.pone.0025181

Quinn, J.C., Molinek, M., Martynoga, B.S., Zaki, P.A., Faedo, A., Bulfone, A., Hevner, R.F., West, J.D., Price, D.J., et al. (2007). Pax6 controls cerebral cortical cell number by regulating exit from the cell cycle and specifies cortical cell identity by a cell autonomous mechanism. *Developmental biology* 302, 50–65.

Quintana-Urzainqui, I., Kozić, Z., Mitra, S., Tian, T., Manuel, M., Mason, J. O., et al. (2018). Tissue-specific actions of Pax6 on proliferation and differentiation balance in developing forebrain are Foxg1 dependent. *Iscience* 10, 171–191. doi: 10.1016/j.isci.2018.11.031

Ragge, N. K., Brown, A. G., Poloschek, C. M., Lorenz, B., Henderson, R. A., Clarke, M. P., et al. (2005). Heterozygous mutations of OTX2 cause severe ocular malformations. *Am. J. Hum. Genet.* 76, 1008–1022. doi: 10.1086/430721

Rakić, S., Kanatani, S., Hunt, D., Faux, C., Cariboni, A., Chiara, F., et al. (2015). Cdk5 phosphorylation of ErbB4 is required for tangential migration of cortical interneurons. *Cereb. Cortex* 25, 991–1003. doi: 10.1093/cercor/bht290

Rakic, P. (1972). Mode of cell migration to the superficial layers of fetal monkey neocortex. *J. Comp. Neurol.* 145, 61–83. doi: 10.1002/cne.901450105

Ramdzan, Z. M., and Nepveu, A. (2014). CUX1, a haploinsufficient tumour suppressor gene overexpressed in advanced cancers. *Nat. Rev. Cancer* 14, 673–682. doi: 10.1038/nrc3805

Reish, O., Fullston, T., Regev, M., Heyman, E., and Gecz, J. (2009). A novel de novo 27 bp duplication of the ARX gene, resulting from postzygotic mosaicism and leading to three severely affected males in two generations. *Am. J. Med. Genet. Part A* 149, 1655–1660. doi: 10.1002/ajmg.a.32842

Rieckhof, G. E., Casares, F., Ryoo, H. D., Abu-Shaar, M., and Mann, R. S. (1997). Nuclear translocation of extradenticle requires homothorax, which encodes an extradenticle-related homeodomain protein. *Cell* 91, 171–183. doi: 10.1016/s0092-8674(00)80400-6

Rhinn, M., Dierich, A., Shawlot, W., Behringer, R. R., Le Meur, M., and Ang, S.-L. (1998). Sequential roles for Otx2 in visceral endoderm and neuroectoderm for forebrain and midbrain induction and specification. *Development* 125, 845–856. doi: 10.1242/dev.125.5.845

Robertshaw, E., Matsumoto, K., Lumsden, A., and Kiecker, C. (2013). Irx3 and Pax6 establish differential competence for Shh-mediated induction of GABAergic and glutamatergic neurons of the thalamus. *Proc. Natl. Acad. Sci.* 110, E3919–E3926. doi: 10.1073/pnas.1304311110

Robledo, R. F., Rajan, L., Li, X., and Lufkin, T. (2002). The Dlx5 and Dlx6 homeobox genes are essential for craniofacial, axial, and appendicular skeletal development. *Genes Dev.* 16, 1089–1101. doi: 10.1101/gad.988402

Rodríguez-Seguel, E., Alarcón, P., and ómez-Skarmeta, J. L. G. (2009). The Xenopus Irx genes are essential for neural patterning and define the border between prethalamus and thalamus through mutual antagonism with the anterior repressors Fezf and Arx. *Dev. Biol.* 329, 258–268. doi: 10.1016/j.ydbio.2009.02.028

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. (2009). The role of DNA shape in protein–DNA recognition. *Nature* 461, 1248–1253. doi: 10.1038/nature08473

Roy, A., Gonzalez-Gomez, M., Pierani, A., Meyer, G., and Tole, S. (2014). Lhx2 regulates the development of the forebrain hem system. *Cereb. Cortex* 24, 1361–1372. doi: 10.1093/cercor/bhs421

Roychoudhury, K., Salomone, J., Qin, S., Cain, B., Adam, M., Potter, S. S., et al. (2020). Physical interactions between Gsx2 and Ascl1 balance progenitor expansion versus neurogenesis in the mouse lateral ganglionic eminence. *Development* 147:dev185348. doi: 10.1242/dev.185348

Rubenstein, J. L., Martinez, S., Shimamura, K., and Puelles, L. (1994). The embryonic vertebrate forebrain: the prosomeric model. *Science* 266, 578–580. doi: 10.1126/science.7939711

Rudin, C. M., Hann, C. L., Laterra, J., Yauch, R. L., Callahan, C. A., Fu, L., et al. (2009). Treatment of medulloblastoma with hedgehog pathway inhibitor GDC-0449. *New Engl. J. Med.* 361, 1173–1178. doi: 10.1056/NEJMoa0902903

Rudy, B., Fishell, G., Lee, S., and Hjerling-Leffler, J. (2011). Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. *Dev. Neurobiol.* 71, 45–61. doi: 10.1002/dneu.20853

Ruest, L.-B., Hammer, R. E., Yanagisawa, M., and Clouthier, D. E. (2003). Dlx5/6-enhancer directed expression of Cre recombinase in the pharyngeal arches and brain. *Genesis* 37, 188–194. doi: 10.1002/gene.10247

Ryan, A. K., Rosenfeld, M. G., and domain, P. O. U. (1997). family values: flexibility, partnerships, and developmental codes. *Genes Dev.* 11, 1207–1225. doi: 10.1101/gad.11.10.1207

Sagner, A., and Briscoe, J. (2019). Establishing neuronal diversity in the spinal cord: a time and a place. *Development* 146:dev182154. doi: 10.1242/dev.182154

Saha, B., Hari, P., Huilgol, D., and Tole, S. (2007). Dual role for LIM-homeodomain gene Lhx2 in the formation of the lateral olfactory tract. *J. Neurosci.* 27, 2290–2297. doi: 10.1523/JNEUROSCI.5571-06.2007

Salomone, J., Qin, S., Fufa, T. D., Cain, B., Farrow, E., Guan, B., et al. (2021). Conserved Gsx2/Ind homeodomain monomer versus homodimer DNA binding defines regulatory outcomes in flies and mice. *Genes Dev.* 35, 157–174. doi: 10.1101/gad.343053.120

Sandberg, M., Flandin, P., Silberberg, S., Su-Feher, L., Price, J. D., Hu, J. S., et al. (2016). Transcriptional networks controlled by NKX2-1 in the development of forebrain GABAergic neurons. *Neuron* 91, 1260–1275. doi: 10.1016/j.neuron.2016.08.020

Sansregret, L., Gallo, D., Santaguida, M., Leduy, L., Harada, R., and Nepveu, A. (2010). Hyperphosphorylation by cyclin B/CDK1 in mitosis resets CUX1 DNA binding clock at each cell cycle. *J. Biol. Chem.* 285, 32834–32843. doi: 10.1074/jbc.M110.156406

Sansom, S. N., Griffiths, D. S., Faedo, A., Kleinjan, D.-J., Ruan, Y., Smith, J., et al. (2009). The level of the transcription factor Pax6 is essential for controlling the

balance between neural stem cell self-renewal and neurogenesis. *PLoS Genet.* 5:e1000511. doi: 10.1371/journal.pgen.1000511

Santaguida, M., Ding, Q., Bérubé, G., Truscott, M., Whyte, P., and Nepveu, A. (2001). Phosphorylation of the ccaat displacement protein (cdp)/cux transcription factor by cyclin A-Cdk1 modulates its dna binding activity in G2*. *J. Biol. Chem.* 276, 45780–45789. doi: 10.1074/jbc.M107978200

Sakurai, Y., Kurokawa, D., Kiyonari, H., Kajikawa, E., Suda, Y., and Aizawa, S. (2010). Otx2 and Otx1 protect diencephalon and mesencephalon from caudalization into metencephalon during early brain regionalization. *Dev. Biol.* 347, 392–403. doi: 10.1016/j.ydbio.2010.08.028

Saulnier, A., Keruzore, M., De Clercq, S., Bar, I., Moers, V., Magnani, D., et al. (2013). The doublesex homolog Dmrt5 is required for the development of the caudomedial cerebral cortex in mammals. *Cereb. Cortex* 23, 2552–2567. doi: 10.1093/cercor/bhs234

Scardigli, R., Bäumer, N., Gruss, P., Guillemot, F., and Le Roux, I. (2003). Direct and concentration-dependent regulation of the proneural gene Neurogenin2 by Pax6. *Development* 130, 3269–3281. doi: 10.1242/dev.00539

Scherer, S. W., Neufeld, E. J., Lievens, P. M. J., Orkin, S. H., Kim, J., and Tsui, L. C. (1993). Regional localization of the CCAAT displacement protein gene (CUTL1) to 7q22 by analysis of somatic cell hybrids. *Genomics* 15, 695–706. doi: 10.1006/geno.1993.1130

Schonemann, M. D., Ryan, A. K., McEvilly, R. J., O'Connell, S. M., Arias, C. A., Kalla, K. A., et al. (1995). Development and survival of the endocrine hypothalamus and posterior pituitary gland requires the neuronal POU domain factor Brn-2. *Genes Dev.* 9, 3122–3135. doi: 10.1101/gad.9.24.3122

Schmahl, W., Knoedlseder, M., Favor, J., and Davidson, D. (1993). Defects of neuronal migration and the pathogenesis of cortical malformations are associated with Small eye (Sey) in the mouse, a point mutation at the Pax-6-locus. *Acta Neuropathologica* 86, 126–135. doi: 10.1007/BF00334879

Schmidt-Sidor, B., Szymanska, K., Williamson, K., van Heyningen, V., Roszkowski, T., Wierzba-Bobrowicz, T., et al. (2009). Malformations of the brain in two fetuses with a compound heterozygosity for two PAX6 mutations. *Folia Neuropathol.* 47, 372–382.

Schreiber, E., Harshman, K., Kemler, I., Malipiero, U., Schaffner, W., and Fontana, A. (1990). Astrocytes and glioblastoma cells express novel octamer-DNA binding proteins distinct from the ubiquitous Oct-1 and B cell type Oct-2 proteins. *Nucleic Acids Res.* 18, 5495–5503. doi: 10.1093/nar/18.18.5495

Schreiber, E., Himmelmann, A., Malipiero, U., Tobler, A., Stahel, R., and Fontana, A. (1992). Human small cell lung cancer expresses the octamer DNA-binding and nervous system-specific transcription factor N-Oct 3 (brain-2). *Cancer Res.* 52, 6121–6124.

Schulte, D. (2014). Meis: new friends of Pax. *Neurogenesis* 1:e976014. doi: 10.4161/23262133.2014.976014

Schuurmans, C., Armant, O., Nieto, M., Stenman, J. M., Britz, O., Klenin, N., et al. (2004). Sequential phases of cortical specification involve Neurogenin-dependent and-independent pathways. *EMBO J.* 23, 2892–2902. doi: 10.1038/sj.emboj.7600278

Sellers, K., Zyka, V., Lumsden, A. G., and Delogu, A. (2014). Transcriptional control of GABAergic neuronal subtype identity in the thalamus. *Neural Dev.* 9:14. doi: 10.1186/1749-8104-9-14

Sequerra, E. B. (2014). Subventricular zone progenitors in time and space: generating neuronal diversity. *Front. Cell. Neurosci.* 8:434. doi: 10.3389/fncel.2014.00434

Sessa, A., Mao, C.-A., Colasante, G., Nini, A., Klein, W. H., and Broccoli, V. (2010). Tbr2-positive intermediate (basal) neuronal progenitors safeguard cerebral cortex expansion by controlling amplification of pallial glutamatergic neurons and attraction of subpallial GABAergic interneurons. *Genes Dev.* 24, 1816–1826. doi: 10.1101/gad.575410

Sheng, H. Z., Bertuzzi, S., Chiang, C., Shawlot, W., Taira, M., Dawid, I., et al. (1997). Expression of murine Lhx5 suggests a role in specifying the forebrain. *Dev. Dyn.* 208, 266–277. doi: 10.1002/(SICI)1097-0177(199702)208:2<266::AID-AJA13>3.0.CO;2-1

Shi, X., Lin, W., Gao, X., Xie, W., Golden, J. A., and Tao, T. (2020). Identification and validation of the phosphorylation sites on Aristaless-related homeobox protein. *Biosci. Rep.* 40:BSR20194513. doi: 10.1042/BSR20194513

Shi, Y., and Massagué, J. (2003). Mechanisms of TGF-β signaling from cell membrane to the nucleus. *Cell* 113, 685–700. doi: 10.1016/s0092-8674(03)00432-x

Shibata, M., Pattabiraman, K., Lorente-Galdos, B., Andrijevic, D., Kim, S.-K., Kaur, N., et al. (2021). Regulation of prefrontal patterning and connectivity by retinoic acid. *Nature* 598, 483–488. doi: 10.1038/s41586-021-03953-x

Shimizu, T., and Hibi, M. (2009). Formation and patterning of the forebrain and olfactory system by zinc-finger genes Fezf1 and Fezf2. *Dev. Growth Differ.* 51, 221–231. doi: 10.1111/j.1440-169X.2009.01088.x

Shinozaki, K., Miyagi, T., Yoshida, M., Miyata, T., Ogawa, M., Aizawa, S., et al. (2002). Absence of Cajal-Retzius cells and subplate neurons associated with defects of tangential cell migration from ganglionic eminence in Emx1/2 double mutant cerebral cortex. *Development* 129, 3479–3492. doi: 10.1242/dev.129.14.3479

Shinozaki, K., Yoshida, M., Nakamura, M., Aizawa, S., and Suda, Y. (2004). Emx1 and Emx2 cooperate in initial phase of archipallium development. *Mech. Dev.* 121, 475–489. doi: 10.1016/j.mod.2004.03.013

Siehr, M. S., Massey, C. A., and Noebels, J. L. (2020). Arx expansion mutation perturbs cortical development by augmenting apoptosis without activating innate immunity in a mouse model of X-linked infantile spasms syndrome. *Dis. Models Mech.* 13:dmm04251. doi: 10.1242/dmm.042515

Silbereis, J. C., Nobuta, H., Tsai, H.-H., Heine, V. M., McKinsey, G. L., Meijer, D. H., et al. (2014). Olig1 function is required to repress dlx1/2 and interneuron production in Mammalian brain. *Neuron* 81, 574–587. doi: 10.1016/j.neuron.2013.11.024

Simeone, A., Acampora, D., Gulisano, M., Stornaiuolo, A., and Boncinelli, E. (1992). Nested expression domains of four homeobox genes in developing rostral brain. *Nature* 358, 687–690. doi: 10.1038/358687a0

Simeone, A., Acampora, D., Mallamaci, A., Stornaiuolo, A., D'Apice, M. R., Nigro, V., et al. (1993). A vertebrate gene related to orthodenticle contains a homeodomain of the bicoid class and demarcates anterior neuroectoderm in the gastrulating mouse embryo. *EMBO J.* 12, 2735–2747. doi: 10.1002/j.1460-2075.1993.tb05935.x

Simmons, J. L., Pierce, C. J., Al-Ejeh, F., and Boyle, G. M. (2017). MITF and BRN2 contribute to metastatic growth after dissemination of melanoma. *Sci. Rep.* 7:10909. doi: 10.1038/s41598-017-11366-y

Sisodiya, S. M., Free, S. L., Williamson, K. A., Mitchell, T. N., Willis, C., Stevens, J. M., et al. (2001). PAX6 haploinsufficiency causes cerebral malformation and olfactory dysfunction in humans. *Nat. Genet.* 28, 214–216. doi: 10.1038/90042

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C. G., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507, 371–375. doi: 10.1038/nature13138

Son, J. E., Dou, Z., Kim, K.-H., Wanggou, S., Cha, V. S. B., Mo, R., et al. (2021a). Irx3 and Irx5 in Ins2-Cre+ cells regulate hypothalamic postnatal neurogenesis and leptin response. *Nat. Metab.* 3, 701–713. doi: 10.1038/s42255-021-00382-y

Son, J. E., Dou, Z., Wanggou, S., Chan, J., Mo, R., Li, X., et al. (2021b). Ectopic expression of Irx3 and Irx5 in the paraventricular nucleus of the hypothalamus contributes to defects in Sim1 haploinsufficiency. *Sci. Adv.* 7:eabh4503. doi: 10.1126/sciadv.abh4503

Solomon, B. D., Pineda-Alvarez, D. E., Balog, J. Z., Hadley, D., Gropman, A. L., Nandagopal, R., et al. (2009). Compound heterozygosity for mutations in PAX6 in a patient with complex brain anomaly, neonatal diabetes mellitus, and microophthalmia. *Am. J. Med. Genet. Part A* 149, 2543–2546. doi: 10.1002/ajmg.a.33081

Soriano, E., and Del Río, J. A. (2005). The cells of cajal-retzius: still a mystery one century after. *Neuron* 46, 389–394. doi: 10.1016/j.neuron.2005.04.019

Spigoni, G., Gedressi, C., and Mallamaci, A. (2010). Regulation of Emx2 expression by antisense transcripts in murine cortico-cerebral precursors. *PLoS One* 5:e8658. doi: 10.1371/journal.pone.0008658

Srivastava, D., Thomas, T., Lin, Q., Kirby, M. L., Brown, D., and Olson, E. N. (1997). Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND. *Nat. Genet.* 16, 154–160. doi: 10.1038/ng0697-154

Stenman, J., Toresson, H., and Campbell, K. (2003). Identification of two distinct progenitor populations in the lateral ganglionic eminence: implications for striatal and olfactory bulb neurogenesis. *J. Neurosci.* 23, 167–174. doi: 10.1523/JNEUROSCI.23-01-00167.2003

Stiles, J. (2008). *The Fundamentals of Brain Development: Integrating Nature and Nurture*. Cambridge, MA: Harvard University Press.

Stock, D. W., Ellies, D. L., Zhao, Z., Ekker, M., Ruddle, F. H., and Weiss, K. M. (1996). The evolution of the vertebrate Dlx gene family. *Proc. Natl. Acade. Sci.* 93, 10858–10863. doi: 10.1073/pnas.93.20.10858

Stocker, A. M., and O'Leary, D. D. (2016). Emx1 is required for neocortical area patterning. *PLoS One* 11:e0149900. doi: 10.1371/journal.pone.0149900

Stoykova, A., and Gruss, P. (1994). Roles of Pax-genes in developing and adult brain as suggested by expression patterns. *J. Neurosci.* 14, 1395–1412. doi: 10.1523/JNEUROSCI.14-03-01395.1994

Stoykova, A., Fritsch, R., Walther, C., and Gruss, P. (1996). Forebrain patterning defects in small eye mutant mice. *Development* 122, 3453–3465. doi: 10.1242/dev.122.11.3453

Stoykova, A., Gotz, M., Gruss, P., and Price, J. (1997). Pax6-dependent regulation of adhesive patterning, R-cadherin expression and boundary formation in developing forebrain. *Development* 124, 3765–3777. doi: 10.1242/dev.124.19.3765

Stoykova, A., Treichel, D., Hallonet, M., and Gruss, P. (2000). Pax6 modulates the dorsoventral patterning of the mammalian telencephalon. *J. Neurosci.* 20, 8042–8050. doi: 10.1523/JNEUROSCI.20-21-08042.2000

Striano, P., Paravidino, R., Sicca, F., Chiurazzi, P., Gimelli, S., Coppola, A., et al. (2011). West syndrome associated with 14q12 duplications harboring FOXG1. *Neurology* 76, 1600–1602. doi: 10.1212/WNL.0b013e3182194bbf

Strømme, P., Mangelsdorf, M. E., Scheffer, I. E., and Gécz, J. (2002). Infantile spasms, dystonia, and other X-linked phenotypes caused by mutations in Aristaless related homeobox gene, ARX. *Brain Dev.* 24, 266–288. doi: 10.1016/s0387-7604(02)00079-7

Stühmer, T., Anderson, S. A., Ekker, M., and Rubenstein, J. L. R. (2002). Ectopic expression of the Dlx genes induces glutamic acid decarboxylase and Dlx expression. *Development* 129, 245–252. doi: 10.1242/dev.129.1.245

Stumm, R. K., Zhou, C., Ara, T., Lazarini, F., Dubois-Dalcq, M., Nagasawa, T., et al. (2003). CXCR4 regulates interneuron migration in the developing neocortex. *J. Neurosci.* 23, 5123–5130. doi: 10.1523/JNEUROSCI.23-12-05123.2003

Su, Z., Wang, Z., Lindtner, S., Yang, L., Shang, Z., Tian, Y., et al. (2022). Dlx1/2-dependent expression of Meis2 promotes neuronal fate determination in the mammalian striatum. *Development* 149:dev200035. doi: 10.1242/dev.200035

Suda, Y., Kokura, K., Kimura, J., Kajikawa, E., Inoue, F., and Aizawa, S. (2010). The same enhancer regulates the earliest Emx2 expression in caudal forebrain primordium, subsequent expression in dorsal telencephalon and later expression in the cortical ventricular zone. *Development* 137, 2939–2949. doi: 10.1242/dev.048843

Sugimori, M., Nagao, M., Bertrand, N., Parras, C.M., Guillemot, F.o., and Nakafuku, M. (2007). Combinatorial actions of patterning and HLH transcription factors in the spatiotemporal control of neurogenesis and gliogenesis in the developing spinal cord. *Development* 134, 1617–1629. doi: 10.1242/dev.001255

Sugitani, Y., Nakai, S., Minowa, O., Nishi, M., Jishage, K.-i., Kawano, H., et al. (2002). Brn-1 and Brn-2 share crucial roles in the production and positioning of mouse neocortical neurons. *Genes Dev.* 16, 1760–1765. doi: 10.1101/gad.978002

Sumiyama, K., Washio-Watanabe, K., Saitou, N., Hayakawa, T., and Ueda, S. (1996). Class III POU genes: generation of homopolymeric amino acid repeats under GC pressure in mammals. *J. Mol. Evol.* 43, 170–178. doi: 10.1007/BF02338824

Sun, J., Rockowitz, S., Xie, Q., Ashery-Padan, R., Zheng, D., and Cvekl, A. (2015). Identification of in vivo DNA-binding mechanisms of Pax6 and reconstruction of Pax6-dependent gene regulatory networks during forebrain and lens development. *Nucleic Acids Res.* 43, 6827–6846. doi: 10.1093/nar/gkv589

Sun, T., Pringle, N. P., Hardy, A. P., Richardson, W. D., and Smith, H. K. (1998). Pax6 influences the time and site of origin of glial precursors in the ventral neural tube. *Mol. Cell. Neurosci.* 12, 228–239. doi: 10.1006/mcne.1998.0711

Sussel, L., Marin, O., Kimura, S., and Rubenstein, J. (1999). Loss of Nkx2. 1 homeobox gene function results in a ventral to dorsal molecular respecification within the basal telencephalon: evidence for a transformation of the pallidum into the striatum. *Development* 126, 3359–3370. doi: 10.1242/dev.126.15.3359

Symmank, J., Gölling, V., Gerstmann, K., and Zimmer, G. (2019). The transcription factor LHX1 regulates the survival and directed migration of POA-derived cortical interneurons. *Cereb. Cortex* 29, 1644–1658. doi: 10.1093/cercor/bhy063

Szczaluba, K., Nawara, M., Poirier, K., Pilch, J., Gajdulewicz, M., Spodar, K., et al. (2006). Genotype-phenotype associations for ARX gene duplication in X-linked mental retardation. *Neurology* 67, 2073–2075. doi: 10.1212/01.wnl.0000247833.29314.5b

Szucsik, J. C., Witte, D. P., Li, H., Pixley, S. K., Small, K. M., and Potter, S. S. (1997). Altered forebrain and hindbrain development in mice mutant for the Gsh-2 homeobox gene. *Dev. Biol.* 191, 230–242. doi: 10.1006/dbio.1997.8733

Tajima, T., Ohtake, A., Hoshino, M., Amemiya, S., Sasaki, N., Ishizu, K., et al. (2009). OTX2 loss of function mutation causes anophthalmia and combined pituitary hormone deficiency with a small anterior and ectopic posterior pituitary. *J. Clin. Endocrinol. Metab.* 94, 314–319. doi: 10.1210/jc.2008-1219

Takebayashi, H., Nabeshima, Y., Yoshida, S., Chisaka, O., Ikenaka, K., and Nabeshima, Y.-i. (2002). The basic helix-loop-helix factor olig2 is essential for the development of motoneuron and oligodendrocyte lineages. *Curr. Biol.* 12, 1157–1163. doi: 10.1016/s0960-9822(02)00926-0

Takebayashi, H., Yoshida, S., Sugimori, M., Kosako, H., Kominami, R., Nakafuku, M., et al. (2000). Dynamic expression of basic helix-loop-helix Olig family members: implication of Olig2 in neuron and oligodendrocyte differentiation and identification of a new member, Olig3. *Mech. Dev.* 99, 143–148. doi: 10.1016/s0925-4773(00)00466-4

Takuma, N., Sheng, H. Z., Furuta, Y., Ward, J. M., Sharma, K., Hogan, B., et al. (1998). Formation of Rathke's pouch requires dual induction from the diencephalon. *Development* 125, 4835–4840. doi: 10.1242/dev.125.23.4835

Tao, W., and Lai, E. (1992). Telencephalon-restricted expression of BF-1, a new member of the HNF-3/fork head gene family, in the developing rat brain. *Neuron* 8, 957–966. doi: 10.1016/0896-6273(92)90210-5

Taylor, M. D., Northcott, P. A., Korshunov, A., Remke, M., Cho, Y.-J., Clifford, S. C., et al. (2012). Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathologica* 123, 465–472. doi: 10.1007/s00401-011-0922-z

Tekki-Kessaris, N., Woodruff, R., Hall, A. C., Gaffield, W., Kimura, S., Stiles, C. D., et al. (2001). Hedgehog-dependent oligodendrocyte lineage specification in the telencephalon. *Development* 128, 2545–2554. doi: 10.1242/dev.128.13.2545

Terrinoni, A., Pagani, I., Zucchi, I., Chiaravalli, A., Serra, V., Rovera, F., et al. (2011). OTX1 expression in breast cancer is regulated by p53. *Oncogene* 30, 3096–3103. doi: 10.1038/onc.2011.31

Thakurela, S., Tiwari, N., Schick, S., Garding, A., Ivanek, R., Berninger, B., et al. (2016). Mapping gene regulatory circuitry of Pax6 during neurogenesis. *Cell Discovery* 2, 1–22. doi: 10.1038/celldisc.2015.45

Theil, T., Alvarez-Bolado, G., Walter, A., and Ruther, U. (1999). Gli3 is required for Emx gene expression during dorsal telencephalon development. *Development* 126, 3561–3571. doi: 10.1242/dev.126.16.3561

Theil, T., Aydin, S.l., Koch, S., Grotewold, L., and Rüther, U. (2002). Wnt and Bmp signalling cooperatively regulate graded Emx 2 expression in the dorsal telencephalon. *Development* 129, 3045–3054. doi: 10.1242/dev.129.13.3045

Thomson, J. A., Murphy, K., Baker, E., Sutherland, G. R., Parsons, P. G., Sturm, R. A., et al. (1995). The brn-2 gene regulates the melanocytic phenotype and tumorigenic potential of human melanoma cells. *Oncogene* 11, 691–700.

Thorwarth, A., Schnittert-Hübener, S., Schrumpf, P., Müller, I., Jyrch, S., Dame, C., et al. (2014). Comprehensive genotyping and clinical characterisation reveal 27 novel NKX2-1 mutations and expand the phenotypic spectrum. *J. Med. Genet.* 51, 375–387. doi: 10.1136/jmedgenet-2013-102248

Tian, E., Kimura, C., Takeda, N., Aizawa, S., and Matsuo, I. (2002). Otx2 is required to respond to signals from anterior neural ridge for forebrain specification. *Dev. Biol.* 242, 204–223. doi: 10.1006/dbio.2001.0531

Toresson, H., Mata de Urquiza, A., Fagerstrom, C., Perlmann, T., and Campbell, K. (1999). Retinoids are produced by glia in the lateral ganglionic eminence and regulate striatal neuron differentiation. *Development* 126, 1317–1326. doi: 10.1242/dev.126.6.1317

Toresson, H., Parmar, M., and Campbell, K. (2000). Expression of Meis and Pbx genes and their protein products in the developing telencephalon: implications for regional differentiation. *Mech. Dev.* 94, 183–187. doi: 10.1016/s0925-4773(00)00324-5

Toresson, H. K., and Campbell, K. (2001). A role for Gsh1 in the developing striatum and olfactory bulb of Gsh2 mutant mice. *Development* 128, 4769–4780. doi: 10.1242/dev.128.23.4769

Truscott, M., Denault, J. B., Goulet, B., Leduy, L., Salvesen, G. S., and Nepveu, A. (2007). Carboxyl-terminal proteolytic processing of CUX1 by a caspase enables

transcriptional activation in proliferating cells. *J. Biol. Chem.* 282, 30216–30226. doi: 10.1074/jbc.M702328200

Truscott, M., Raynal, L., Wang, Y., Bérubé, G., Leduy, L., and Nepveu, A. (2004). The n-terminal region of the CCAAT displacement protein (CDP)/cux transcription factor functions as an autoinhibitory domain that modulates DNA binding*. *J. Biol. Chem.* 279, 49787–40794. doi: 10.1074/jbc.M409484200

Tu, X.-P., Li, H., Chen, L.-S., Luo, X.-N., Lu, Z.-M., Zhang, S.-Y., et al. (2020). OTX1 exerts an oncogenic role and is negatively regulated by miR129-5p in laryngeal squamous cell carcinoma. *BMC Cancer* 20:794. doi: 10.1186/s12885-020-07279-1

Tucker, E. S., Polleux, F., and LaMantia, A. S. (2006). Position and time specify the migration of a pioneering population of olfactory bulb interneurons. *Dev. Biol.* 297, 387–401. doi: 10.1016/j.ydbio.2006.05.009

Tufo, C., Poopalasundaram, S., Dorrego-Rivas, A., Ford, M. C., Graham, A., and Grubb, M. S. (2022). Development of the mammalian main olfactory bulb. *Development* 149:dev200210. doi: 10.1242/dev.200210

Turner, G., Partington, M., Kerr, B., Mangelsdorf, M., and Gecz, J. (2002). Variable expression of mental retardation, autism, seizures, and dystonic hand movements in two families with an identical ARX gene mutation. *Am. J. Med. Genet.* 112, 405–411. doi: 10.1002/ajmg.10714

Tyas, D. A., Pearson, H., Rashbass, P., and Price, D. J. (2003). Pax6 regulates cell adhesion during cortical development. *Cereb. Cortex* 13, 612–619. doi: 10.1093/cercor/13.6.612

Valerius, M. T., Li, H., Stock, J. L., Weinstein, M., Kaur, S., Singh, G., et al. (1995). Gsh-1: a novel murine homeobox gene expressed in the central nervous system. *Dev. Dyn.* 203, 337–351. doi: 10.1002/aja.1002030306

Verret, L., Mann, E. O., Hang, G. B., Barth, A. M. I., Cobos, I., Ho, K., et al. (2012). Inhibitory interneuron deficit links altered network activity and cognitive dysfunction in alzheimer model. *Cell* 149, 708–721. doi: 10.1016/j.cell.2012.02.046

Volk, D. W., Edelson, J. R., and Lewis, D. A. (2014). Cortical inhibitory neuron disturbances in schizophrenia: role of the ontogenetic transcription factor Lhx6. *Schizophr. Bull.* 40, 1053–1061. doi: 10.1093/schbul/sbu068

Volk, D. W., Matsubara, T., Li, S., Sengupta, E. J., Georgiev, D., Minabe, Y., et al. (2012). Deficits in transcriptional regulators of cortical parvalbumin neurons in schizophrenia. *Am. J. Psychiatry* 169, 1082–1091. doi: 10.1176/appi.ajp.2012.12030305

von Engelhardt, J., Khrulev, S., Eliava, M., Wahlster, S., and Monyer, H. (2011). 5-HT(3A) receptor-bearing white matter interstitial GABAergic interneurons are functionally integrated into cortical and subcortical networks. *J. Neurosci.* 31:16844. doi: 10.1523/JNEUROSCI.0310-11.2011

Vue, T. Y., Aaker, J., Taniguchi, A., Kazemzadeh, C., Skidmore, J. M., Martin, D. M., et al. (2007). Characterization of progenitor domains in the developing mouse thalamus. *J. Comp. Neurol.* 505, 73–91. doi: 10.1002/cne.21467

Waclaw, R. R., Ehrman, L. A., Pierani, A., and Campbell, K. (2010). Developmental origin of the neuronal subtypes that comprise the amygdalar fear circuit in the mouse. *J. Neurosci.* 30:6944. doi: 10.1523/JNEUROSCI.5772-09.2010

Walcher, T., Xie, Q., Sun, J., Irmler, M., Beckers, J., Öztürk, T., et al. (2013). Functional dissection of the paired domain of Pax6 reveals molecular mechanisms of coordinating neurogenesis and proliferation. *Development* 140, 1123–1136. doi: 10.1242/dev.082875

Walther, C., Guenet, J.-L., Simon, D., Deutsch, U., Jostes, B., Goulding, M. D., et al. (1991). Pax: a murine multigene family of paired box-containing genes. *Genomics* 11, 424–434. doi: 10.1016/0888-7543(91)90151-4

Wang, B., Long, J. E., Flandin, P., Pla, R., Waclaw, R. R., Campbell, K., et al. (2013). Loss of Gsx1 and Gsx2 function rescues distinct phenotypes in Dlx1/2 mutants. *J. Comp. Neurol.* 521, 1561–1584. doi: 10.1002/cne.23242

Wang, L., Bluske, K. K., Dickel, L. K., and Nakagawa, Y. (2011a). Basal progenitor cells in the embryonic mouse thalamus - their molecular characterization and the role of neurogenins and Pax6. *Neural Dev.* 6, 35–35. doi: 10.1186/1749-8104-6-35

Wang, Y., Dye, C. A., Sohal, V., Long, J. E., Estrada, R. C., Roztocil, T., et al. (2010). Dlx5 and Dlx6 regulate the development of parvalbumin-expressing cortical interneurons. *J. Neurosci.* 30, 5334–5345. doi: 10.1523/JNEUROSCI.5963-09.2010

Wang, Y., Li, G., Stanco, A., Long, J. E., Crawford, D., Potter, G. B., et al. (2011b). CXCR4 and CXCR7 have distinct functions in regulating interneuron migration. *Neuron* 69, 61–76. doi: 10.1016/j.neuron.2010.12.005

Wanaka, A., Matsumoto, K., Kashihara, Y., Furuyama, T., Tanaka, T., Mori, T., et al. (1997). LIM-homeodomain gene family in neural development. *Dev. Neurosci.* 19, 97–100. doi: 10.1159/000111191

Way, J. C., and Chalfie, M. (1988). mec-3, a homeobox-containing gene that specifies differentiation of the touch receptor neurons in *C. elegans*. *Cell* 54, 5–16. doi: 10.1016/0092-8674(88)90174-2

Webb, A. E., Pollina, E. A., Vierbuchen, T., Urbán, N., Ucar, D., Leeman, D. S., et al. (2013). FOXO3 shares common targets with ASCL1 genome-wide and inhibits ASCL1-dependent neurogenesis. *Cell Rep.* 4, 477–491. doi: 10.1016/j.celrep.2013.06.035

Weinschutz Mendes, H., Taktek, M., Duret, T., and Ekker, M. (2020). Expression of dlx genes in the normal and regenerating brain of adult zebrafish. *PLoS One* 15:e0229549.

Weiss, L. A., and Nieto, M. (2019). The crux of Cux genes in neuronal function and plasticity. *Brain Res.* 1705, 32–42. doi: 10.1016/j.brainres.2018.02.044

Weiss, J. B., Von Ohlen, T., Mellerick, D. M., Dressler, G., Doe, C. Q., and Scott, M. P. (1998). Dorsoventral patterning in the Drosophila central nervous system: the intermediate neuroblasts defective homeobox gene specifies intermediate column identity. *Genes Dev.* 12, 3591–3602. doi: 10.1101/gad.12.22.3591

Westphal, D.S., Riedhammer, K.M., Kovacs-Nagy, R., Meitinger, T., Hoefele, J., and Wagner, M. (2018). A de novo missense variant in pou3f2 identified in a child with global developmental delay. *Neuropediatrics* 49, 401–404. doi: 10.1055/s-0038-1669926

Whitman, M. C., and Greer, C. A. (2009). Adult neurogenesis and the olfactory system. *Prog. Neurobiol.* 89, 162–175. doi: 10.1016/j.pneurobio.2009.07.003

Wichterle, H., Turnbull, D. H., Nery, S., Fishell, G., and Alvarez-Buylla, A. (2001). In utero fate mapping reveals distinct migratory pathways and fates of neurons born in the mammalian basal forebrain. *Development* 128, 3759–3771. doi: 10.1242/dev.128.19.3759

Winterbottom, E. F., Illes, J. C., Faas, L., and Isaacs, H. V. (2010). Conserved and novel roles for the Gsh2 transcription factor in primary neurogenesis. *Development* 137, 2623–2631. doi: 10.1242/dev.047159

Won, H., de La Torre-Ubieta, L., Stein, J. L., Parikshak, N. N., Huang, J., Opland, C. K., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* 538, 523–527 doi: 10.1038/nature19847

Wortham, M., Guo, C., Zhang, M., Song, L., Lee, B.-K., Iyer, V. R., et al. (2014). Chromatin accessibility mapping identifies mediators of basal transcription and retinoid-induced repression of OTX2 in medulloblastoma. *PLoS One* 9:e107156. doi: 10.1371/journal.pone.0107156

Wright, A. G., Demyanenko, G. P., Powell, A., Schachner, M., Enriquez-Barreto, L., Tran, T. S., et al. (2007). Close homolog of L1 and neuropilin 1 mediate guidance of thalamocortical axons at the ventral telencephalon. *J. Neurosci.* 27, 13667–13679. doi: 10.1523/JNEUROSCI.2888-07.2007

Wong, C. C., Martincorena, I., Rust, A. G., Rashid, M., Alifrangis, C., Alexandrov, L. B., et al. (2014). Inactivating CUX1 mutations promote tumorigenesis. *Nat. Genet.* 46, 33–38. doi: 10.1038/ng.2846

Xu, Q., Cobos, I., De La Cruz, E., Rubenstein, J. L., and Anderson, S. A. (2004). Origins of cortical interneuron subtypes. *J. Neurosci.* 24, 2612–2622. doi: 10.1523/jneurosci.5667-03.2004

Xu, Q., Wonders, C. P., and Anderson, S. A. (2005). Sonic hedgehog maintains the identity of cortical interneuron progenitors in the ventral telencephalon. *Development* 132, 4987–4998. doi: 10.1242/dev.02090

Xu, Y., Xi, J., Wang, G., Guo, Z., Sun, Q., Lu, C., et al. (2021). PAUPAR and PAX6 sequentially regulate human embryonic stem cell cortical differentiation. *Nucleic Acids Res.* 49, 1935–1950. doi: 10.1093/nar/gkab030

Xuan, S., Baptista, C. A., Balas, G., Tao, W., Soares, V. C., and Lai, E. (1995). Winged helix transcription factor BF-1 is essential for the development of the cerebral hemispheres. *Neuron* 14, 1141–1152. doi: 10.1016/0896-6273(95)90262-7

Yang, Y., Hwang, C. K., D'Souza, U. M., Lee, S. H., Junn, E., and Mouradian, M. M. (2000). Three-amino acid extension loop homeodomain proteins Meis2 and TGIF differentially regulate transcription. *J. Biol. Chem.* 275, 20734–20741. doi: 10.1074/jbc.M908382199

Yang, Y., Shen, W., Ni, Y., Su, Y., Yang, Z., and Zhao, C. (2017). Impaired interneuron development after Foxg1 disruption. *Cereb. Cortex* 27, 793–808. doi: 10.1093/cercor/bhv297

Yoshihara, S., Omichi, K., Yanazawa, M., Kitamura, K., and Yoshihara, Y. (2005). Arx homeobox gene is essential for development of mouse olfactory system. *Development* 132, 751–762. doi: 10.1242/dev.01619

Yoshida, M., Suda, Y., Matsuo, I., Miyamoto, N., Takeda, N., Kuratani, S., et al. (1997). Emx1 and Emx2 functions in development of dorsal telencephalon. *Development* 124, 101–111. doi: 10.1242/dev.124.1.101

Yoon, S. O., and Chikaraishi, D. M. (1994). Isolation of two E-box binding factors that interact with the rat tyrosine hydroxylase enhancer. *J. Biol. Chem.* 269, 18453–18462. doi: 10.1016/s0021-9258(17)32330-x

Younossi-Hartenstein, A., Green, P., Liaw, G.-J., Rudolph, K., Lengyel, J., and Hartenstein, V. (1997). Control of early neurogenesis of thedrosophilabrain by the head gap genestll, otd, ems, andbtd. *Dev. Biol.* 182, 270–283. doi: 10.1006/dbio.1996.8475

Ypsilanti, A. R., Pattabiraman, K., Catta-Preta, R., Golonzhka, O., Lindtner, S., Tang, K., et al. (2021). Transcriptional network orchestrating regional patterning of cortical progenitors. *Proc. Natl. Acad. Sci.* 118:e2024795118 doi: 10.1073/pnas.2024795118

Yuan, F., Chen, X., Fang, K.-H., Wang, Y., Lin, M., Xu, S.-B., et al. (2018). Induction of human somatostatin and parvalbumin neurons by expressing a single transcription factor LIM homeobox 6. *Elife* 7:e37382. doi: 10.7554/eLife.37382

Yun, K., Fischman, S., Johnson, J., de Angelis, M. H., Weinmaster, G., and Rubenstein, J. L. R. (2002). Modulation of the notch signaling by Mash1 and Dlx1/2regulates sequential specification and differentiation of progenitor cell types in the subcortical telencephalon. *Development* 129, 5029–5040.

Yun, K., Garel, S., Fischman, S., and Rubenstein, J. L. R. (2003). Patterning of the lateral ganglionic eminence by the Gsh1 and Gsh2 homeobox genes regulates striatal and olfactory bulb histogenesis and the growth of axons through the basal ganglia. *J. Comp. Neurol.* 461, 151–165. doi: 10.1002/cne.10685

Yun, K., Potter, S., and Rubenstein, J. (2001). Gsh2 and Pax6 play complementary roles in dorsoventral patterning of the mammalian telencephalon. *Development* 128, 193–205. doi: 10.1242/dev.128.2.193

Zagozewski, J., Shahriary, G. M., Morrison, L. C., Saulnier, O., Stromecki, M., Fresnoza, A., et al. (2020). An OTX2-PAX3 signaling axis regulates Group 3 medulloblastoma cell fate. *Nat. Commun.* 11, 1–18. doi: 10.1038/s41467-020-17357-4

Zakrzewska, M., Grešner, S. M., Zakrzewski, K., Zalewska-Szewczyk, B., and Liberski, P. P. (2013). Novel gene expression model for outcome prediction in paediatric medulloblastoma. *J. Mol. Neurosci.* 51, 371–379. doi: 10.1007/s12031-013-0016-6

Zerucha, T., Stühmer, T., Hatch, G., Park, B. K., Long, Q., Yu, G., et al. (2000). A highly conserved enhancer in the Dlx5/Dlx6 intergenic region is the site of cross-regulatory interactions between dlx genes in the embryonic forebrain. *J. Neurosci.* 20:709. doi: 10.1523/JNEUROSCI.20-02-00709.2000

Zhao, Y., Guo, Y.-J., Tomac, A. C., Taylor, N. R., Grinberg, A., Lee, E. J., et al. (1999). Isolated cleft palate in mice with a targeted mutation of the LIM homeobox gene lhx8. *Proc. Natl. Acad. Sci.* 96, 15002–15006. doi: 10.1073/pnas.96.26.15002

Zhao, Y., Flandin, P., Long, J. E., Cuesta, M. D., Westphal, H., and Rubenstein, J. L. (2008). Distinct molecular pathways for development of telencephalic interneuron subtypes revealed through analysis of Lhx6 mutants. *J. Comp. Neurol.* 510, 79–99. doi: 10.1002/cne.21772

Zhao, Y., Marín, O., Hermesz, E., Powell, A., Flames, N., Palkovits, M., et al. (2003). The LIM-homeobox gene Lhx8 is required for the development of many cholinergic neurons in the mouse forebrain. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9005–9010. doi: 10.1073/pnas.1537759100

Zhou, Q. P., Le, T. N., Qiu, X., Spencer, V., de Melo, J., Du, G., et al. (2004). Identification of a direct Dlx homeodomain target in the developing mouse forebrain and retina by optimization of chromatin immunoprecipitation. *Nucleic Acids Res.* 32, 884–892. doi: 10.1093/nar/gkh233

Ziffra, R. S., Kim, C. N., Ross, J. M., Wilfert, A., Turner, T. N., Haeussler, M., et al. (2021). Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* 598, 205–213. doi: 10.1038/s41586-021-03209-8

Zimmer, C., Tiveron, M.-C., Bodmer, R., and Cremer, H. (2004). Dynamics of Cux2 expression suggests that an early pool of SVZ precursors is fated to become upper cortical layer neurons. *Cereb. Cortex* 14, 1408–1420. doi: 10.1093/cercor/bhh102

Zuccoli, G., Yannes, M. P., Nardone, R., Bailey, A., and Goldstein, A. (2015). Bilateral symmetrical basal ganglia and thalamic lesions in children: an update. *Neuroradiology* 57, 973–989. doi: 10.1007/s00234-015-1568-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Leung, George, Roussel, Faux, Wigle and Eisenstat. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 29 April 2022
doi: 10.3389/fgene.2022.854211

Identification of Enhancer RNA CDK6-AS1 as a Potential Novel Prognostic Biomarker in Gastric Cancer

Shifeng Yang, Xiaoming Zou, Hao Yang, Jiacheng Li, Ange Zhang, Lisha Zhang, Changjian Li, Lei Zhu and Zhen Ma*

Department of Gastrointestinal Surgery, The Second Affiliated Hospital of Harbin Medical University, Harbin, China

Background: This study aimed to confirm the role of enhancer RNAs (eRNAs) in gastric cancer and their clinical utility.

Methods: We used Cox survival and relevance analysis to identify the candidate eRNAs in gastric cancer and performed Gene Ontology and Reactome pathway enrichment to determine the potential functions of eRNAs. Correlation between eRNA, tumor-infiltrating immune cells, and drug sensitivity was then analyzed.

Results: *CDK6-AS1*, a long non-coding RNA cyclin-dependent kinase 6, may serve as a poor potential prognostic biomarker candidate in gastric cancer with a positive correlation with its target gene *CDK6*. The low *CDK6-AS1* expression group showed more frequent mutated driver genes than the high expression group. Moreover, *CDK6-AS1* is involved in a key oncogenic pathway of the cell cycle and RNA transcription. *CDK6-AS1* also shows dysregulations and associations with prognosis at the pan-cancer level. This eRNA may also be associated with immune cell infiltration and drug sensitivity.

Conclusion: *CDK6-AS1* may be a potential prognostic biomarker and chemotherapeutic drug sensitivity predictor in gastric cancer.

Keywords: CDK6-AS1, CDK6, enhancer RNA, gastric cancer, prognostic biomarker

OPEN ACCESS

Edited by:
Wei-Lin Jin,
First Hospital of Lanzhou University, China

Reviewed by:
Atrayee Bhattacharya,
Dana–Farber Cancer Institute, United States
Kamini Singh,
Albert Einstein College of Medicine, United States
Yigan Zhang,
First Hospital of Lanzhou University, China

****Correspondence:***
Xiaoming Zou
zou4930@163.com

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: *13 January 2022*
Accepted: *01 April 2022*
Published: *29 April 2022*

Citation:
Yang S, Zou X, Yang H, Li J, Zhang A, Zhang L, Li C, Zhu L and Ma Z (2022) Identification of Enhancer RNA CDK6-AS1 as a Potential Novel Prognostic Biomarker in Gastric Cancer. Front. Genet. 13:854211. doi: 10.3389/fgene.2022.854211

INTRODUCTION

Gastric cancer was the world's fifth most commonly diagnosed cancer type and the sixth cause of cancer mortality in 2018, responsible for 1,033,701 newly diagnosed cases and 782,685 deaths worldwide (Bray et al., 2018). During the past few decades, gastric cancer has maintained a high case fatality rate of 75% throughout most of the world and is the main contributor to the global disability-adjusted life-year burden (Soerjomataram et al., 2012; Thrift and El-Serag, 2020). Recently, the prognosis of gastric cancer has improved and its treatment technology has significantly improved (The Cancer Genome Atlas Research Network, 2014; Digklia and Wagner, 2016).

The past few decades have witnessed the rapid progress of knowledge about the role noncoding RNAs play in a wide range of cancers (Martens-Uzunova et al., 2014). An increasing number of researchers have paid attention to eRNAs in the mediation of gene transcription (Natoli and Andrau, 2012; Andersson et al., 2014). Enhancer RNAs can

TABLE 1 | The clinical parameters in TCGA gastric cancer cohort.

Covariates	Type	Percent
Gender	Female	134 (35.73%)
	Male	241 (64.27%)
Age	≤60y	121 (32.61%)
	>60y	250 (67.39%)
Grade	G1	10 (2.67%)
	G2	137 (36.53%)
	G3	219 (58.4%)
	Unknown	9 (2.4%)
M_stage	M0	330 (88%)
	M1	25 (6.67%)
	Unknown	20 (5.33%)
N_stage	N0	111 (29.6%)
	N1	97 (25.87%)
	N2	75 (20%)
	N3	74 (19.73%)
	Unknown	18 (4.8%)
T_stage	T1	19 (5.07%)
	T2	80 (21.33%)
	T3	168 (44.8%)
	T4	100 (26.67%)
	Unknown	8 (2.13%)
Clinical Stage	Stage II	111 (29.6%)
	Stage III	150 (40%)
	Stage IV	38 (10.13%)
	Stage I	53 (14.13%)
	Unknown	23 (6.13%)
Race	Asian	74 (19.73%)
	Black or African American	11 (2.93%)
	Native Hawaiian or other Pacific Islander	1 (0.27%)
	Unknown	51 (13.6%)
	White	238 (63.47%)

independently activate enhancer activity, and cooperate with other transcription-related factors to initiate the formation of the enhancer-promoter loop, thereby activating the expression of downstream genes and pathways (Kaikkonen et al., 2013; Kim and Shiekhattar, 2015). Dysregulation of eRNAs, specifically in the oncogenic signaling pathway, could result in the formation of a wide range of human cancers (Zhang et al., 2019). For instance, Kallikrein-related peptidase 3 eRNA in prostate cancer was found to promote the transcription of the downstream androgen receptor gene and promote cancer cell proliferation (Hsieh et al., 2014). Recent studies have also shown that the dysregulation of some eRNAs could serve as prognostic biomarkers in a range of cancers such as head and neck squamous cell carcinoma and lung and colon adenocarcinoma (Gu et al., 2019; Miracco et al., 2021; Xiao et al., 2021). The predictive factors of gastric cancer have not been fully identified, and the underlying functions associated with the tumor microenvironment (TME) cells and chemosensitivity.

In this study, we aim to identify potential prognostic eRNAs in gastric cancer, specifically focusing on cyclin-dependent kinase 6 (CDK6)-AS1 and its target gene. We performed pathway enrichment analyses to explore the potential function *CDK6-AS1* may have on tumorigenesis. Furthermore, we validated *CDK6-AS1* expression and the overall survival at a pan-cancer level and analyzed the correlation between eRNA, tumor-infiltrating immune cells, and drug sensitivity.

MATERIALS AND METHODS

Data Collection and Processing

Information from 33 datasets was downloaded from the University of California, Santa Cruz (UCSC) Xena The Cancer Genome Atlas (TCGA) hub (https://xena.ucsc.edu) (Goldman et al., 2020). The dataset included 407 gastric cancer tissues, 32 normal tissues, and 9951 other tumors from different types of cancers, and the RNA expression matrix was transformed to $\log_2$ (FPKM+1). Ensemble transcript IDs were converted to their corresponding GENCODE v19 using the Gene Transfer Format (GTF) annotation files from humans. The enhancer RNAs and target gene information were obtained from the putative literature, which was previously identified by the PreSTIGE method (Corradin et al., 2014).

Identification of Predictive eRNAs in Gastric Cancer

To avoid bias, patients with a survival time of less than 1 month were excluded. A total of 375 patients passed the quality control and were used in the following analysis as shown in **Table 1**. The survival-associated eRNAs were screened using the Cox regression model, with age, sex, and tumor stage adjusted as covariates. We set $p < 0.05$ as the significance cut-off value.

Analysis of Significantly Mutated Genes

The R package maftools were used to compare the mutant frequencies of significantly mutated genes between *CDK6-AS1* as high- and low-expression groups. Mutant types including frame shift, deletion, splice site, frameshift insertion, missense mutation, nonsense mutation, multiple hist, and in-frame deletion were considered in the analysis.

Gene Enrichment Analysis

Gene Ontology (GO) functional analysis was performed using the clusterProfiler package in R software, and a Reactome pathway analysis of eRNA-related coding genes was performed based on co-expression analysis. Specifically, the GO analysis revealed the function of the biology process (BP), cell component (CC), and molecular function (MF). To avoid accumulation of type-I errors, enrichment items meeting the false discovery rate (FDR) < 0.05 were considered significant.

Validation in TCGA Pan-Cancer Cohort

The expression data of *CDK6-AS1* and its target gene *CDK6* were obtained at the pan-cancer level as previously described, and patients were classified into low- and high-expression groups according to the median value of *CDK6-AS1* expression. The Cox regression method was used to compare the overall survival difference between the two groups. Covariates of sex, age, and

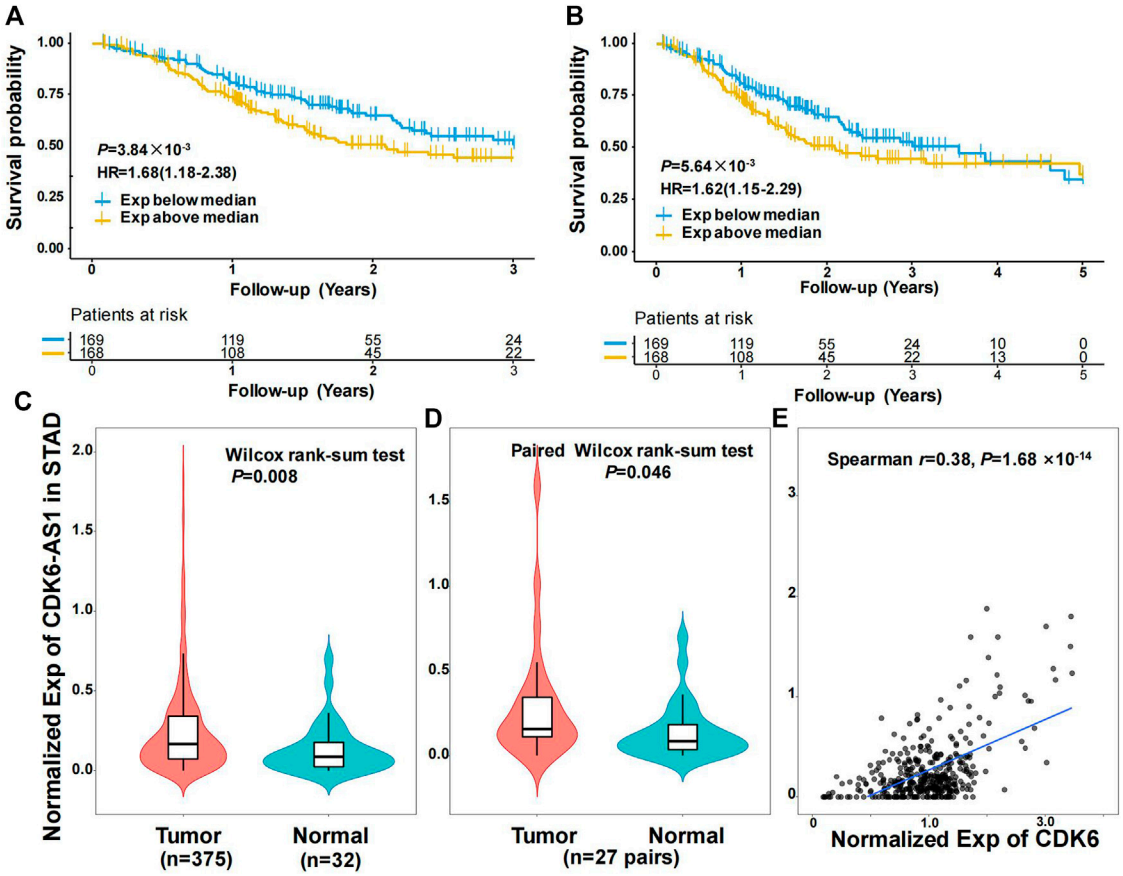


FIGURE 1 | Characteristics of eRNA *CDK6-AS1* in gastric cancer. **(A)** Kaplan–Meier 3-year overall survival curve for gastric cancer patients with *CDK6-AS1* low and high expression. **(B)** Kaplan–Meier 5-year overall survival curve for gastric cancer patients with *CDK6-AS1* low and high expression. **(C)** Differential expression of *CDK6-AS1* between unpaired tumor and adjacent normal tissues. **(D)** Differential expression of *CDK6-AS1* between paired tumor and adjacent normal tissues. **(E)** The correlation between the *CDK6-AS1* and its target gene, *CDK6* expression levels.

tumor stage were adjusted in the Cox model, and Spearman's coefficient was applied to the correlation analysis.

Construction and Validation of CDK6-AS1-Related Prognosis

Univariate Cox regression analysis and Kaplan-Meier analysis were used to screen 6 genes co-expressed with CDK6-AS1. Gastric cancer patients in the TCGA data set were randomly divided into the training set and internal test set. The aforementioned six genes were used for the LASSOCox regression analysis. By using the cross-verification error curve, the best tuning parameter λ is selected through the minimum 10-fold cross-verification in the training set. Based on these six genes, a risk-scoring model is established. Risk score = 0.2637*CTHRC1 + 0.0132*PFN2 + 0.1384*PRSS35 + 0.0355*RTN4 - 0.072*SMPD3 + 0.5459*SYCP2L.

GC Patients in the internal test set and an external cohort were divided into a high-risk group and a low-risk group by the optimal cut-off value of the risk score. The overall survival (OS) rates between the high-risk and the low-risk groups were analyzed by the Kaplan–Meier OS analysis. A two-sided log-rank $p < 0.05$ was considered significant. The time-dependent prognostic value of the prognostic signature was evaluated using the R package"survival ROC." Area under the curve (AUC) values were used to evaluate the time-dependent prognostic values of the prognostic signature. An AUC >0.60 was considered to be acceptable.

Analysis of Immune Cell Infiltrates

To evaluate the relationship between tumor-infiltrating lymphocytes (TIL) and the expression of *CDK6-AS1* in gastric cancer, we estimated the expressed fraction of TIL cells using the ssGSEA algorithm by comparing the gastric cancer gene expression matrix with those of the signatures from nine reported TIL cell types (Li et al., 2016). The relationship of the proportion matrix for the nine TIL cells with *CDK6-AS1* was calculated by Spearman's correlation analysis.

Prediction of Chemosensitivity

The R package pRRophetic (Geeleher et al., 2017), based on the pharmacogenomics database of the Cancer Genome Project

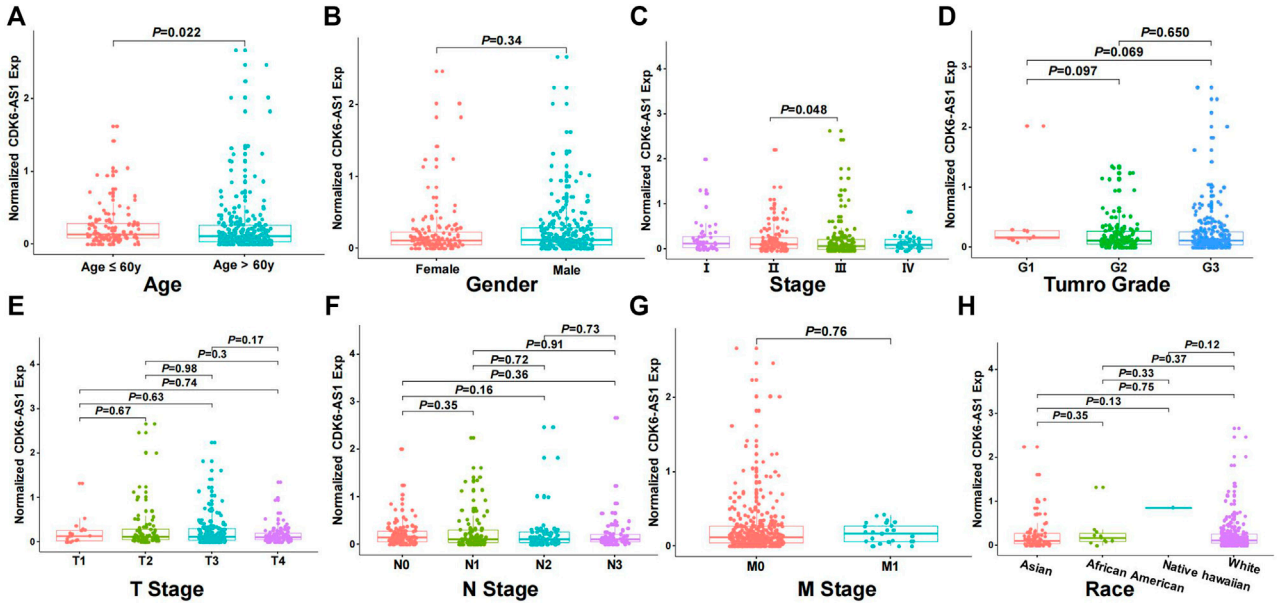


FIGURE 2 | The relationship between *CDK6-AS1* expression and clinical features. **(A–H)** The expression of *CDK6-AS1* among patients with age ≥60 years and <60 years, male and female, I–IV clinical stages, G 1–3 grades, T 1–4 stages, N 0–3 stages, M 0–1 stages, and different races.

(CGP) cell line data and the Cancer Cell Line Encyclopedia (CCLE), was used to predict chemotherapeutic sensitivity for gastric cancer by an estimation of IC50 (the maximal inhibitory concentration). Default settings were used for the prediction model, including "stomach cancer" for reference tissue type and "cvFold = 10" for ridge regression model training.

Statistical Analysis

R software (Version 3.6.2) was used to perform analyses in this study. The statistical results are expressed as mean ± standard deviation (M ± SD), and the data comparison of the two groups was analyzed with the Wilcoxon rank-sum test. A value of $p < 0.05$ was used to determine the statistical significance.

RESULTS

Screening of Key eRNAs in Gastric Cancer

Twenty-three eRNAs were identified, eight of which met the criteria (Spearman r ≥ 0.3 and *FDR* < 0.05) and were included (**Supplementary Table S1**). Of these, *CDK6-AS1* exhibited the lowest Cox model *p*-value and was therefore considered a candidate marker. Patients in the *CDK6-AS1* high-expression group had a shorter survival than those in the low-expression group (3-year OS:HR = 1.68, $p = 3.84 \times 10^{-3}$; 5-year OS:HR = 1.62, $p = 5.64 \times 10^{-3}$, **Figures 1A,B**). In addition, *CDK6-AS1* shows a higher expression in unpaired and paired tumor tissues compared to normal tissues (unpaired: $p = 8.00 \times 10^{-3}$, paired: $p = 0.046$, **Figures 1C,D**). A positive correlation between *CDK6-AS1* and its target gene CDK6 was observed (Spearman r = 0.38, $p = 1.68 \times 10^{-14}$). The connections between the clinical features of gastric cancer patients and the *CDK6-AS1* expression were further investigated. It was found that *CDK6-AS1* had a higher expression in patients aged <60 years ($p = 0.022$, **Figure 2A**). *CDK6-AS1* was significantly linked to the clinical stage (III vs. II, $p = 0.048$, **Figure 2C**). Other clinical characteristics were not clearly correlated with the *CDK6-AS1* expression ($p > 0.05$, **Figures 2B, D–H**). As driver gene mutations are crucial to tumor growth, the frequencies of significantly mutated genes were compared between the high- and low-*CDK6-AS1* expression groups. It was noted that several classic gastric cancer driver genes were more frequently mutated in the low-*CDK6-AS1* expression group than in the high-expression group, such as ARID1A and PIK3CA (**Figures 3A,B**, **Supplementary Table S2**).

Pathway Enrichment Analysis of *CDK6-AS1* Co-Expressed Genes

To further explore the function and related pathways *CDK6-AS1* involved in gastric cancer, a co-expression analysis between *CDK6-AS1* and other protein-coding genes in 375 TCGA gastric cancer cases was performed. It was found that 595 transcripts presented a significant correlation with *CDK6-AS1* (Spearman r > 0.30 & FDR <0.05). GO and Reactome enrichment analyses were performed. Results from TOP10GO pathways in BP, MF, and CC are shown in **Figure 4A**, **Supplementary Table S3**. In BP, the terms were mainly related to RNA transportation. In MF, the terms were related to exoribonuclease activity. In CC, the term is involved in the nuclear chromosomal region. Enrichment from the Reactome pathway database indicated that *CDK6-AS1* related co-expressed genes were mainly involved in the cell cycle and mitosis (**Figure 4B**, **Supplementary Table S5**), and key signals, in tumor cell proliferation, as shown by Spearman's correlation (r > 0.3 are

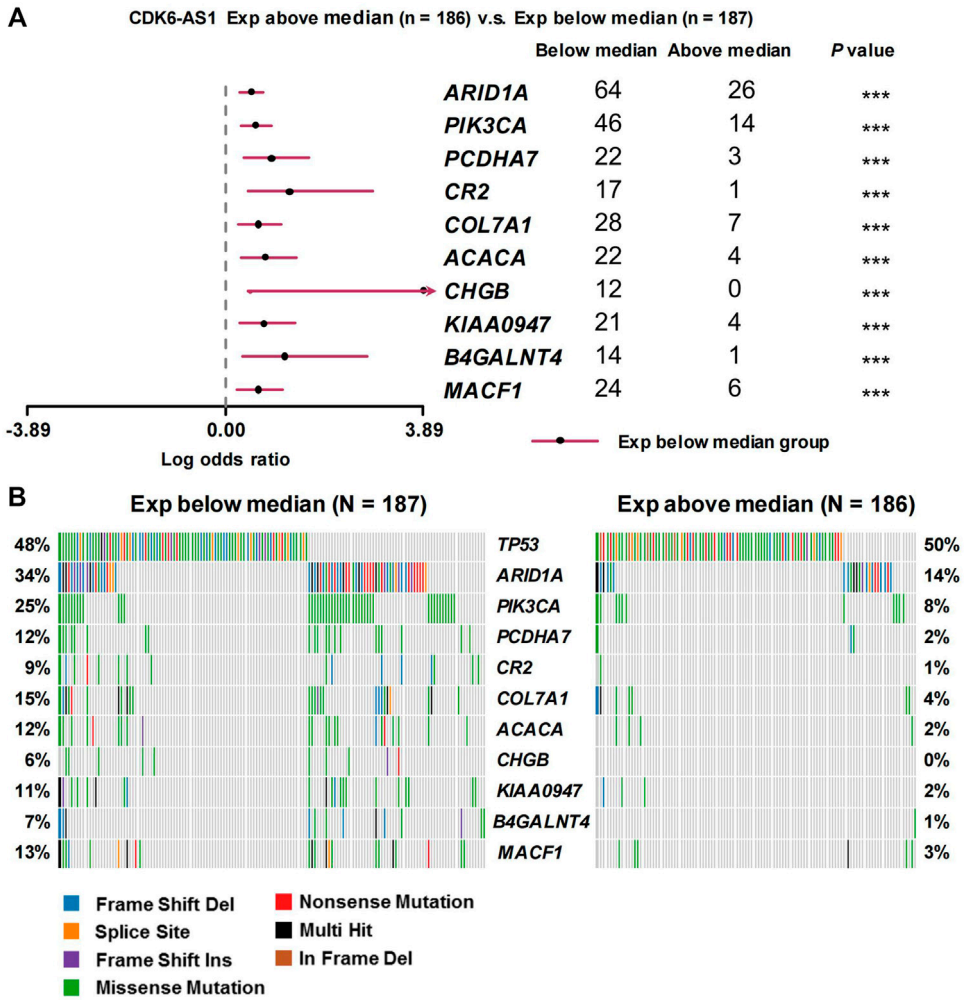


FIGURE 3 | Frequencies of mutated genes between different *CDK6-AS1* expression group. **(A)** Forest plot showing the top frequencies of mutated genes in the high- and low-*CDK6-AS1* expression groups. **(B)** Waterfall plot side by side for comparison of different mutated genes and mutated types between the high- and low-*CDK6-AS1* expression groups.

showed in **Supplementary Table S5**). Taken together, *CDK6-AS1* and its related genes may be involved in gene transcription and cell cycle processes that are essential for malignant progression.

Pan-Cancer Analysis of *CDK6-AS1*

To determine *CDK6-AS1* expression, prognosis, and correlation with its target gene at the pan-cancer level, we analyzed 33 tumor cohorts from the TCGA database. *CDK6-AS1* displayed higher expression in tumor tissues than in adjacent normal tissues in 16 tumor types: COAD, DLBC, ESCA, GBM, HNSC, KIRP, LAML, LIHC, LUSC, OV, PAAD, PCPG, READ, SARC, THYM, and UCS. Meanwhile, six types showed higher expression in normal tissues than in malignant ones: ACC, BRCA, KICH, PCPG, and TGCT (**Figure 5A**). In terms of survival analysis, high expression of *CDK6-AS1* was related to poor prognosis in BLCA, HNSC, KIRC, LGG, LUAD, MESO, and THCA (**Figures 5B–H**), while showing better prognosis in UVM (**Figure 5I**). Further analysis was performed to determine the correlation between *CDK6-AS1* expression and its target gene, *CDK6*. We found that *CDK6-AS1* was correlated with its target in 29 tumor types (**Supplementary Table S6**). Taken together, when *CDK6-AS1* is dysregulated it could influence the prognosis in a range of different cancer types. Representative immunohistochemical staining was used for gastric cancer tumor-infiltrating target gene *CDK6*. Scale bar, 50 mm (**Figure 5J**). Western blot results showed that CDK6 was significantly expressed in metastatic gastric cancer cells MKN-45 and in the *in situ* gastric cancer cell line HGC-27, while GES-1 was not significantly expressed in normal gastric mucosa epithelial cells (**Figure 5K**).

Construction and Verification of Prognostic Features Related to CDK6-AS1

In order to identify predictive genes and construct a prognostic model, six genes related to DEG and OS co-expressed with CDK6-AS1 were crossed to obtain 6 differentially expressed

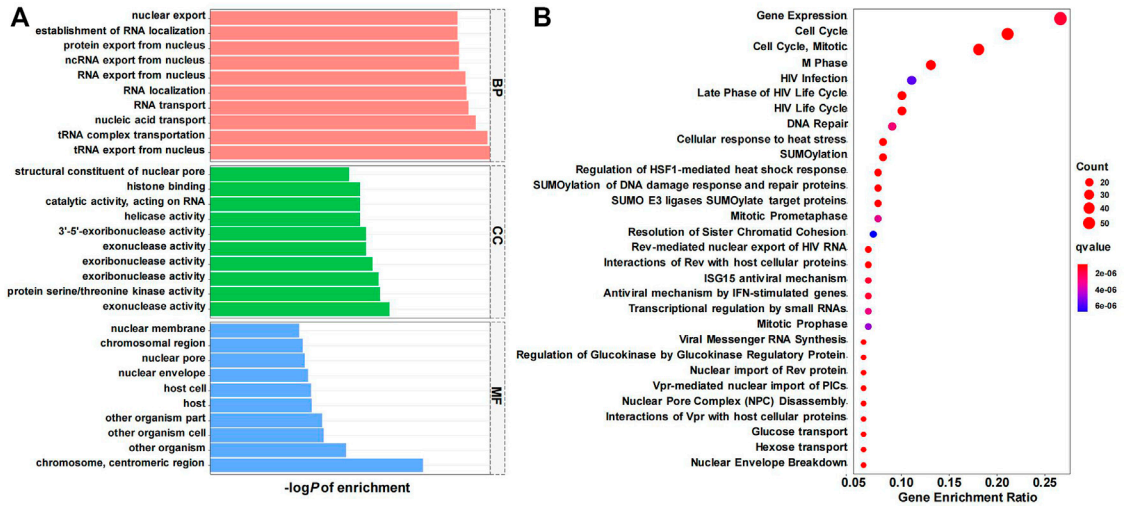


FIGURE 4 | Pathway enrichment analysis of *CDK6-AS1* co-expressed genes. **(A)** GO (Gene Ontology) enrichment result in biological process (BP), cellular component (CC), and molecular function (MF) categories. **(B)** Reactome enrichment result. Top 30 items were shown.

genes related to CDK6-AS1 and OS. Then the six genes were analyzed by LASSOCox regression and the tuning parameter lambda (λ) was selected by using the cross-validation error curve. The prognostic models were constructed when the λ value was minimum (**Figure 6A**) (Risk score = 0.2637*CTHRC1 + 0.0132*PFN2 + 0.1384*PRSS35 + 0.0355*RTN4 - 0.072*SMPD3 + 0.5459*SYCP2L, and their LASSO coefficient curves are shown in **Figure 6B**. The relationship between survival status/risk score, mRNA expression heat map of 6 genes, and survival time (days)/risk score showed that the prognostic model had a good prognostic effect, and the OS of gastric cancer patients in the high-risk group was worse than that in the low-risk group ($p < 0.001$) (**Figures 6C–F**).

Correlation Between *CDK6-AS1* and TIL

TIL are important players in the TME and have been reported to influence gastric cancer survival rates. Therefore, the association between *CDK6-AS1* expression and the proportion of infiltrating lymphocytes was analyzed. Higher *CDK6-AS1* expression showed negative tendencies with almost all inferred immune cell enrichment scores (**Figures 7A–I**) and was among Th_{elper} (Spearman r = −0.20, e < 0.001, **Figure 7C**), T_{reg} (Spearman r = −0.10, e = 0.044, **Figure 7D**), and neutrophil cell clusters (Spearman's r = −0.14, $p = 0.008$, **Figure 7I**). Taken together, it appears that high-*CDK6-AS1* expression may hamper immune TME cell infiltration, both in innate and adaptive cell clusters, which may have an effect on anti-tumor immunity.

Prediction of Drug Sensitivity

Since chemo-sensitivity or -resistance is related to the clinical prognosis of gastric cancer, we explored the chemosensitivity of the high- and low-*CDK6-AS1* expression groups. The ridge regression model was used to predict individual drug sensitivities. A commonly used chemotherapy drug in gastric cancer therapy, cisplatin, showed greater sensitivity in the high-*CDK6-AS1* expression group than in the low-expression group ($p = 0.024$, **Figure 8A**). Conversely, paclitaxel showed higher sensitivity in the low-expression group ($p = 0.040$). We also explored the expression of immunotherapy and targeted therapeutic markers. PD-L1 expression was higher in the low-*CDK6-AS1* expression group, which may indicate anti-PD-L1 therapy (**Figure 8B**). We can also prove that CDK6-AS1 is related to immunotherapy by TMB analysis, and the results show that the immunotherapy effect is better in the group with low expression of CDK6-AS1 (**Figure 8C**).

DISCUSSION

In our study, we found that *CDK6-AS1* may serve as an independent poor prognostic biomarker candidate in gastric cancer, with a positive correlation to its target gene, *CDK6*. The low-*CDK6-AS1* expression group showed more frequently mutated driver genes than the high-expression group. Moreover, *CDK6-AS1* is involved in key oncogenic pathways such as the cell cycle and RNA transcription. *CDK6-AS1* also shows dysregulation and is associated with prognosis at the pan-cancer level. We also verified by constructing a prognostic model and found that the model had a good prognostic effect, and the OS of patients with gastric cancer in the high-risk group was worse than that in the low-risk group.

Some researchers had demonstrated that eRNAs worked by regulating target genes to form a chromatin loop (Natoli and Andrau, 2012; Bresnick and Johnson, 2019). Research has shown that the eRNA of ACTRT1 can lower the expression of target genes and promote the development of cancer (Bal et al., 2017).

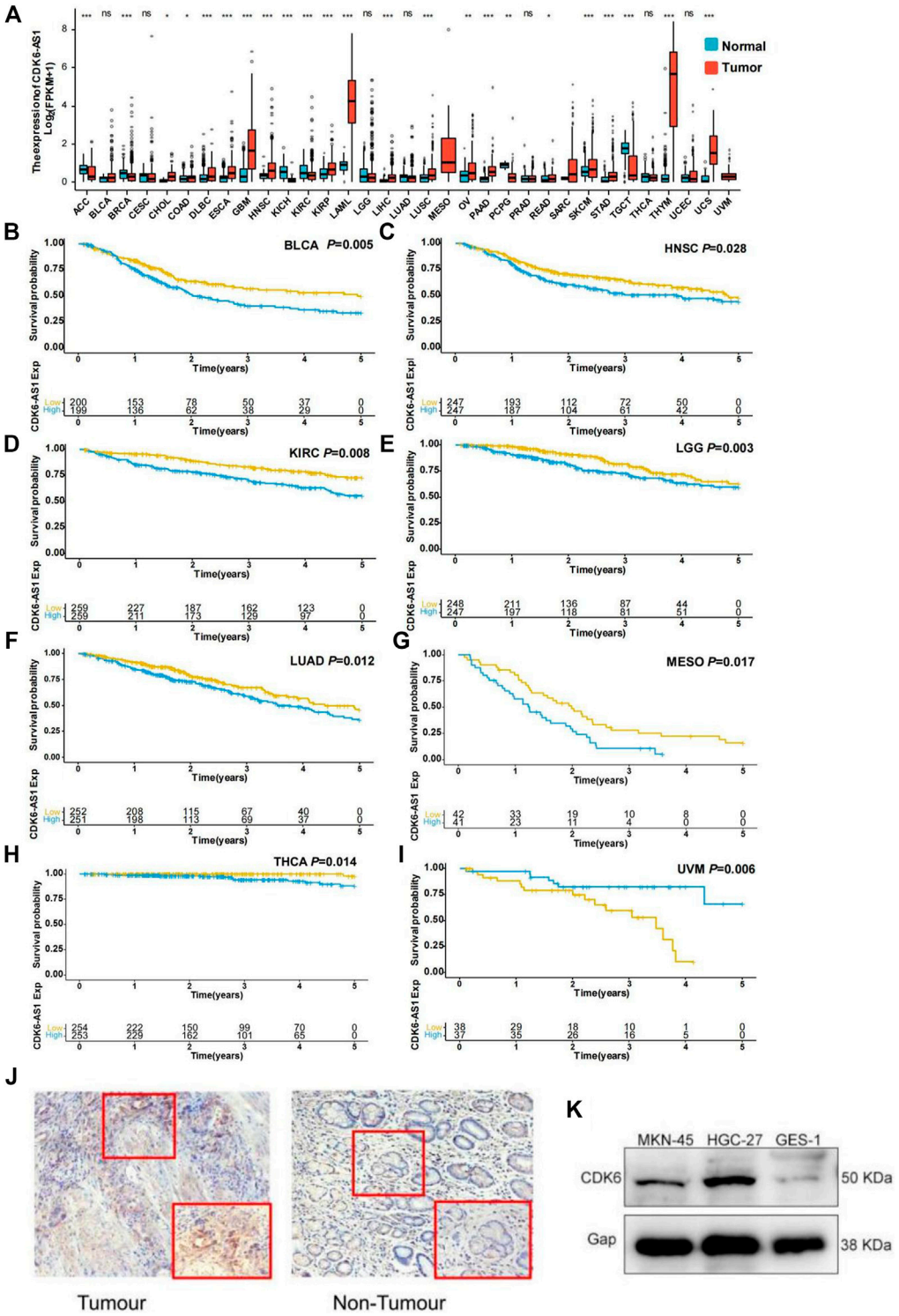


FIGURE 5 | Expression and survival validation in TCGA pan-cancer cohort. **(A)** Differential expression validation of *CDK6-AS1* in TCGA pan-cancer cohort. *p e 0.05; **$p < 0.01$; ***$p < 0.001$; ns: not significant with $p > 0.05$. **(B–I)** The prognostic effect of *CDK6-AS1* in other TCGA cancer types. Types with Cox $p > 0.05$ are shown. **(J)** Representative immunohistochemical staining for gastric cancer tumor-infiltrating target gene *CDK6*. **(K)** The *CDK6* was expressed in gastric cancer cells MKN-45, HGC-27, and gastric mucosa epithelial cells GES- 1.

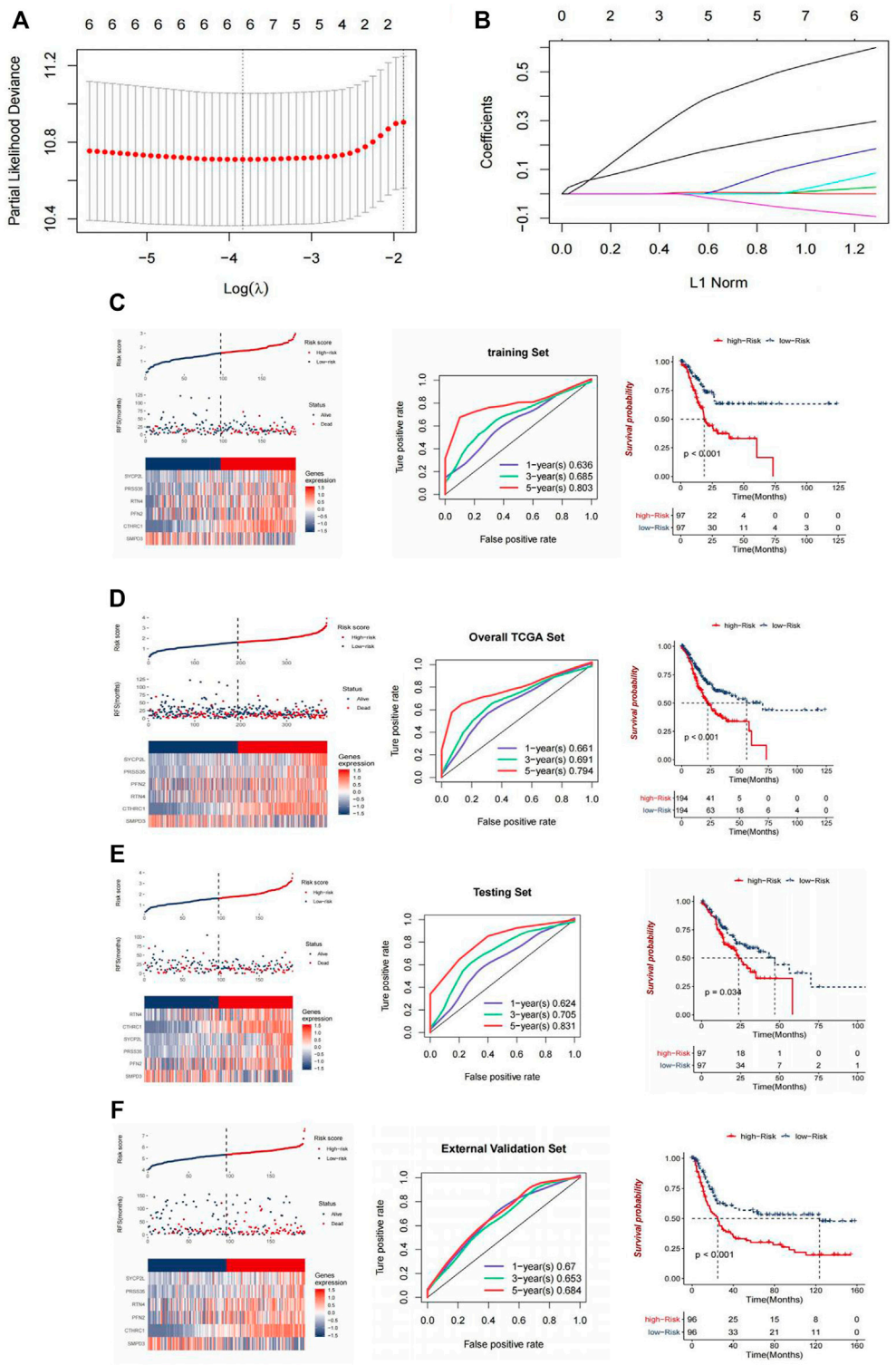


FIGURE 6 | Construction and validation of CDK6-AS1. **(A)** Tuning parameter lambda (λ) selected by cross-validation error curve. **(B)** LASSO coefficient profiles of six genes. **(C,D)** Relationship between the survival status/risk score rank, mRNA expression heat map of 6 genes and survival time (days)/risk score rank. Left: survival status/risk score rank, mRNA expression heat map of 6 genes. Middle: Time-dependent ROC curves for OS of the TCGA-LAML training data set. The AUC was assessed at 1, 3, and 5 years. Right: Kaplan-Meier OS analysis of gastric cancer patients in low-risk and high-risk groups. *p* value was calculated using the log-rank test. $p < 0.001$. **(E,F)** Validation using an internal set **(E)** and an external set **(F)**.

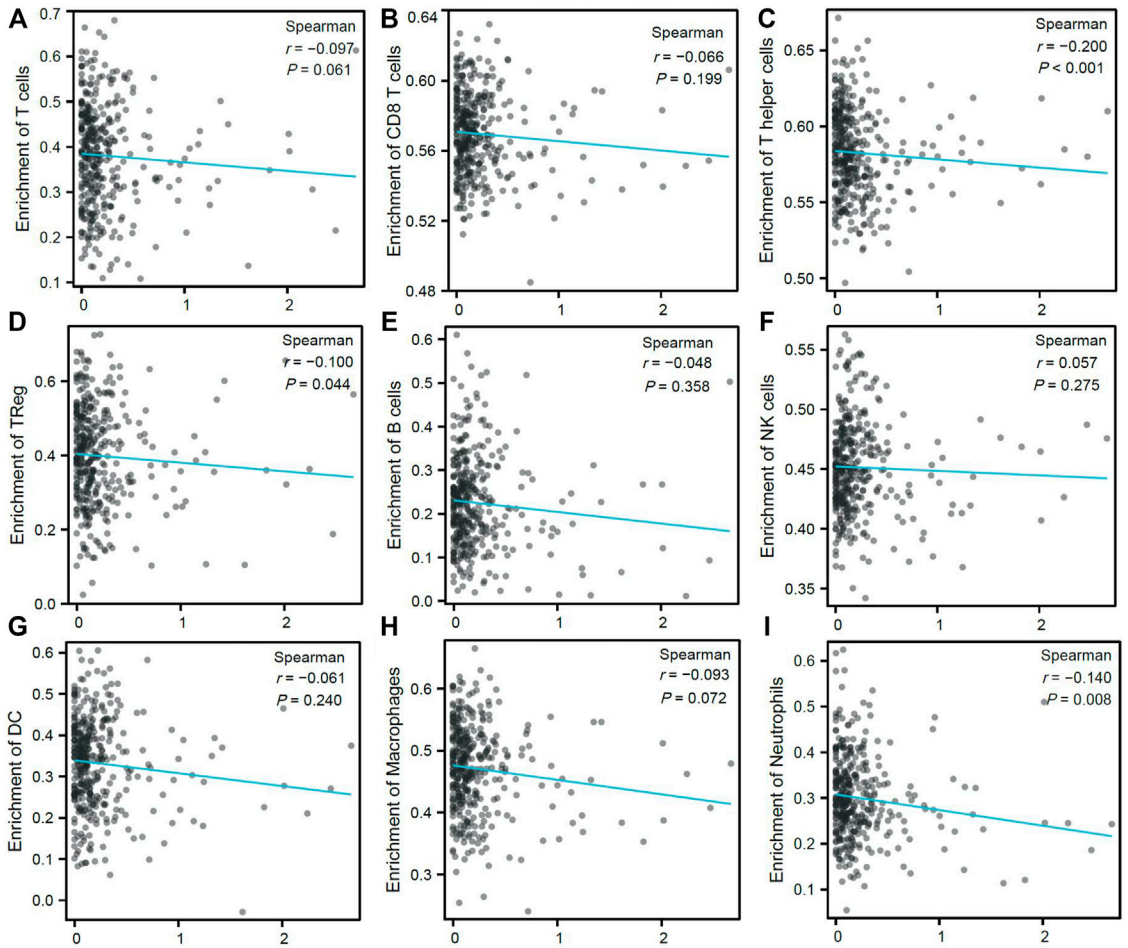


FIGURE 7 | Correlation between expression of *CDK6-AS1* and immune cell infiltration proportion. **(A–I)** Correlation between expression of *CDK6-AS1* and ssGSEA inferred immune cell infiltration of T, CD8+T, T helper, T_{reg}, B, NK, DC, macrophage, and neutrophil cells.

In our study, it was also confirmed that *CDK6-AS1* was associated with survival in eight types of tumors (BLCA, HNSC, KIRC, LGG, LUAD, MESO, THCA, and UVM), and moreover, *CDK6-AS1* expression was correlated with that of its target gene, *CDK6*, in 29 tumor types. Thus, we suggest that *CDK6-AS1* acts as an independent predictor of gastric cancer.

CDK6 can form complex D-type cyclins (D1, D2, and D3) and progresses to the early G1 phase (Nebenfuehr et al., 2020). CDK4/6-cyclin D-complexes are regulated by Cip/Kip proteins, which affect the nuclear translocation of complexes (Song et al., 2020; Nardone et al., 2021). Some studies have reported that CDK6 overexpression could affect lymphoma, leukemia, and other malignancies. Although there is no relationship between mutations in CDK6 and diseases, CDK6 has served as a hub gene in acute myeloid leukemia (Malumbres and Barbacid, 2009; Scheicher et al., 2015). The cyclin D-CDK4/6 axis is commonly expressed in breast cancer. Another effect of CDK4/6 inhibitors is anti-tumor immunity. Zhang et al. suggested that CDK4 negatively regulates programmed cell death ligand 1 (PD-L1) protein stability; moreover, CDK4 and PD-L1 levels negatively correlate with tumor treatment (Scheicher et al., 2015; Goel et al., 2017).

Since TME is important during tumor progress, we studied *CDK6-AS1* expression in infiltrating immune cell fractions. Intriguingly, we found that lower *CDK6-AS1* expression was positively correlated with the proportion of antitumor immune cells, such as T_{helper} cells. This could be partially explained by the finding that the low-*CDK6-AS1* expression group had more mutated genes than the high-expression group. More frequently mutated genes indicate an increased mutational burden and cancer neoantigens (Xu et al., 2020). Neoantigens can serve as targets for immune recognition and recruitment (Garcia-Garijo et al., 2019). The relationship between *CDK6-AS1* and neoantigens warrants further study.

Our study had some limitations. First, the sample size of this research was small, and more clinical research and data are needed. Second, further research and greater detailing is required on the function and role of *CDK6-AS1* in gastric cancer.

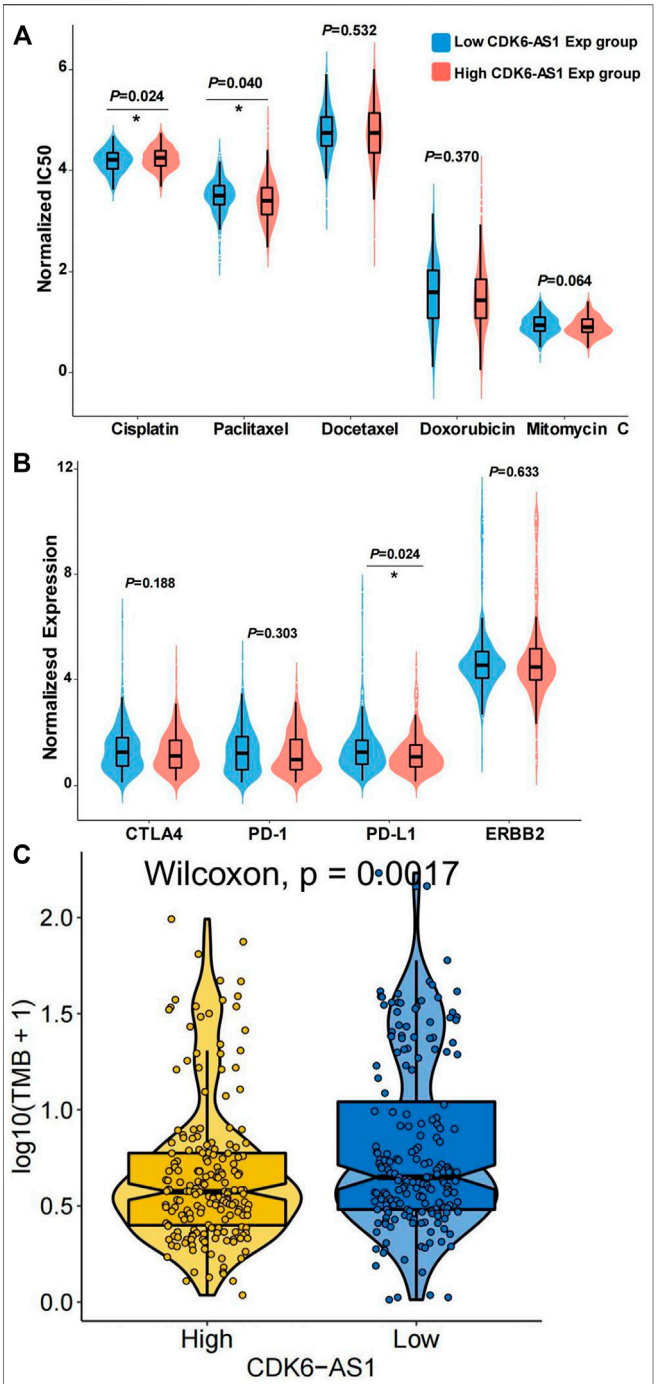


FIGURE 8 | Drug sensitivity prediction in low- and high-*CDK6-AS1* expression group. **(A)** Chemosensitivity prediction of five commonly used drugs in antineoplastic therapy in the low- and high-*CDK6-AS1* expression groups. IC50: Half-maximum inhibitory concentration. **(B)** Expression of immunotherapy and HER2-targeted therapy markers in the low and high *CDK6-AS1* expression groups. **(C)** TMB analysis showed that the immunotherapy effect is better in the group with low expression of *CDK6-AS1*.

CONCLUSION

CDK6-AS1 may serve as a poor independent prognostic biomarker candidate for gastric cancer, demonstrating a positive correlation with its target gene, *CDK6*. Moreover, *CDK6-AS1* is involved in key oncogenic pathways such as the cell cycle and RNA transcription. *CDK6-AS1* also shows dysregulation and is associated with prognosis at the pan-cancer level. These eRNAs may also be associated with immune cell infiltration and drug sensitivity.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Second Affiliated Hospital of Harbin Medical University. The patients/participants provided their written informed consent to participate in this study. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SY and XZ designed the study. HY, JL and AZ performed the experiment. LEZ, and ZM analyzed the data. SY and LIZ wrote the manuscript. All authors reviewed the manuscript.

FUNDING

This project was supported by the National Natural Science Foundation of China (81672355).

ACKNOWLEDGMENTS

The authors want to thank all the participants in the research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.854211/full#supplementary-material

REFERENCES

Andersson, R., Gebhard, C., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., et al. (2014). An Atlas of Active Enhancers across Human Cell Types and Tissues. *Nature* 507 (7493), 455–461. doi:10.10.1038/nature12787

Bal, E., Park, H.-S., Belaid-Choucair, Z., Kayserili, H., Naville, M., Madrange, M., et al. (2017). Mutations in ACTRT1 and its Enhancer RNA Elements lead to Aberrant Activation of Hedgehog Signaling in Inherited and Sporadic Basal Cell Carcinomas. *Nat. Med.* 23 (10), 1226–1233. doi:10.1038/nm.4368

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer J. Clinicians* 68 (6), 394–424. doi:10.3322/caac.21492

Bresnick, E. H., and Johnson, K. D. (2019). Blood Disease-Causing and -suppressing Transcriptional Enhancers: General Principles and GATA2 Mechanisms. *Nat. Commun.* 3 (13), 2045–2056. doi:10.1182/bloodadvances.2019000378

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal·lari, R., et al. (2014). Combinatorial Effects of Multiple Enhancer Variants in Linkage Disequilibrium Dictate Levels of Gene Expression to Confer Susceptibility to Common Traits. *Genome Res.* 24 (1), 1–13. doi:10.1101/gr.164079.113

Digklia, A., and Wagner, A. D. (2016). Advanced Gastric Cancer: Current Treatment Landscape and Future Perspectives. *Wjg* 22 (8), 2403–2414. doi:10.3748/wjg.v22.i8.2403

Garcia-Garijo, A., Fajardo, C. A., and Gros, A. (2019). Determinants for Neoantigen Identification. *Front. Immunol.* 10, 1392. doi:10.3389/fimmu.2019.01392

Geeleher, P., Zhang, Z., Wang, F., Gruener, R. F., Nath, A., Morrison, G., et al. (2017). Discovering Novel Pharmacogenomic Biomarkers by Imputing Drug Response in Cancer Patients from Large Genomics Studies. *Genome Res.* 27 (10), 1743–1751. doi:10.1101/gr.221077.117

Goel, S., DeCristo, M. J., Watt, A. C., BrinJones, H., Sceneay, J., Li, B. B., et al. (2017). CDK4/6 Inhibition Triggers Anti-tumour Immunity. *Nature* 548 (7668), 471–475. doi:10.1038/nature23465

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and Interpreting Cancer Genomics Data via the Xena Platform. *Nat. Biotechnol.* 38 (6), 675–678. doi:10.1038/s41587-020-0546-8

Gu, X., Wang, L., Boldrup, L., Coates, P. J., Fahraeus, R., Sgaramella, N., et al. (2019). AP001056. 1, A Prognosis-Related Enhancer RNA in Squamous Cell Carcinoma of the Head and Neck. *Cancers (Basel)* 11 (3), 347. doi:10.3390/cancers11030347

Hsieh, C.-L., Fei, T., Chen, Y., Li, T., Gao, Y., Wang, X., et al. (2014). Enhancer RNAs Participate in Androgen Receptor-Driven Looping that Selectively Enhances Gene Activation. *Proc. Natl. Acad. Sci. U.S.A.* 111 (20), 7319–7324. doi:10.1073/pnas.1324151111

Kaikkonen, M. U., Spann, N. J., Heinz, S., Romanoski, C. E., Allison, K. A., Stender, J. D., et al. (2013). Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. *Mol. Cel.* 51 (3), 310–325. doi:10.1016/j.molcel.2013.07.010

Kim, T.-K., and Shiekhattar, R. (2015). Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* 162 (5), 948–959. doi:10.10.1016/j.cell.2015.08.008

Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., et al. (2016). Comprehensive Analyses of Tumor Immunity: Implications for Cancer Immunotherapy. *Genome Biol.* 17 (1), 174. doi:10.1186/s13059-016-1028-7

Malumbres, M., and Barbacid, M. (2009). Cell Cycle, CDKs and Cancer: a Changing Paradigm. *Nat. Rev. Cancer* 9 (3), 153–166. doi:10.1038/nrc2602

Martens-Uzunova, E. S., Böttcher, R., Croce, C. M., Jenster, G., Visakorpi, T., and Calin, G. A. (2014). Long Noncoding RNA in Prostate, Bladder, and Kidney Cancer. *Eur. Urol.* 65 (6), 1140–1151. doi:10.1016/j.eururo.2013.12.003

Miracco, C., Giannicola, R., Giordano, A., Pirtoli, L., Cheng, L., Han, T., et al. (2021). TBX5-AS1, an Enhancer RNA, Is a Potential Novel Prognostic Biomarker for Lung Adenocarcinoma. *Int. J. ofmolecular Sci.* 21 (1), 794. doi:10.1186/s12885-021-08517-w

Nardone, V., Barbarino, M., Angrisani, A., Correale, P., Pastina, P., Cappabianca, S., et al. (2021). CDK4, CDK6/cyclin-D1 Complex Inhibition and Radiotherapy for Cancer Control: A Role for Autophagy. *Ijms* 22 (16), 8391. doi:10.3390/ijms22168391

Natoli, G., and Andrau, J.-C. (2012). Noncoding Transcription at Enhancers: General Principles and Functional Models. *Annu. Rev. Genet.* 46, 1–19. doi:10.1146/annurev-genet-110711-155459

Nebenfuehr, S., Kollmann, K., and Sexl, V. (2020). The Role of CDK6 in Cancer. *Int. J. Cancer* 147 (11), 2988–2995. doi:10.1002/ijc.33054

Scheicher, R., Hoelbl-Kovacic, A., Bellutti, F., Tigan, A.-S., Prchal-Murphy, M., Heller, G., et al. (2015). CDK6 as a Key Regulator of Hematopoietic and Leukemic Stem Cell Activation. *Blood* 125 (1), 90–101. doi:10.1182/blood-2014-06-584417

Soerjomataram, I., Lortet-Tieulent, J., Parkin, D. M., Ferlay, J., Mathers, C., Forman, D., et al. (2012). Global burden of Cancer in 2008: a Systematic Analysis of Disability-Adjusted Life-Years in 12 World Regions. *The Lancet* 380 (9856), 1840–1850. doi:10.1016/s0140-6736(12)60919-2

Song, Z., Wei, Z., Wang, Q., Zhang, X., Tao, X., Wu, N., et al. (2020). The Role of DOT1L in the Proliferation and Prognosis of Gastric Cancer. *Biosci. Rep.* 40 (1), 40. doi:10.1042/BSR20193515

The Cancer Genome Atlas Research Network (2014). Comprehensive Molecular Characterization of Gastric Adenocarcinoma. *Nature* 513 (7517), 202–209. doi:10.1038/nature13480

Thrift, A. P., and El-Serag, H. B. (2020). Burden of Gastric Cancer. *Clin. Gastroenterol. Hepatol.* 18 (3), 534–542. doi:10.1016/j.cgh.2019.07.045

Xiao, J., Liu, Y., Yi, J., and Liu, X. (2021). LINC02257, an Enhancer RNA of Prognostic Value in Colon Adenocarcinoma, Correlates with Multi-Omics Immunotherapy-Related Analysis in 33 Cancers. *Front. Mol. Biosci.* 8, 646786. doi:10.3389/fmolb.2021.646786

Xu, P., Luo, H., Kong, Y., Lai, W.-F., Cui, L., and Zhu, X. (2020). Cancer Neoantigen: Boosting Immunotherapy. *Biomed. Pharmacother.* 131, 110640. doi:10.1016/j.biopha.2020.110640

Zhang, Z., Lee, J. H., Ruan, H., Ye, Y., Krakowiak, J., Hu, Q., et al. (2019). Transcriptional Landscape and Clinical Utility of Enhancer RNAs for eRNA-Targeted Therapy in Cancer. *Nat. Commun.* 10 (1), 4562. doi:10.1038/s41467-019-12543-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Yang, Zou, Yang, Li, Zhang, Zhang, Li, Zhu and Ma. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Toxicology

ORIGINAL RESEARCH
published: 09 May 2022
doi: 10.3389/ftox.2022.821116

Comparative Toxicotranscriptomics of Single Cell RNA-Seq and Conventional RNA-Seq in TCDD-Exposed Testicular Tissue

Alex Haimbaugh[1], *Danielle Meyer*[1,2,3], *Camille Akemann*[1], *Katherine Gurdziel*[4†] *and Tracie R. Baker*[1,2,3*†]

[1]*Department of Pharmacology, School of Medicine, Wayne State University, Detroit, MI, United States,* [2]*Institute of Environmental Health Sciences, Wayne State University, Detroit, MI, United States,* [3]*Department of Environmental and Global Health, University of Florida, Gainesville, FL, United States,* [4]*Genome Sciences Core, Office of the Vice President for Research, Wayne State University, Detroit, MI, United States*

OPEN ACCESS

Edited by:
Yuhong Fan,
Georgia Institute of Technology, United States

Reviewed by:
Ramji K. Bhandari,
University of North Carolina at Greensboro, United States
Kevin Gerrish,
National Institute of Environmental Health Sciences (NIH), United States

***Correspondence:**
Tracie Baker
tracie.baker@ufl.edu

†These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Toxicogenomics, a section of the journal Frontiers in Toxicology

Received: *23 November 2021*
Accepted: *03 March 2022*
Published: *09 May 2022*

Citation:
Haimbaugh A, Meyer D, Akemann C, Gurdziel K and Baker TR (2022) Comparative Toxicotranscriptomics of Single Cell RNA-Seq and Conventional RNA-Seq in TCDD-Exposed Testicular Tissue. Front.Toxicology 4:821116. doi: 10.3389/ftox.2022.821116

In this report, we compare the outcomes and limitations of two methods of transcriptomic inquiry on adult zebrafish testes exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) during sexual differentiation: conventional or bulk RNA-seq (bulk-seq) and single cell RNA sequencing (scRNA-seq) data. scRNA-seq has emerged as a valuable tool for uncovering cell type-specific transcriptome dynamics which exist in heterogeneous tissue. Our lab previously showed the toxicological value of the scRNA-seq pipeline to characterize the sequelae of TCDD exposure in testes, demonstrating that loss of spermatids and spermatozoa, but not other cell types, contributed to the pathology of infertility in adult male zebrafish exposed during sexual differentiation. To investigate the potential for technical artifacts in scRNA-seq such as cell dissociation effects and reduced transcriptome coverage, we compared bulk-sequenced and scRNA-seq-paired samples from control and TCDD-exposed samples to understand what is gained and lost in scRNA-seq vs bulk-seq, both transcriptomically and toxicologically. We hypothesized that the testes may be sensitive to tissue disruption as they contain multiple cell types under constant division and/or maturation, and that TCDD exposure may mediate the extent of sensitivity. Thus, we sought to understand the extent to which this dissociation impacts the toxicological value of data returned from scRNA-seq. We confirm that the required dissociation of individual cells from intact tissue has a significant impact on gene expression, affecting gene pathways with the potential to confound toxicogenomics studies on exposures if findings are not well-controlled and well-situated in context. Additionally, a common scRNA-seq method using cDNA amplified from the 3' end of mRNA under-detects low-expressing transcripts including transcription factors. We confirm this, and show TCDD-related genes may be overlooked by scRNA-seq, however, this under-detection effect is not mediated by TCDD exposure. Even so, scRNA-seq generally extracted toxicologically relevant information better than the bulk-seq method in the present study. This report aims to inform future experimental design for transcriptomic investigation in the growing field of toxicogenomics by demonstrating the differential information extracted from sequencing cells—despite being from the same tissue and

exposure scheme—is influenced by the specific protocol used, with implications for the interpretation of exposure-induced risk.

Keywords: single cell RNA-seq, TCDD, pseudo bulk-seq, bulk-seq, scRNA-seq dropout, exposure, transcriptome

INTRODUCTION

The transcriptome is the complete picture of all expressed genes in a cell and in what quantity, during a specific developmental stage and/or experimental condition. The field of transcriptomic research has expanded at a rapid rate since the first attempts in the 1990s, gaining granularity and precision with each new advance. What could first only tell us limited transcript abundance in a tissue (SAGE), can now impart entire transcriptomes in high-throughput studies with single-cell resolution. Transcriptome analysis is the most widely used tool in the field of toxicogenomics (TGx) to study gene (dys) regulation in a biological system following chemical exposure (Federico et al., 2020). Due to its importance, there has been a worldwide call for standardization of "omics" data from the TGx community, as guidelines for analysis have not been formally established (Federico et al., 2020). Successful TGx requires best practices in experimental design, data processing techniques, and validation assays in order to produce reliable transcriptomic data which can be efficiently interpreted to serve downstream analyses, such as safety assessments used in regulatory decisions.

With each new method of exploring the gene expression landscape at an increasingly granular level, there are advantages and disadvantages to take into account when planning an experiment. Some methods are more accessible and therefore quite common in labs, such as microarray and RNA-seq. The concept behind microarray is that RNA molecules in a sample can be reversed transcribed to cDNA, and these cDNA sequences can be captured, if present, by oligo probes on the microarray chip. This method certainly advanced the field as microarrays became quite detailed, encompassing entire genomes; gene batteries specifically for toxicology are even available (Lettieri, 2006). RNA sequencing (RNA-seq) uses the same concept of cDNA sequencing as microarray, but allows untargeted exploration of the transcriptomic landscape in a sample of any species for which the genome is annotated. Thus, the advent of RNA-seq revolutionized the field, heralding the next-generation sequencing (NGS) era of "discovery-driven" research. The more complete picture offered by RNA-seq allowed for more in-depth analysis of mechanism of action (MOA), pathway enrichment analysis, and, when combined with phenotypic endpoints, phenotypic anchoring such as biomarkers of disease or prognosis. Together, these valuable insights can inform chemical toxicity and risk assessments. However, despite the broadened opportunities RNA-seq allows, a major limitation to precisely identifying biological relevance with RNA-seq has been the processing of heterogeneous tissues as homogeneous entities, which produces a gene expression signature that is essentially a composite of the responses of the different cell types comprising the tissue. Science is well aware that various cells in a tissue can be highly differentiated and exhibit markedly different expression profiles. Toxicology can benefit from this more precise information as various cell types in diverse tissue can respond quite differently to the same toxicant (Hsu et al., 2020; Wang et al., 2021).

The newer transcriptomic analysis method of single cell RNA-Seq (scRNA-seq) (Tang et al., 2009) mitigates this issue by capturing the complex profiles of the singular cells constituting a tissue. There are various specific scRNA-seq methods, depending on the specific knowledge (Chen et al., 2019), but the overarching goal is acquiring the transcriptome of each cell in a tissue. scRNA-seq builds upon conventional RNA-seq (referred to as bulk-seq; representing the homogeneous, intact state of the tissue) research and is comparable in the basic concept. The sequencing that occurs at the single cell level is essentially the same process cells undergo in bulk-seq; the difference lies in the ability to computationally extricate the results for each cell rather than an averaged output of the entire sample. The advantageous difference of increased granularity is tantamount—scRNA-seq faithfully represents cellular heterogeneity by distinguishing gene expression profiles between different cell types, and even between the same cell type in different cellular states (Lähnemann et al., 2020). The advances in medicine made possible by scRNA-seq are invaluable, e.g., scRNA-seq enabled Zou et al. (2021) to develop a prognostic signature in gastric adenocarcinoma, Deng et al. (2020) identified characteristics predicting CAR T cell therapy efficacy. Toxicology can follow suit to develop targeted therapies for the myriad of multi-cell-type health problems associated with environmental exposure to chemicals, including cancer (Snyder, 2012), asthma (Sachdeva et al., 2019), and infertility (Canipari et al., 2020; Selvaraju et al., 2020).

With this exciting opportunity come inevitable challenges. While scRNA-seq data is structurally similar to bulk-seq data, differences in the tissue- and data preprocessing steps affect replicability, with scRNA-seq more prone to artifacts (e.g., technical or other non-biological sources of variation) (Federico et al., 2020). Contributing to artifacts are tissue dissociation, dropout, and complex quality control measures. The fact that scRNA-seq requires, by definition, single cells that must be individuated from the intact tissue of an organism potentially confounds results. This requirement impacts not only the number of cells available for analysis (those that survive dissociation), but also the physical separation of cells from their network can have an unwanted biological impact on surviving cells. While *in situ* sequencing (the act of sequencing RNA on, for example, a slide of tissue preserved in its natural environment) is possible, the throughput limitations render it undesirable. A more commonly used method for singularizing cells is enzymatic digestion, such as used in the

present study. The act of disruption from the native environment and the digestion process (which changes cellular shape among other facets) can potentially result in an altered cellular state, e.g., stress response, and thus findings unrepresentative of the actual cellular mechanisms *in vivo*. However, other methods of tissue dissociation that may prevent such microenvironmental shifts, such as laser capture microscopy and microdissection capillary pipette, are much lower-throughput (Hwang et al., 2018). Dropout events (failure to detect part or all of a transcriptome) and high (Ramsköld et al., 2012; Brennecke et al., 2013; Lun et al., 2016), can permeate due to low capture efficiency, the general stochastic nature of gene expression, and the very low input available (pg-ng) compared to bulk-seq (ng-g). Additionally, workflow pipelines for scRNA-seq and bulk-seq diverge in the complex QC measures required for scRNA-seq data that are not required in bulk-seq analysis. For example, "doublets" and poor quality cells (abnormally low read counts, high mitochondrial mRNA content) must be removed from analysis.

One way to parse artifacts from biological relevance is to validate scRNA-seq data by comparing it to bulk-seq data. In order to make scRNA-seq data comparable to bulk-seq data, a pseudo bulk-seq dataset can be derived from collapsing the reads counts from the various identified clusters (cell types) of scRNA-seq data to mimic *in silico* an average expression profile, such as that produced by bulk-seq on intact tissue (Lähnemann et al., 2020). Theoretically, the results from a pseudo bulk-seq dataset would resemble those from a bulk-seq dataset. However, technical artifacts are introduced by the act of sequencing cells on a microfluidic device rather than at once from a suspension of dissociated cells. A common 3′-end counting droplet-based method of scRNA-seq, 10x Genomics Chromium, is used here. With the expanding field of TGx and increased use of scRNA-seq, we were interested in determining what information was gained and lost compared to the more common, accessible bulk-seq method, in terms of transcriptome coverage, data analysis, and conclusions that can be drawn from both methods.

The field can benefit from seeing these data from different perspectives; apparent inconsistencies are not necessarily misleading and in fact can be quite informative (Qiu, 2020). However, understanding the sources of divergent information and their implications is critical. Our previous toxicological work with scRNA-seq in zebrafish (*Danio rerio*) testes revealed differential cell type population alterations in response to exposure to the classic environmental contaminant 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) (Haimbaugh et al., 2022), which is known to induce male fertility defects in both fish and humans (King Heiden et al., 2009; Baker et al., 2013; Baker et al., 2016; Eskenazi et al., 2018; Bruner-Tran et al., 2019). One marked difference was a loss of spermatids and spermatozoa; in these cells, pathways of apoptosis and sperm disorder were upregulated. We show that these intriguing pathway

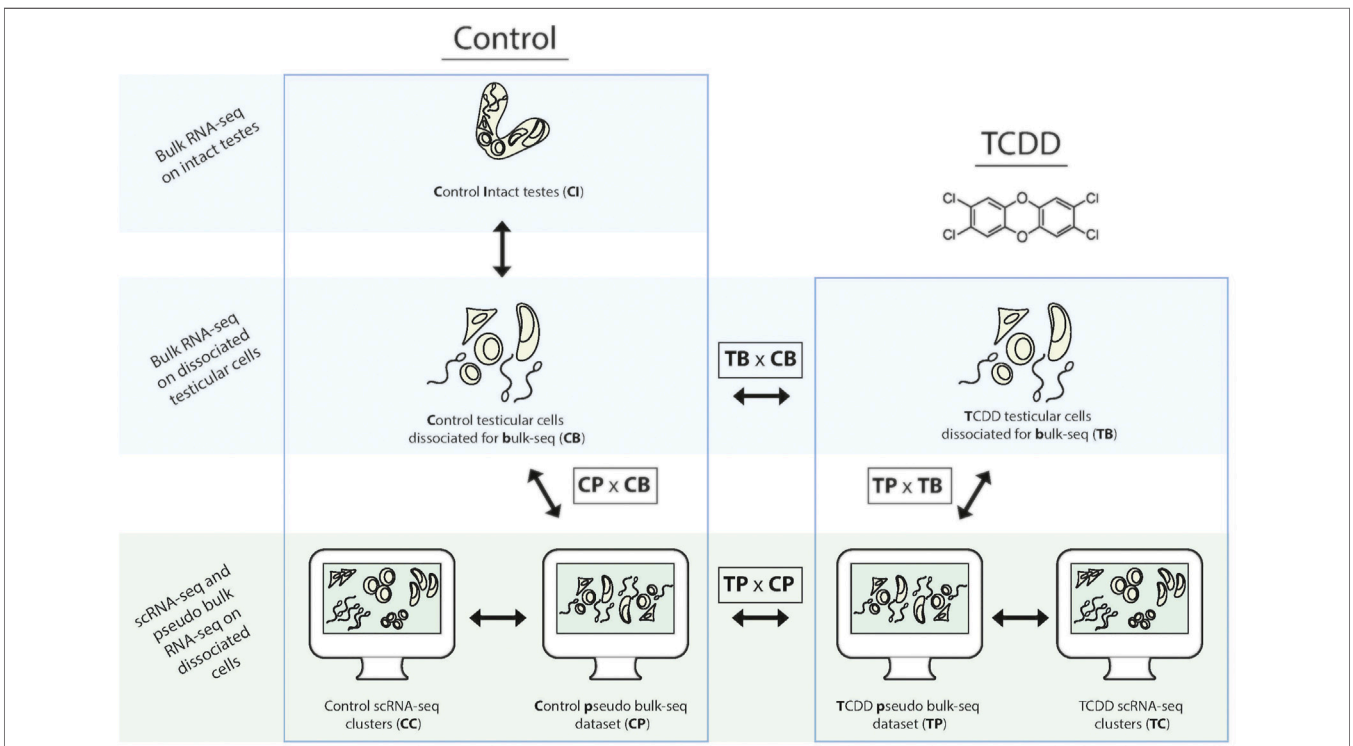


FIGURE 1 | Schematic of the comparative transcriptomic analyses from each method and associated tissue preparation. To determine baseline effects of dissociation, paired samples of intact and dissociated testes were taken for bulk-seq. Bulk-seq of control (CB) and TCDD-exposed dissociated testes (TB) (middle panel) were compared to pseudo bulk-seq of paired dissociated samples that underwent scRNA-seq (CP, TB) (lower panel, middle images). Pseudo bulk-seq samples were compared to their scRNA-seq cluster complement (lower panel, outer images).

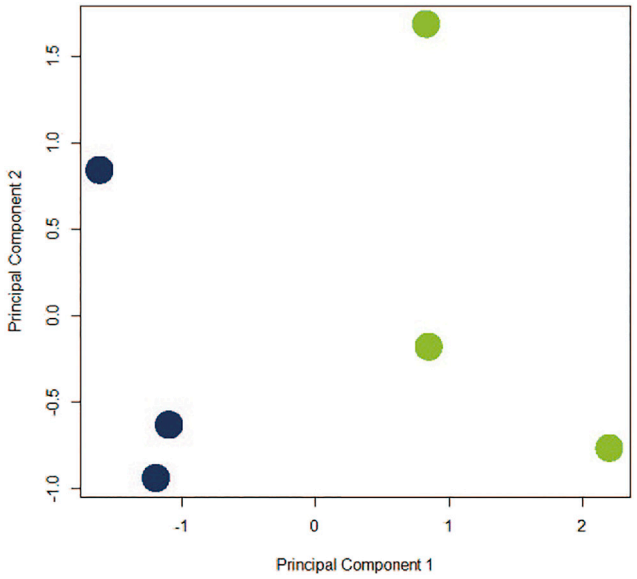


FIGURE 2 | PCA plot depicting co-clustering of gene expression in dissociated replicates (blue) and intact replicates (green). Samples clustered distinctly along PC1 based on method of preparation (negative and positive, respectively).

upregulations resulting from sequencing at the single cell level are not detected with bulk-seq. scRNA-seq detected the change in sperm proportion from about 20% in controls to around 4% in TCDD-exposed cells. In bulk-seq, sperm suffer a similarly reduced physical representation in the sample, but the resulting low signal from these populations may be occluded by competing signals from the overwhelming majority of other cell types. Under bulk-seq conditions, certain aspects of the toxicological profile may be lost to noise, such as toxic effects to relatively few cells. However, we demonstrate that bulk-seq detected immune system modulation that scRNA-seq did not, likely due to cell dropout of immune cells in scRNA-seq. The immune cell population in testes is quite small. We did not detect any immune cells in our scRNA-seq data, and in other scRNA-seq datasets, they comprise just 3% of the overall cell population on average in controls. The required cell dissociation step to perform scRNA-seq can cause cell death, further reducing the number of detected immune cells. scRNA-seq loses about a third of the cell input (10x Genomics Inc., 2021a). Further, by standard bioinformatic processing of scRNA-seq data, cell type populations under 30 cells will not be registered. Minor or unstable cell types such as immune cells may not survive the sample preparation and data processing of scRNA-seq. Early-life exposure to TCDD does not appear to affect adult sperm vulnerability to dissociation, but this potential limitation could apply for other toxicants or other, more acute exposure schemes.

In this study we seek to characterize and compare the TGx profile(s) of early-life TCDD exposure, *via* two complementary methods: pseudo bulk-seq of scRNA-seq data, and bulk-seq of dissociated cells. Abbreviations for the various analyses are provided in **Figure 1** as reference (i.e., pseudo bulk-seq comparison of TCDD-exposed and control samples is abbreviated TPxCP, while the same comparison in bulk-seq is TBxCB). We first confirm the known fact that indeed, dissociation causes transcriptomic disruption (CDxCI); however, this disruption does not appear to significantly affect TCDD-exposed cells more so than controls (CDxCI vs. TBxCB). Single cell sequencing does have a differential effect on TCDD samples compared to the bulk-seq method (TPxTB vs. CPxCB), despite identical tissue preparation of controls. The toxicotranscriptomic testicular profiles of TCDD exposure delivered by bulk-seq (TBxCB) do differ from that of pseudo bulk-seq (TPxCP), presenting significantly better transcriptome coverage, and especially that of low-expressing transcripts and from cell types of smaller populations; at the same time, giving more global information and potentially less toxicologically relevant information. This report provides perspective to inform methodological design in TGx by demonstrating that differential information extracted from sequencing cells—despite being from the same tissue and exposure scheme—is influenced by the specific protocol used, with implications for the interpretation of exposure-induced risk.

RESULTS AND DISCUSSION

The Impact of Protocol Choice (Dissociation) on Gene Expression (CDxCI)

The assumption in measuring any experimental variable where all other factors (i.e., tissue preparation) are held constant is that any necessary technical manipulation will affect both groups equally and thus avoid confounding the results. Tissue dissociation is expected to produce a broad stress response in cells. Mitochondrial activity (van den Brink et al., 2017), heat shock response (O'Flanagan et al., 2019), cell death (Adam et al., 2017; Wu et al., 2018) and their associated gene expression profiles are known effects of dissociation. Tissue dissociation is a well-established "necessary evil" ingrained in common scRNA-seq protocols. Cells must be individuated for sequencing, yet dissociation of the tissue (whether mechanical, chemical, enzymatic, etc.) induces myriad transcriptional artifacts due to the disrupted cell microenvironment. Machado et al. (2021) found 10–50% of the transcriptome is altered by dissociation alone, and Wu et al. (2018) scRNA-seq data returned a cluster consisting primarily of dissociation-induced artifacts. To characterize the unavoidable impact of dissociation on zebrafish testes, we compared bulk-seq data from enzymatically dissociated testes with intact testes. Following dissociation, 435 differentially expressed genes (DEGs) ($p < 0.01$, LFC ≥ 1, ≤ -1) were significantly upregulated; 348 downregulated. **Figure 2** depicts the PCA plot of dissociated and intact samples. Pathway analysis of DEGs confirmed expected differences, such as upregulation of apoptosis, necrosis, and degeneration pathways; downregulation of proliferation and growth pathways (**Tables 1**, **2**).

Dissociation is generally expected to affect cell types equally within an experiment, but exposure to environmental contaminants can compromise cell health, rendering exposed

TABLE 1 | Ingenuity Pathway Analysis table of disease and Biological Functions from CDxCI DEGs ($p < 0.01$, log fold change >1/<−1). From "Diseases or Functions Annotations'' column, after removing any redundancies, functions without a z-score, or a z-score between −1 and 1. Mapped IDs: 9865; unmapped IDs: 4922.459 analysis-ready molecules: 214 upregulated and 245 downregulated. Ranked by z-score. Left: upregulated pathways. Right: downregulated pathways.

Rank	Upregulated diseases or Functions Annotation	*p*-value	Bias-corrected z-score	Rank	Downregulated diseases or functions annotation	*p*-value	Bias-corrected z-score
1	Necrosis of epithelial tissue	1.15E-06	3.14	48	Cell proliferation of T lymphocytes	0.00119	−1.00
3	Cell death of epithelial cells	1.46E-05	2.99	54	Invasion of tissue	3.61E-08	−1.13
7	Necrosis	8.51E-07	2.79	56	Activation of cells	0.00341	−1.19
9	Apoptosis of epithelial cells	5.81E-06	2.76	73	Growth of genital organ	3.76E-05	−1.65
13	Organ Degeneration	0.00353	2.18	74	Cell movement of epithelial cells	6.98E-05	−1.70
14	Congenital malformation of	6.46E-07	2.05	75	Migration of epithelial cells	9.97E-05	−1.72
15	Dysgenesis	0.000439	1.96	76	Malignant genitourinary solid	0.00415	−1.74
16	Apoptosis	5.94E-10	1.93	79	Growth of organism	4.66E-05	−1.84
22	Aplasia or hypoplasia	0.000818	1.83	80	Proliferation of gonadal cells	7.97E-05	−1.90
30	Oxidation of fatty acid	0.00444	1.62	88	Vasculogenesis	4.54E-07	−2.71
34	Degeneration of testis	0.00283	1.38	—	—	—	—
39	Steroidogenesis of hormone	1.75E-05	1.15	—	—	—	—
40	Congenital malformation of	0.000175	1.14	—	—	—	—
41	Hypoplasia of genital organ	0.000104	1.13	—	—	—	—

TABLE 2 | Ingenuity Pathway Analysis table of Canonical Pathways from CDxCI DEGs ($p < 0.01$, log fold change >1/<-1). Redundant pathways, pathways with no z-score or z-score between -1 and one removed. Mapped IDs: 9865; unmapped IDs: 4922.459 analysis-ready molecules: 245 upregulated and 214 downregulated. Left: upregulated pathways. Right: Downregulated pathways. Ranked by z-score.

Rank	Upregulated ingenuity canonical pathways	−Log (*p*-value)	z-score	Rank	Downregulated ingenuity canonical pathways	−Log (*p*-value)	z-score
1	Cyclins and Cell Cycle Regulation	2.62	2.65	4	Agrin Interactions at Neuromuscular Junction	4.6	−2.65
2	Aryl Hydrocarbon Receptor Signaling	3.27	1.13	5	ILK Signaling	2.04	−3.16
3	Estrogen-mediated S-phase Entry	6.47	1.13	—	—	—	—

cells more vulnerable to threats. In our scRNA-seq data, the contributions of TCDD-exposed sperm population to the overall dataset were dramatically reduced (about 4%) compared to controls (about 23%; a 79% reduction in representation). We previously explored the possibility of this shrinkage being due to the technical requirement of cell dissociation (Haimbaugh et al., 2022). Histology of intact tissue confirmed an apoptosis-driven decrease of 11% in healthy sperm cell representation following exposure, certainly not to the extent observed in scRNA-seq. While apoptosis without dissociation is occurring, we assume the sperm cells that dropped out of scRNA-seq data were toxicologically impaired to the point where dissociation induced cell death in a greater swath of the population. ScRNA-seq has been criticized for its high dropout rate, and we interpret our results as indicating toxicant-related stress could increase the dropout rate in TGx studies.

In bulk-seq exposure data (TBxCB) however, dissociation appears to have minimal impact. The changes evident in TBxCB most likely represent TCDD exposure, and not the tissue preparation method. When examining genes in the Ingenuity Pathway Analysis (IPA) Canonical Pathways of both CDxCI and TBxCB, only three genes are shared of the 113 total genes involved. It is worth mentioning expression of those three overlapping genes (*aldh3b1*, *pklr* and *sat1*), or genes in similar pathways, may be affected by dissociation. An experimental model investigating these genes *via* scRNA-seq may be confounded by dissociation effects. Future scRNA-seq exposure studies should include a dissociation reference in order to parse baseline dissociation effects from toxicologically-relevant changes. In this same vein, "Aryl Hydrocarbon Receptor (AhR) Signaling Pathway" and "Estrogen-mediated S-phase Entry" were upregulated simply by dissociation (CDxCI). AhR and estrogen signaling are a canonical signature of TCDD exposure, and have been thoroughly studied, exclusive of dissociation (Safe et al., 1998; Wilson and Safe, 1998). These findings in context underscore the importance of properly controlling for potential method-specific artifacts in toxicological experiments, which may have to be adjusted for the specific toxicant under study.

CPxCB: The Droplet-Based Technical Effect

The benefit of scRNA-seq is that population composition can be captured since the transcriptome of each cell in a tissue is represented. At the same time, an established limitation of scRNA-seq is dropout of certain cells and of mRNA detection. Cell dropout can have a physical or technical basis, and transcript dropout has a bioinformatic (QC) basis. The underrepresentation of smaller or rare cell types in scRNA-seq is hypothesized by Denisenko et al. (2019) as being due to differing resistance to cell lysis on the flow cell (thus preempting sequencing) or simply due to irregular cell size or shape preventing entry to the flow cell. Once on the flow cell, scRNA-seq preferentially detects more

TABLE 3 | Ingenuity Pathway Analysis table of Canonical Pathways from CPxCB DEGs ($p < 0.01$, log fold change >1/<-1). Redundant pathways, pathways with no z-score or z-score between -1 and one removed. Mapped: 2374; unmapped: 594.1,465 analysis-ready molecules: 735 upregulated and 730 downregulated. Ranked by z-score. Left: upregulated pathways. Right: Downregulated pathways. Ranked by z-score.

Rank	Upregulated ingenuity canonical pathways	-Log (*p*-value)	z-score	Rank	Downregulated ingenuity canonical pathways	-Log (p- value)	z-score
1	Oxidative Phosphorylation	3.62	2.24	34	Coronavirus Pathogenesis Pathway	3.62	−1.06
2	Kinetochore metaphase	4.71	2.06	35	BAG2 Signaling Pathway	2.86	−1.26
3	Assembly of RNA	3.52	1.73	36	Aldosterone Signaling in Epithelial	3.75	−1.26
4	Spliceosomal Cycle	2.34	1.67	37	3-phosphoinositide Biosynthesis	2.03	−1.278
5	Cyclins and Cell Cycle	2.16	1.51	38	Androgen Signaling	2.72	−1.34
6	Superpathway of Cholesterol	2.74	1.41	39	Estrogen Receptor Signaling	5.14	−1.54
7	Pyrimidine	3.8	1.41	40	Cysteine Biosynthesis III	2.23	−1.63
8	Cholesterol Biosynthesis I	2.33	1.34	41	RAN Signaling	2.82	−2.45
9	Cholesterol Biosynthesis II	2.33	1.34	42	Unfolded protein response	2.74	−2.71
10	Cholesterol Biosynthesis III	2.33	1.34	42	Unfolded protein response	2.74	−2.71

abundant transcripts, overlooking low-expressing genes (Kharchenko et al., 2014), and has a strong 3′ end bias (Zhang et al., 2019). As for quality control filters, clusters and the cells constituting them are pre-defined by parameters set by the researchers. If these requirements are not met, they are excluded from analysis. The standard definition of a cluster is that it consists of >30 cells. A small cell population's presence may be grouped into a similar cluster, erasing its unique identity and dropping out of the dataset. A cell is defined as having >500 features (genes) detected. Transcriptome coverage is known to be less extensive in scRNA-seq, with a bias towards detecting longer and more highly expressed transcripts (Macosko et al., 2015; Phipson et al., 2017; Davies et al., 2021). A low-expressing cell or one expressing short transcripts such as transcription factors (TFs), then, could be erroneously removed from analysis as an artifact. Any of these factors could result in artifactual gene expression differences between pseudo bulk-seq and bulk-seq. At least a third of cells are expected to escape capture in the 10x pipeline (10x Genomics, Inc., 2021a). Dropout is assumed proportional across cell types, thus a collapsed scRNA-seq dataset (pseudo bulk-seq) is expected to resemble the bulk-seq dataset from its biological and technical counterpart. Comparing pseudo bulk-seq to bulk-seq data can be used to account for the scRNA-seq cell and transcript dropout rate. The only difference between pseudo bulk-seq on dissociated cells, and bulk-seq on dissociated cells from the same sample, is the technical influence from the introduction of the 10x microfluidic system, where cell lysis and cDNA synthesis occur in oil droplets containing individual cells, rather than as a traditional cell suspension of all cell types at once. We call this technical effect the droplet effect, after the multiple scRNA-seq technologies of encapsulating the cells along with reagents and barcodes in an oil droplet. The assumption is that any differences between the resulting datasets of each method would thus be attributed to the droplet effect. We characterized how the transcriptome profiles of each differ in terms of DEGs and low-expressor coverage.

Of note, the scRNA-seq pipeline appears to have a significant impact on differential gene expression. When comparing control pseudo bulk-seq datasets to control bulk-seq datasets from identical samples, we observed 1,102 significantly upregulated DEGs; 1,029 significantly downregulated. These genes were involved in upregulating pathways of oxidative phosphorylation, cholesterol biosynthesis and cell cycle regulation (**Table 3**). The most highly upregulated pathway was of oxidative phosphorylation (OXPHOS). One known drawback of scRNA-seq is the cellular stress the required tissue dissociation induces, which is routinely quantified during QC as % mitochondrial content, including OXPHOS genes. The mitochondria release mtRNA under duress, and thus signals an unhealthy, unrepresentative cell. The recommendation is that <10% mitochondrial transcriptional content is required to move forward with a scRNA-seq dataset. Cholesterol biosynthesis can be increased by cell dissociation (Volpe et al., 1985). Both pseudo bulk-seq and bulk-seq samples were dissociated in the same way at the same time, however, there is a short but unquantifiable amount of time between loading cells onto the flow cell and transcriptome capture. Bulk-seq cells were immediately placed in lysis reagent. During this short period between dissociation of both samples and the act of single cell sequencing, there is additional time for these cells to respond to dissociation, which could explain the pathway upregulation in identically-prepared samples. Cell cycle dysregulation is a known, non-biologically relevant source of variation in scRNA-seq data. It is often regressed out of scRNA-seq datasets (Luecken and Theis, 2019; Hérault et al., 2021).

As expected, transcriptome coverage was decreased in pseudo bulk-seq, detecting only 6,847 RefSeq mRNA Accession IDs to bulk-seq's 14,160. We next tested the idea that scRNA-seq data underrepresents shorter and/or less abundant mRNAs. First, we surveyed control transcriptomes for low expressor coverage by defining a low expressing transcript as possessing between one read count and ≤1% of the maximum normalized read count sum within a dataset, and then calculated the percent transcriptome coverage of these low expressors for each method. By setting a filter for bulk-seq data to match the pseudo bulk-seq read count sum lower limit, we investigated how bulk-seq datasets would be hypothetically diminished by the higher read count threshold. The average coverage by bulk-seq then fell by about 30% ($p = 0.003522$, 1-tailed *t*-test) (**Table 4**). Thus, if the bulk-seq transcript detection sensitivity matched pseudo bulk-seq sensitivity, a sizable population of bulk-seq transcripts would

TABLE 4 | Coverage of low expressors is reduced in bulk-seq when scRNA-seq filters apply. Low expressor reads counts are an average across all control samples (CD, CI, CB).

		Detected transcripts	Detected transcripts (significant DEGs)
No filter	Low expressor reads count range	1–823,127	1–178,033
	Low expressor coverage	51.81%	24.45%
Filter	Low expressor reads count range	106–823,127	106–178,033
	Low expressor coverage	36.62%	7.20%
—	*p*-value	0.003522	0.045294

TABLE 5 | Detection of transcription factors is reduced in pseudo bulk-seq. TBxCB: TCDD bulk-seq vs control bulk-seq. TPxCP: TCDD pseudo bulk-seq vs control pseudo bulk-seq. Cells pertaining to significant DEGs are shaded. Chi-squared test.

	Detected transcripts	Detected transcription factors	Detected transcripts (significant DEGs)	Detected transcription factors (significant DEGs)
TBxCB	14,160	1,393 (9.8%)	310	63 (20.3%)
TPxCP	6,847	487 (7.1%)	1,099	44 (4%)
p-value	NA	6.84e-10	NA	3.14e-21

remain undetected. This difference is more important when dealing with only significant DEGs. While the pseudo bulk-seq dataset does not contain low expressors (as defined as 1% of the maximum count sum), about 25% of low expressors are significant DEGs in the bulk-seq dataset. When the filter was applied to match the pseudo bulk-seq sensitivity, this dropped to around 7% ($p = 0.045294$, 1-tailed *t*-test). The presumed absence of these genes in single cell datasets could affect interpretation of the results.

Second, using transcription factors (TFs) as a proxy for both short and low-expressing transcripts, we searched the pseudo bulk-seq and bulk-seq dataset for the available list of 3,068 *Danio rerio* TFs (ATFDBv3.0) (AnimalTFDB3, 2021). Fewer TFs were detected in the pseudo bulk-seq dataset compared to bulk-seq, and this slight trend is exaggerated for significant DEGs (**Table 5**). Further, none of these significant TFs are shared between the two methods. In fact, the entire population of TFs detected by pseudo bulk-seq is a significantly different population than the significant bulk-seq DEGs ($p = 2.175e-05$, chi-squared test), with only seven genes overlapping. Thus, a study using solely one method would receive a one-dimensional representation of the data, and may not cover TFs of interest. "Missing" genes in either dataset, by virtue of their absence, could influence the toxicological interpretation of experimental findings.

While scRNA-seq pipleine itself is not assumed to affect unexposed, healthy cells any differently than cells exposed to environmental contaminants, it is important to determine if the established baseline changes in DEGs, gene ontology, and low expressor coverage resulting from the droplet effect in controls (CPxCB) would remain in TCDD-exposed samples (TPxTB). If TCDD exposure exerted no influence on scRNA-seq processing, a high overlap would be expected among each aspect of CPxCB and TPxTB. The number and fold change direction of significant DEGs held steady in TPxTB

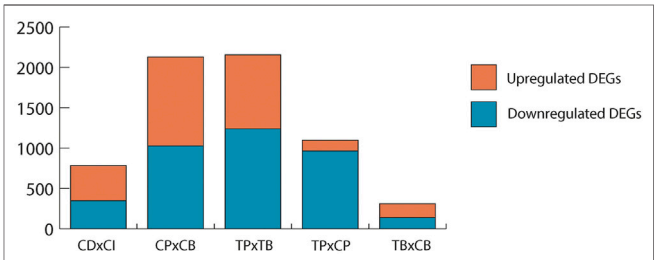


FIGURE 3 | Significant DEGs ($p < 0.01$, LFC ≥1 or ≤ −1) in each comparison. Upregulated: LFC ≥1; downregulated: LFC ≤ −1. CDxCI: paired control samples of dissociated and intact testes used for bulk-seq. CPxCB: paired control samples of dissociated testes used for pseudo bulk-seq and bulk-seq. TPxTB: paired TCDD-exposed samples of dissociated testes used for pseudo bulk-seq and bulk-seq. TPxCP: pseudo bulk-seq analysis of paired dissociated control and TCDD-exposed testes. TBxCB: bulk-seq analysis of paired dissociated control and TCDD-exposed testes.

comparisons (1,239 up-; 917 downregulated) as compared to CPxCB (**Figure 3**), with considerable overlap (about 60%).

The core pathways associated with these DEGs (**Table 6**) were also similar to pathways of CPxCB. About 40% of CPxCB and TPxTB pathways overlap, including oxidative phosphorylation, cholesterol biosynthesis, and cell cycle regulation upregulation. These overlapping main pathways demonstrate the general influence of the droplet effect on gene expression from otherwise identically prepared samples and are not toxicologically relevant to TCDD exposure. In fact, cholesterol biosynthesis has been thoroughly shown to decrease following TCDD exposure (Fletcher et al., 2005; Sato et al., 2008; Tanos et al., 2012; Fader et al., 2017). However, it is important to note that 60% of pathways were unique to TCDD; other chemicals may have more or less of an effect. Bulk-seq detection of low-expressing transcripts and

TABLE 6 | Ingenuity Pathway Analysis table of Canonical Pathways from TPxTB DEGs ($p < 0.01$, log fold change >1/<-1). Redundant pathways, pathways with no z-score or z-score between -1 and one removed. Mapped IDs: 3550; unmapped IDs: 838.184 analysis-ready molecules: 103 upregulated and 81 downregulated. Ranked by z-score. Left: upregulated pathways. Right: Downregulated pathways.

Rank	Upregulated ingenuity canonical pathways	z-score	Rank	Downregulated ingenuity canonical pathways	z-score
1	Oxidative Phosphorylation	2.52	12	p70S6K Signaling	−1.07
2	Cholesterol Biosynthesis I	2	13	mTOR Signaling	−1.07
3	Cholesterol Biosynthesis II (*via* 24,25-dihydrolanosterol)	2	14	Insulin Receptor Signaling	−1.21
4	Cholesterol Biosynthesis III (*via* Desmosterol)	2	15	Remodeling of Epithelial Adherens Junctions	−1.34
5	Superpathway of Cholesterol Biosynthesis	1.89	16	Gluconeogenesis I	−1.41
6	Cell Cycle Control of Chromosomal Replication	1.60	17	D-myo-inositol-5-phosphate Metabolism	−1.41
7	Assembly of RNA polymerase II Complex	1.07	18	HIF1α Signaling	−1.53
—	—	—	19	Glycolysis I	−1.63
—	—	—	21	Sirtuin Signaling Pathway	−1.71
—	—	—	22	3-phosphoinositide Biosynthesis	−1.79
—	—	—	23	Superpathway of Inositol Phosphate Compounds	−1.96
—	—	—	24	Androgen Signaling	−2
—	—	—	25	Coronavirus Pathogenesis Pathway	−2.04
—	—	—	26	Estrogen Receptor Signaling	−2.14
—	—	—	27	Autophagy	−2.35
—	—	—	28	IGF-1 Signaling	−2.50

TFs is not affected by TCDD exposure, before or after filters were applied to match pseudo bulk-seq coverage. Coverage reduction is simply a result of the pseudo bulk-seq method and is not mediated by exposure. By these three measures, TCDD exposed testes cells do not appear differentially susceptible to the technical effects of the scRNA-seq pipeline.

Clustering Reveals Differential Susceptibility to scRNA-Seq Dropout (TPxCP)

Despite no differences in technical dropout between exposed cells and control cells, clustering of the TCDD sample by scRNA-seq revealed near-total physical dropout of spermatids and sperm, beyond what is attributed to exposure. We have shown in intact tissue that a significant percent of sperm and spermatid populations apoptose in response to TCDD exposure (Haimbaugh et al., 2022), but the remaining healthy sperm and spermatids contribute similarly to the overall population as control sperm and spermatids. In scRNA-seq clusters, these two populations are diminished by about 80%. Other cell types are not affected as drastically, in fact, the proportion of spermatogonial stem cells did not change. Dropout in scRNA-seq is assumed to be distributed proportionally across all cell types in a sample. However, in testes cells with a history of TCDD exposure, dropout is differentially experienced by late germ cells. These late germ cells are known to be under strain, as many are undergoing apoptosis. The stress background of a tissue may influence cell dropout in the scRNA-seq pipeline as the added mitochondrial duress and increased latency to lysis on the flow cell could be an insurmountable affront to already-unstable cells from both exposure and the combination of exposure and dissociation mentioned above.

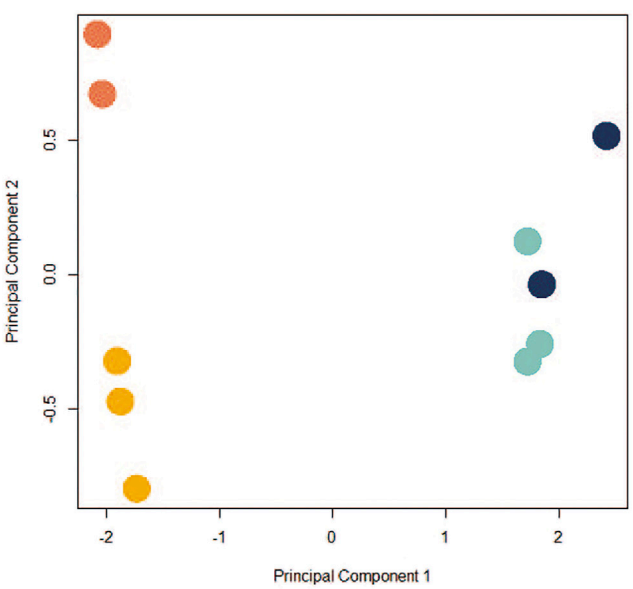


FIGURE 4 | PCA plot depicting co-clustering in CBxTB and CPxTP. Control bulk-seq (CB) replicates: light orange; TCDD-exposed bulk-seq (TB) replicates: dark orange; control pseudo bulk-seq (CP) replicates: light blue; TCDD-exposed pseudo bulk-seq (TP) replicates: dark blue. Samples clustered distinctly along PC1 based on method of preparation (bulk-seq (CB, TB) or pseudo bulk-seq (CP, TP); negative and positive, respectively), independently of exposure status. Control and exposed replicates from either preparation clustered along PC2 (exposed replicates tended to cluster above the control replicates).

Toxicotranscriptomic Profile Representation in Bulk-Seq vs. Psuedo Bulk-Seq (TBxCB vs. TPxCP)

Given the ways dissociation and individual sequencing can affect control and exposed cells (CPxCB, TPxTB), and the growing

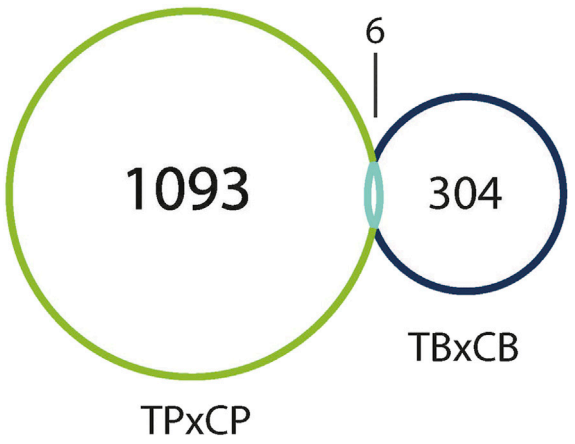


FIGURE 5 | Venn diagram of DEGs (both up- and downregulated) in TCDD-exposed pseudo bulk-seq preparation (left), and bulk-seq preparation (right) as compared to control. Pseudo bulk-seq detected 1,093 DEGs that were undetected in bulk-seq. Conversely, 304 genes were detected using bulk-seq, but not pseudo bulk-seq. 6 DEGs were detected by both methods.

interest in using scRNA-seq for TGx studies, we next examined the toxicologically-relevant differences in profiles between bulk-seq (TBxCB) and pseudo bulk-seq (TPxCP) exposure datasets. Since pseudo bulk-seq is a collapsed transcriptome of every individual cell, it is expected to resemble the bulk-seq averaged transcriptome. Any differences in the two datasets can help estimate the cell populations undergoing dropout in scRNA-seq. If the pattern of cell dropout does not meet the expectation that all cell types will be affected equally, this suggests the sequencing process itself contributed to the sacrifice of a cell population. In the bulk-seq comparison, 310 genes were significantly differentially expressed (171 upregulated and 139 down-; $p < 0.01$; LFC ≥ 1 or ≤ -1) between TCDD-exposed and control; the pseudo bulk-seq comparison contained 1,099 significant DEGs (134 up- and 965 down-). The greater number of DEGs in scRNA-seq is likely a product of noise due to both cell and transcript dropout. Disproportionate cell-type dropout in scRNA-seq TCDD samples produces a composition of cells unlike that remaining in the bulk-seq TCDD samples, where dropout does not apply. The increased incomparability of cell type populations embeds noise in the system. Transcript dropout from lower transcriptome coverage in scRNA-seq results in a decreased signal:noise ratio. **Figure 4** shows the differences between the control and TCDD samples for each method.

Interestingly, only six of these DEGs overlapped between the two preparations (**Figure 5**). A common use of scRNA-seq is identification of cell-type markers of interest for further experimentation. Populations expressing a particular marker are then isolated from new tissue, and bulk-seq is performed to allow for deeper sequencing coverage. The markedly different DEG profiles between pseudo bulk-seq and bulk-seq could mislead marker identification and subsequent experiments. Despite low overlap in DEGs, it is possible the transcriptomic profiles of each method converged on toxicologically-relevant pathways, therefore, we compared the overarching functions represented by each method (**Tables 7**, **8**). The 43 pathways generated from pseudo bulk-seq data seem to convey more specific toxicological functions (teratozoospermia, impaired cilia formation), and it is clear the reproductive system has been affected. With the 55 bulk-seq pathways, it is clear basic cellular functions are under distress (ion homeostasis, apoptosis, ROS production), but without prior knowledge, it would be difficult to assume the tissue in question, as they range over less informative pathways. Additionally, the pseudo bulk-seq results of sperm disorder, oligozoospermia, etc., reflect the phenotypic infertility and the lowered male-mediated fertilization rates we have observed following TCDD exposure. There were three exact overlapping pathway annotations from pseudo bulk-seq and bulk-seq including cancer of secretory structure, and two non-specific cancer pathways. The prominence of sperm-related pathways in pseudo bulk-seq data and absence of such in the bulk-seq pathway list may be explained by the observed dropout of sperm cells in scRNA-seq. The sperm-specific signal may have been overridden by the abundance of other cell-type signals in bulk-seq, such as from spermatocytes (45% of all cells, Haimbaugh et al., 2022). In fact, meiosis-related pathways are well-represented in the pseudo bulk-seq results but absent from bulk-seq. This lack of focus could also be due to the relatively low number of IPA-ready molecules for bulk-seq (206) compared to pseudo bulk-seq (696). The full list of IPA results is **Supplemental File S1**.

Bulk-seq pathways included a potential aspect of the exposure that the pseudo bulk-seq dataset did not: immune system damage or suppression. The most downregulated pathways in bulk-seq are immune-related, while in scRNA-seq, there is no indication of changes in immune function. TCDD is a known immunotoxicant (Warren, 2000; Marshall and Kerkvliet, 2010). Immune cells, while serving important functions in the testes, are such a small population they are often not detected in scRNA-seq experiments, or at very low representation averaging about 3% (Green et al., 2018; Guo et al., 2018; Wang et al., 2018; Jung et al., 2019; Sohni et al., 2019; Shami et al., 2020; Yang et al., 2020). We did not detect immune clusters in our scRNA-seq dataset. It follows, then, that without enriching a sample for immune cells prior to scRNA-seq, immune system information may be lost. The revelation of this loss between our bulk and single cell data reiterates the importance of hypothesis-driven research. ScRNA-seq in general has been rightly touted as a useful tool in discovery-based research which can be mined for relevant information. However, with the testes it may be the case that when studying environmental contaminants known or suspected to be immunotoxic, the *a priori* decision to enrich for immune cells in a scRNA-seq study would produce a more realistic and meaningful dataset.

scRNA-Seq Clusters Deliver Crucial Toxicological Details (TCxCC)

As we have seen with the comparison of bulk-seq (TBxCB) to pseudo bulk-seq (TPxCP), the narrowing of focus from a

TABLE 7 | Ingenuity Pathway Analysis table of disease and Biological Functions from TPxCP DEGs ($p < 0.01$, log fold change >1/<−1). From "Diseases or Functions Annotation'' column, after removing any redundancies, functions without a z-score, or a z-score between −1 and 1. Mapped IDs: 4,926; unmapped IDs: 1,921 (raw data), after settings of logFC >1 or < −1/$p < 0.01$.696 analysis-ready molecules: 89 upregulated and 607 downregulated. Ranked by z-score. Left: upregulated pathways. Right: downregulated pathways.

Rank	Upregulated diseases or functions annotation	*p*-value	Activation z-score	Rank	Downregulated diseases or functions annotation	*p*-value	Activation z-score
1	Sperm disorder	2.63E-05	3.45	67	Ploidy of cells	1.07E-05	−1.08
2	Carcinoma	2.39E-31	3.30	78	Recombination	1.78E-06	−1.15
7	Nonhematologic malignant neoplasm	1.74E-34	2.79	79	DNA recombination	1.78E-06	−1.15
8	Extracranial solid tumor	3.88E-31	2.74	80	Recombination of cells	6.22E-05	−1.22
9	Tumorigenesis of epithelial neoplasm	0.000777	2.71	81	Cell movement of sperm	2.07E-08	−1.34
10	Non-melanoma solid tumor	9.45E-31	2.68	83	Cycling of centrosome	0.000515	−1.45
11	Teratozoospermia	2.81E-05	2.65	84	Formation of cilia	1.12E-26	−1.47
14	Genitourinary tumor	0.000236	2.51	85	Homologous recombination	6.43E-05	−1.53
15	Genitourinary carcinoma	6.42E-05	2.43	87	Meiosis of germ cells	2.81E-05	−1.72
18	Laterality defect	6.43E-08	2.42	88	Smoothened signaling pathway	3.83E-12	−1.73
19	Malignant genitourinary solid tumor	0.00013	2.38	89	Cell surface receptor linked signal transduction	2.72E-05	−1.73
20	Heterotaxy or ciliopathy	1.31E-33	2.24	90	Formation of cellular protrusions	1.88E-06	−1.76
22	Regional congenital anomaly	4.73E-07	2.22	91	Organization of cytoplasm	1.09E-07	−2.19
24	Oligozoospermia	0.000117	2.14	92	Organization of cytoskeleton	4.15E-06	−2.19
25	Solid tumor	4.03E-31	2.09	93	Microtubule dynamics	5.28E-06	−2.18

TABLE 8 | Ingenuity Pathway Analysis table of disease or Functions from TBxCB DEGs ($p < 0.01$, log fold change >1/<−1). From "Diseases or Functions Annotations" column, after removing any redundancies, functions without a z-score, or a z-score between −1 and 1. Mapped IDs: 9,868/unmapped IDs: 4,924 (raw data), after settings of log fold change >1 or < −1, $p < 0.01$.206 analysis-ready molecules: 114 upregulated and 92 downregulated. Ranked by z-score. Left: upregulated pathways. Right: downregulated pathways.

Rank	Upregulated diseases or functions annotation	*p*-value	Activation z-score	Rank	Downregulated diseases or functions annotation	*p*-value	Activation z-score
1	Neurotransmission	2.66E-07	2.53	41	Ductal carcinoma	0.00093	−1.67
2	Transport of cation	0.00014	2.20	42	Extraintestinal functional disorder	8.1E-06	−1.74
3	Transport of ion	1.57E-05	2.07	43	Morbidity or mortality	1.52E-05	−1.79
4	Transport of inorganic cation	0.000174	1.96	44	Migration of endothelial cells	0.000904	−1.90
5	Transport of metal ion	0.000314	1.96	45	Cell movement of antigen presenting cells	0.000257	−2.15
6	Transport of molecule	1.77E-06	1.77	46	Cell movement of macrophages	0.000878	−2.56
7	Ion homeostasis of cells	3.39E-06	1.63	47	Cell movement of phagocytes	0.000715	−2.58
8	Growth of organism	0.000509	1.61	48	Cellular infiltration by phagocytes	0.000381	−2.63
9	Colon tumor	0.000743	1.43	49	Cellular infiltration by leukocytes	0.00106	−2.73
10	Transport of metal	0.000106	1.43	50	Cellular infiltration by myeloid cells	0.000528	−2.76
11	Depolarization of cells	9.93E-05	1.39	51	Cell movement of leukocytes	0.000371	−2.76
12	Synthesis of reactive oxygen species	0.000636	1.34	52	Leukocyte migration	0.000582	−2.76
13	Incidence of tumor	7.45E-10	1.32	53	Cellular infiltration by blood cells	0.000574	−2.85
14	Apoptosis	5.7E-10	1.27	54	Cellular infiltration	0.000735	−3.01
15	Genital tumor	7.13E-07	1.27	55	Cell movement of myeloid cells	0.00102	−3.15

broad-ranging question of global expression, to a more specific inquiry that takes into account heterogeneity of a tissue, can impact the biological or toxicological understanding of the results. The same appreciation of granularity applies when comparing a pseudo bulk-seq dataset (TPxCP) to its scRNA-seq cluster counterparts (TCxCC). Clusters are distinct entities providing unique information about cell types in a tissue, and are the main deliverable of scRNA-seq. Two indispensable advantages of clustering are receiving the both proportions of cell type (or cellular state) constituting a tissue, and the gene expression signature of each cluster. Pseudo bulk-seq is not meant to provide that information. Pseudo bulk-seq data collapses scRNA-seq datasets to obtain a broad overview of those extricated clusters. Collapsing the transcriptomic dimensionality from individual clusters into one generalized dataset is also useful in checking for dropout as described above, but will naturally fail to retain all cell-specific changes. This loss of specificity is greater when comparing clusters to bulk-seq data, despite bulk-seq containing the same cell types. This bulk-to-cluster comparison is still important, however, to estimate dropout, as discussed above. Here we compare representations of the transcriptome from scRNA-seq clusters to pseudo bulk-seq and to bulk-seq data following TCDD exposure.

In previous work, we captured ten scRNA-seq clusters spanning all testicular cell types in control fish testes (spermatogonial stem cells (SSC), spermatogonia (SPG), four stages of spermatocytes (SPC), round and elongating spermatids, and two sperm clusters) (Haimbaugh et al., 2022). In TCDD-exposed fish testes, only eight clusters remained--sperm and late spermatid populations were decimated. A total of 980 genes were significantly differentially expressed among each of the eight control and TCDD clusters. 574 of these DEGs were not included in pseudo bulk-seq DEGs; only 243 DEGs overlapped. These differing representations result from the collapse of transcriptomic information across all cells in pseudo bulk-seq. Spermatogenesis is a multi-step process, where transcriptional programs that are turned on in earlier stages must be silenced in later stages as the cell's needs fluctuate. Each subsequent developmental stage in the male germ cell trajectory is quite different from the one preceding it. In SSCs, self-renewal and mitosis occur. Some of those progeny will remain SSCs, and some will differentiate to SPG. SPG proliferate and then meiose into SPCs, which undergo a second meiosis to form round spermatids. These round spermatids undergo dramatic architectural elongation and compaction to produce the mature spermatozoa. This can lead to the effect of some transcript expression levels "cancelling out," as one subpopulation expresses, for example, a mitosis program where DNA must be accessible for synthesis, while another compacts chromatin quite tightly to shape sperm.

This divergent transcriptional representation in clusters is exaggerated with bulk-seq. Bulk-seq, being an average expression profile of all cells in the suspension, has the same issues as pseudo bulk-seq would in terms of transcriptional programs "cancelling out," but with the added noise of the dropout manifesting in the clusters. The uneven dropout in scRNA-seq results in a different sample composition than that of bulk-seq, as described above. Here, 800 DEGs were unique to clusters, with only 17 DEGs overlapping with bulk-seq DEGs.

Without the cluster information, it would've been difficult to predict the widespread apoptosis from bulk-seq or pseudo bulk-seq alone, as pathway analysis returned apoptosis pathways with either weak (<1) z-scores, or only one apoptosis result of 55 total pathways (1.8%), respectively. From the clusters we could determine the cell population proportions changed from sperm and spermatids contributing about 30% of the control population, to about 4% after exposure. As a result of the population shift, SPG and early SPC made up about 80% of the TCDD population, whereas in controls they constituted about 30%. From this population shift we were able to hypothesize either a failure in spermiogenesis, or cell death of sperm. Cell death of fully developed sperm was confirmed by immunohistochemistry (Haimbaugh et al., 2022).

SUMMARY

Acceleration in the field of transcriptomics has brought about myriad useful, high-powered technologies for investigating every aspect of gene expression. Deliberately choosing the method to best understand the specific research question becomes complicated, yet remains critical. This is pertinent for any discipline, including the growing field of TGx. The implications for regulatory toxicology cast a layer of added urgency to this task. We demonstrate here the nuances associated with two common transcriptomic assays (bulk-seq and scRNA-seq), using an early-life TCDD exposure model. Due to cell and transcript dropout combined with reduced transcriptome coverage in scRNA-seq, these two assays offer incomparable DEG profiles. As scRNA-seq is often used for cell-type marker discovery for future deeper sequencing with bulk-seq, the opposing DEG profiles could complicate marker enrichment and subsequent interpretation of the results. Despite originating from the same tissue and same exposure, the pathways these DEGs contribute to were also divergent, with scRNA-seq pathways offering insight into biological mechanisms of sperm loss following TCDD, while bulk-seq presented a profile of general immune dysregulation. Both pieces of evidence, however, are true. These different perspectives reinforce the need for validation efforts using other methods including phenotypes, histology, and behavior analysis, to supplement transcriptomic findings from different analytical tools.

There are efforts to make bulk-seq and scRNA-seq more cohesive. Spatial transcriptomics, while suffering from low throughput, addresses the question of dropout: no dissociation, microfluidic droplet or complex data analysis are required (10x Genomics, Inc., 2021b). Deconvolution techniques are actively being developed, where scRNA-seq data is used to estimate proportions of cell type populations in homogeneous bulk-seq samples (Hunt et al., 2018; Newman et al., 2019; Wang et al., 2019). A potential limitation to this approach is that all cell populations may not be present in the scRNA-seq matrix; we have shown this in testes. The ability to extract information from bulk-seq with scRNA-seq-level specificity would preempt scRNA-seq-specific artifacts and reduce experimental costs dramatically.

Exposure-induced risk is notoriously difficult to determine. The "exposome," or the multiple exposures each individual ever encounters over their lifetime (Wild, 2005), to-date represents an epistemic limit of toxicology, and a profound caveat in epidemiological studies. Total environmental control can be accomplished only in model organisms. Even with this level of control, the toxicotranscriptomic outcomes of exposure can be differentially represented according to the investigational protocol. With this evident variability, it is critical for researchers to produce reliable, replicable data for toxicological risk interpretation in humans. Zebrafish in particular are uniquely positioned to investigate questions of both aquatic and human toxicology, due to their high potential for translatability. We show the transcriptome of adult zebrafish testes is susceptible to early-life TCDD exposure, which can differentially present as sperm death or immunotoxicity, depending on the assay and its associated strengths and artifacts. Future toxicotranscriptomic studies in other tissues, species, toxicants, or exposure schemes may also find differential results between bulk-seq and scRNA-seq. These differential results are not necessarily misleading, and can in

fact enhance the TGx field's understanding of the cellular and transcriptional states complexly affected by exposure.

METHODS

Fish Husbandry

Zebrafish (AB strain) were maintained as described in Meyer et al. (2018). Briefly, fish were fed twice daily and kept on a 14:10 h light/dark cycle (Westerfield, 2000) in buffered, recirculating, reverse osmosis water with temperatures maintained at 27–30 ˚C. Animal use protocols were approved by the Institutional Animal Care and Use Committees at Wayne State University and the University of Wisconsin-Madison, according to the National Institutes of Health Guide to the Care and Use of Laboratory Animals (Protocol No. M00489).

TCDD Exposure

TCDD (>99% purity) (Chemsyn, Concord, ON, Canada) was used as a 0.4 ng/ml stock solution in dimethyl sulfoxide (DMSO) Zebrafish were exposed as previously described (Henry et al., 1997; Baker et al., 2013). Briefly, fish were exposed at three wpf and again at seven wpf to water-borne TCDD (50 pg/ml) or vehicle (0.1% DMSO) for 1 h each time in small glass beakers with gentle rocking. Fish were raised in beakers with daily water changes of 40–60% at a density of five fish per 400 ml beaker between 3 and 6 weeks, and five fish per 800 ml beaker between 6 and 9 weeks post-fertilization. All results are derived from three independent TCDD exposure experiments done in successive blocks.

Testes Isolation: Intact Control Testes (CI)

Adult (1-year-old (+/- 1 month)) male zebrafish were euthanized in tricaine methanesulfonate (1.67 mg/ml) (Fisher Scientific, Waltham, MA, United States) for 10 minutes. Testes were dissected and excess adipose tissue removed in ice cold 1x PBS (Gibco, Waltham, MA, United States). Testes were placed in 300 µL RNALater (Thermo Fisher, Waltham, MA, United States) for 48 h. RNALater was then drained and tissue stored at -80 ˚C.

Testes Isolation and Enzymatic Dissociation of Testes: Control Dissociated (CD); TCDD and Control Dissociated for 10x Sequencing (TP, CP); TCDD and Control Dissociated for Bulk Sequencing (TB, CB)

Adult (1.5-year-old (+/- 1 month) (TP, CP, TB, CB) or 1-year-old (+/- 1 month) (CD)) male zebrafish were euthanized in tricaine methanesulfonate (1.67 mg/ml) (Fisher Scientific, Waltham, MA) for 10 minutes. For CD replicates (n = 3), only one testis was dissected for dissociation, while the contralateral testis remained intact for paired bulk-seq. For all other replicates, both testes were dissected (CB/CP: n = 3; TB/TP: n = 2). Testes were dissected and excess adipose tissue removed in ice cold 1x PBS (Gibco, Waltham, MA, United States). Testes were minced, then centrifuged for 5 min at 500 g. PBS was removed, and 100 uL of digestion media (100 uL Leibovitz's L-15 medium (MilliporeSigma, Burlington, MA, United States), one uL bovine serum albumin (New England BioLabs, Ipswich, MA, United States), one uL DNAseI (Zymo Research, Irvine, CA, United States), and 1 mg collagenase Type II (Worthington Biochemical Corporation, Lakewood, NJ, United States)) was added. Tissue was shaken at 280 rpm for 1.5 h, with manual disruption *via* wide-bore pipetting every 15 min. Cells were centrifuged for 5 min at 500 g, digestion media aspirated, and cells resuspended in PBS. Dead cells were removed with a Dead Cell Removal Kit (Miltenyi Biotec, Bergisch Gladbach, Germany). Cell viability of 90% was determined using the BioVision Live/Dead Cell Viability Assay Kit (BioVision Inc., Milpitas, CA, United States), according to manufacturer's instructions. Cells were immediately 1) loaded for 10x sequencing (CP (n = 3), TP (n = 2)), or 2) placed in Qiazol (Qiagen, Hilden, Germany) and frozen at −80˚C (CB (n = 3), TB (n = 2)). Approximately 5,000 cells were loaded for scRNA-seq per replicate (CP, TP). The remaining cells (approximately two million) from the testes suspension were reserved for bulk seq (CB, TB). All cells (approximately one million) from each CD testis (n = 3) were used for bulk-seq,

RNA Isolation

RNA was isolated from testes using the RNeasy Lipid Tissue Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's specifications. Briefly, samples were homogenized in Qiazol (Qiagen, Hilden, Germany), RNA was separated from organic material with chloroform-isoamyl alcohol mixture (≥99.5%) (Millipore Sigma, Burlington, MA, United States), RNA was purified on a filter and eluted with RNAse-free water. RNA concentration was measured with Qubit 3.0 Fluorometer (Invitrogen, Carlsbad, CA, United States). Isolated RNA was stored at -80 ˚C.

3′-End Library Preparation, Sequencing, and Alignment

3′ mRNA-seq libraries were prepared from isolated RNA using QuantSeq 3' mRNA-Seq Library Prep Kit FWD for Illumina (Lexogen, Vienna, Austria). Samples were normalized to 40 ng/µL (total input of 200 ng in 5 µL) and amplified at 17 cycles. Libraries were quantified using a Qubit 3.0 Fluorometer and Qubit® dsDNA Broad Range Assay Kit (Invitrogen, Carlsbad, CA, United States), and run on an Agilent TapeStation 2200 (Agilent Technologies, Santa Clara, CA, United States) for quality control. The samples were sequenced on a HiSeq 2500 (Illumina, San Diego, CA, United States) in rapid mode (single-end 50 bp reads). Reads were aligned to *D. rerio* (genome assembly GRCz11 (danRer11)) using the BlueBee Genomics Platform (BlueBee, Rijswijk, Netherlands). Raw data and processed files were uploaded to the NCBI GEO database (GSE193758).

10x Library Preparation and Sequencing

Single cell transcriptome profiles were generated using the 10x Chromium Controller v2 chemistry following the Chromium Single Cell 3′ protocol. We acquired 180 million reads per sample,

or ~120,000 reads/cell. Raw data and processed files were uploaded to the NCBI GEO database (awaiting approval).

10x Data Processing

Cell Ranger was used to align sequencing reads to the zebrafish reference genome (dR10) which was constructed using the mkref command (Zheng et al., 2017). Count data was imported to Seurat (version 4.0.4) for quality control (QC) filtering, clustering, dimensionality reduction, visualization, and differential gene expression (Satija et al., 2015; Hao et al., 2021). Each sample was filtered to cells containing at least 500 features with clusters requiring a minimum of 30 cells. Samples were merged prior to normalization and clustering (resolution 0.3). Differentially expressed genes between conditions for each cluster were identified using the "FindMarkers" function.

Ingenuity Pathway Analysis

The functional pathways in each comparison were generated through the use of IPA (Qiagen Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis). Genes with significant differential expression, as defined by a log fold change of ≥1 or ≤ −1, and a p-value<0.01, were uploaded into IPA software, using RefSeq IDs as identifiers.

Apoptosis Assay

Immunohistochemistry was performed by the Wayne State University Biobank and Correlative Sciences Core. Formalin-fixed paraffin-embedded sections of bisected zebrafish were de-waxed and rehydrated in a xylene-ethanol-water series. Endogenous peroxides were removed by a methanol/1.2% hydrogen peroxide incubation at room temperature for 25 min. HIER antigen retrieval was done with a pH6 citrate buffer and the BIOCARE Decloaking Chamber (Concord, CA, United States). A 40 min blocking step with SuperBlock Blocking Buffer (Thermo Scientific, Waltham, MA, United States) was performed prior to adding the primary antibody. Detection was obtained using GBI Labs (Bothell, WA, United States) DAB Chromogen Kit and counterstained with Mayer's hematoxylin. Sections were then dehydrated through a series of ethanol to xylene washes and coverslipped with Permount (Fisher Scientific, Waltham, MA, United States). A 1:100 dilution of Cleaved caspase three antibody (9664S) antibody (Cell Signaling, Danvers, MA, United States) was used overnight at 4 °C.

The authors analyzed CC3 labeling to determine presence and/or extent of apoptosis (control fish: $N = 3$; TCDD-exposed: $N = 3$). We obtained up to three distinct images from replicate testes slides at ×40 magnification, and manually quantified a quadrant of each image (control = 8 quadrants, TCDD = 9 quadrants). Significance of the percent apoptotic cells per cell type between controls and TCDD images was measured *via* student's two-tailed t-test.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI GEO (GSE193758).

ETHICS STATEMENT

The animal study was reviewed and approved by Wayne State IACUC.

AUTHOR CONTRIBUTIONS

TB and KG are senior authors and conceived the project; AH and KG conceptualized the content; AH, DM, CA, and KG were involved in the investigation; AH and KG performed the data analysis and wrote the manuscript; KG and TB edited the manuscript and assisted with data analysis; Funding and resources were provided by TB.

FUNDING

This work was supported by the National Institute of Environmental Health Sciences (R01 ES030722 to TB; F31 ES030278 to DM), and the WSU Center for Urban Responses to Environmental Stressors (P30 ES020957 to DM and TB). Additional funding was provided by the National Science Foundation (Grant No. 1735038 to CA), and by the NIH Center grant (P30 CA022453 to KG) to the Karmanos Cancer Institute at Wayne State University and the Perinatology Research Branch of the National Institutes of Child Health and Development at Wayne State University.

ACKNOWLEDGMENTS

We would like to thank the Wayne State University Applied Genomics Technology Center for providing 10x Genomics services and the use of Ingenuity Pathway Analysis software; and the WSU Biobank and Correlative Sciences Core for performing the cleaved caspase-3 staining.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/ftox.2022.821116/full#supplementary-material

REFERENCES

10x Genomics, Inc. (2021a). *Getting Started: Single Cell 3' Gene Expression (Rev A) [pdf] (p. 3)*. Pleasanton, CA. Retrieved from https://assets.ctfassets.net/an68im79xiti/5vFcknZbAerfkkeZFPwKsH/8694762f1f04aa8a829bac18c61cbf16/CG000360_GettingStartedSingleCell3-GeneExpression_RevA.pdf. (Accessed December 1, 2021).

10x Genomics, Inc. (2021b). *Inside Visium Spatial Capture Technology (Rev D) [pdf]*. Pleasanton, CA. Retrieved from https://pages.10xgenomics.com/rs/446-PBO-704/images/10x_BR060_Inside_Visium_Spatial_Technology.pdf. (Accessed December 1, 2021).

Adam, M., Potter, A. S., and Potter, S. S. (2017). Psychrophilic Proteases Dramatically Reduce Single Cell RNA-Seq Artifacts: A Molecular Atlas of Kidney Development. *Development* 144 (19), 3625–3632. doi:10.1242/dev.151142

AnimalTFDB3 (2021). Animal Transcription Factor Database. Retrieved 20 November 2021, from http://bioinfo.life.hust.edu.cn/AnimalTFDB/#!/download. (Accessed August 11, 2021).

Baker, B. B., Yee, J. S., Meyer, D. N., Yang, D., and Baker, T. R. (2016). Histological and Transcriptomic Changes in Male Zebrafish Testes Due to Early Life Exposure to Low Level 2,3,7,8-Tetrachlorodibenzo-P-Dioxin. *Zebrafish* 13 (5), 413–423. doi:10.1089/zeb.2016.1275

Baker, T. R., Peterson, R. E., and Heideman, W. (2013). Early Dioxin Exposure Causes Toxic Effects in Adult Zebrafish. *Toxicol. Sci.* 135 (1), 241–250. doi:10.1093/toxsci/kft144

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for Technical Noise in Single-Cell RNA-Seq Experiments. *Nat. Methods* 10 (11), 1093–1095. doi:10.1038/nmeth.2645

Bruner-Tran, K. L., Mokshagundam, S., Barlow, A., Ding, T., and Osteen, K. G. (2019). Paternal Environmental Toxicant Exposure and Risk of Adverse Pregnancy Outcomes. *Curr. Obstet. Gynecol. Rep.* 8 (3), 103–113. doi:10.1007/s13669-019-00265-w

Canipari, R., De Santis, L., and Cecconi, S. (2020). Female Fertility and Environmental Pollution. *Int. J. Environ. Res. Public Health* 17 (23), 8802. doi:10.3390/ijerph17238802

Chen, G., Ning, B., and Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.* 10, 317. doi:10.3389/fgene.2019.00317

Davies, P., Jones, M., Liu, J., and Hebenstreit, D. (2021). Anti-bias Training for (sc) RNA-Seq: Experimental and Computational Approaches to Improve Precision. *Brief. Bioinformatics* 22 (6), bbab148. doi:10.1093/bib/bbab148

Deng, Q., Han, G., Puebla-Osorio, N., Ma, M. C. J., Strati, P., Chasen, B., et al. (2020). Characteristics of Anti-CD19 CAR T Cell Infusion Products Associated with Efficacy and Toxicity in Patients with Large B Cell Lymphomas. *Nat. Med.* 26 (12), 1878–1887. doi:10.1038/s41591-020-1061-7

Denisenko, E., Guo, B. B., Jones, M., Hou, R., de Kock, L., Lassmann, T., et al. (2019). Systematic Assessment of Tissue Dissociation and Storage Biases in Single-Cell and Single-Nucleus RNA-Seq Workflows. *Genome Biol.* 21 (1), 130. doi:10.1186/s13059-020-02048-6

Eskenazi, B., Warner, M., Brambilla, P., Signorini, S., Ames, J., and Mocarelli, P. (2018). The Seveso Accident: A Look at 40 Years of Health Research and Beyond. *Environ. Int.* 121, 71–84. doi:10.1016/j.envint.2018.08.051

Fader, K. A., Nault, R., Zhang, C., Kumagai, K., Harkema, J. R., and Zacharewski, T. R. (2017). 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD)-elicited Effects on Bile Acid Homeostasis: Alterations in Biosynthesis, Enterohepatic Circulation, and Microbial Metabolism. *Sci. Rep.* 7 (1), 5921. doi:10.1038/s41598-017-05656-8

Federico, A., Serra, A., Ha, M. K., Kohonen, P., Choi, J.-S., Liampa, I., et al. (2020). Transcriptomics in Toxicogenomics, Part II: Preprocessing and Differential Expression Analysis for High Quality Data. *Nanomaterials* 10 (5), 903. doi:10.3390/nano10050903

Fletcher, N., Wahlström, D., Lundberg, R., Nilsson, C. B., Nilsson, K. C., Stockling, K., et al. (2005). 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) Alters the mRNA Expression of Critical Genes Associated with Cholesterol Metabolism, Bile Acid Biosynthesis, and Bile Transport in Rat Liver: A Microarray Study. *Toxicol. Appl. Pharmacol.* 207 (1), 1–24. doi:10.1016/j.taap.2004.12.003

Green, C. D., Ma, Q., Manske, G. L., Shami, A. N., Zheng, X., Marini, S., et al. (2018). A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq. *Develop. Cel.* 46 (5), 651–667. e10. doi:10.1016/j.devcel.2018.07.025

Guo, J., Grow, E. J., Mlcochova, H., Maher, G. J., Lindskog, C., Nie, X., et al. (2018). The Adult Human Testis Transcriptional Cell Atlas. *Cell Res.* 28 (12), 1141–1157. doi:10.1038/s41422-018-0099-2

Haimbaugh, A., Akemann, C., Meyer, D. N., Gurdziel, K., and Baker, T. R. (2022). Insight into 2,3,7,8-tetrachlorodibenzo-p-dioxin-induced Disruption of Zebrafish Spermatogenesis via Single Cell RNA-seq. PNAS Nexus, accepted.

Hao, Y., Hao, S., Andersen-Nissen, E. W. M. M., III, Mauck, W. M., Zheng, S., Butler, A., et al. (2021). Integrated Analysis of Multimodal Single-Cell Data. *Cell* 184, 3573–3587. doi:10.1016/j.cell.2021.04.048

Henry, T. R., Spitsbergen, J. M., Hornung, M. W., Abnet, C. C., and Peterson, R. E. (1997). Early Life Stage Toxicity of 2,3,7,8-Tetrachlorodibenzo-P-Dioxin in Zebrafish (Danio rerio). *Toxicol. Appl. Pharmacol.* 142 (1), 56–68. doi:10.1006/taap.1996.8024

Hérault, L., Poplineau, M., Mazuel, A., Platet, N., Remy, É., and Duprez, E. (2021). Single-cell RNA-Seq Reveals a Concomitant Delay in Differentiation and Cell Cycle of Aged Hematopoietic Stem Cells. *BMC Biol.* 19 (1), 19. doi:10.1186/s12915-021-00955-z

Hsu, K. S., Goodale, B. C., Ely, K. H., Hampton, T. H., Stanton, B. A., and Enelow, R. I. (2020). Single-cell RNA-Seq Analysis Reveals that Prenatal Arsenic Exposure Results in Long-Term, Adverse Effects on Immune Gene Expression in Response to Influenza A Infection. *Toxicol. Sci.* 176 (2), 312–328. doi:10.1093/toxsci/kfaa080

Hunt, G. J., Freytag, S., Bahlo, M., and Gagnon-Bartsch, J. A. (2018). Dtangle: Accurate and Robust Cell Type Deconvolution. *Bioinformatics* 35 (12), 2093–2099. doi:10.1093/bioinformatics/bty926

Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA Sequencing Technologies and Bioinformatics Pipelines. *Exp. Mol. Med.* 50 (8), 1–14. doi:10.1038/s12276-018-0071-8

Jung, M., Wells, D., Rusch, J., Ahmad, S., Marchini, J., Myers, S. R., et al. (2019). Unified Single-Cell Analysis of Testis Gene Regulation and Pathology in Five Mouse Strains. *Elife* 8, e43966. doi:10.7554/elife.43966

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian Approach to Single-Cell Differential Expression Analysis. *Nat. Methods* 11 (7), 740–742. doi:10.1038/nmeth.2967

King Heiden, T. C., Spitsbergen, J., Heideman, W., and Peterson, R. E. (2009). Persistent Adverse Effects on Health and Reproduction Caused by Exposure of Zebrafish to 2,3,7,8-Tetrachlorodibenzo-P-Dioxin during Early Development and Gonad Differentiation. *Toxicol. Sci.* 109 (1), 75–87. doi:10.1093/toxsci/kfp048

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven Grand Challenges in Single-Cell Data Science. *Genome Biol.* 21 (1), 31. doi:10.1186/s13059-020-1926-6

Lettieri, T. (2006). Recent Applications of DNA Microarray Technology to Toxicology and Ecotoxicology. *Environ. Health Perspect.* 114 (1), 4–9. doi:10.1289/ehp.8194

Luecken, M. D., and Theis, F. J. (2019). Current Best Practices in Single-cell RNA-seq Analysis: a Tutorial. *Mol. Syst. Biol.* 15 (6), e8746. doi:10.15252/msb.20188746

Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016). A Step-by-step Workflow for Low-Level Analysis of Single-Cell RNA-Seq Data with Bioconductor. *F1000Res* 5, 2122. doi:10.12688/f1000research.9501.2

Machado, L., Geara, P., Camps, J., Dos Santos, M., Teixeira-Clerc, F., Van Herck, J., et al. (2021). Tissue Damage Induces a Conserved Stress Response that Initiates Quiescent Muscle Stem Cell Activation. *Cell Stem Cell* 28 (6), 1125–1135. e7. doi:10.1016/j.stem.2021.01.017

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161 (5), 1202–1214. doi:10.1016/j.cell.2015.05.002

Marshall, N. B., and Kerkvliet, N. I. (2010). Dioxin and Immune Regulation. *Ann. N.Y Acad. Sci.* 1183 (1), 25–37. doi:10.1111/j.1749-6632.2009.05125.x

Meyer, D. N., Baker, B. B., and Baker, T. R. (2018). Ancestral TCDD Exposure Induces Multigenerational Histologic and Transcriptomic Alterations in

Gonads of Male Zebrafish. *Toxicol. Sci.* 164 (2), 603–612. doi:10.1093/toxsci/kfy115

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry. *Nat. Biotechnol.* 37 (7), 773–782. doi:10.1038/s41587-019-0114-2

O'Flanagan, C. H., Campbell, K. R., Zhang, A. W., Kabeer, F., Lim, J. L., Biele, J., et al. (2019). Dissociation of Solid Tumour Tissues with Cold Active Protease for Single-Cell RNA-Seq Minimizes Conserved Collagenase-Associated Stress Responses. *Genome Biol.* 20 (1), 210. doi:10.1186/s13059-019-1830-0

Phipson, B., Zappia, L., and Oshlack, A. (2017). Gene Length and Detection Bias in Single Cell RNA Sequencing Protocols. *F1000Res* 6, 595. doi:10.12688/f1000research.11290.1

Qiu, P. (2020). Embracing the Dropouts in Single-Cell RNA-Seq Analysis. *Nat. Commun.* 11 (1), 1169. doi:10.1038/s41467-020-14976-9

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length mRNA-Seq from Single-Cell Levels of RNA and Individual Circulating Tumor Cells. *Nat. Biotechnol.* 30 (8), 777–782. doi:10.1038/nbt.2282

Sachdeva, K., Do, D. C., Zhang, Y., Hu, X., Chen, J., and Gao, P. (2019). Environmental Exposures and Asthma Development: Autophagy, Mitophagy, and Cellular Senescence. *Front. Immunol.* 10, 2787. doi:10.3389/fimmu.2019.02787

Safe, S., Wang, F., Porter, W., Duan, R., and McDougal, A. (1998). Ah Receptor Agonists as Endocrine Disruptors: Antiestrogenic Activity and Mechanisms. *Toxicol. Lett.* 102-103, 343–347. doi:10.1016/s0378-4274(98)00331-2

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial Reconstruction of Single-Cell Gene Expression Data. *Nat. Biotechnol.* 33, 495–502. doi:10.1038/nbt.3192

Sato, S., Shirakawa, H., Tomita, S., Ohsaki, Y., Haketa, K., Tooi, O., et al. (2008). Low-dose Dioxins Alter Gene Expression Related to Cholesterol Biosynthesis, Lipogenesis, and Glucose Metabolism through the Aryl Hydrocarbon Receptor-Mediated Pathway in Mouse Liver. *Toxicol. Appl. Pharmacol.* 229 (1), 10–19. doi:10.1016/j.taap.2007.12.029

Selvaraju, V., Baskaran, S., Agarwal, A., and Henkel, R. (2020). Environmental Contaminants and Male Infertility: Effects and Mechanisms. *Andrologia* 53 (1), e13646. doi:10.1111/and.13646

Shami, A. N., Zheng, X., Munyoki, S. K., Ma, Q., Manske, G. L., Green, C. D., et al. (2020). Single-Cell RNA Sequencing of Human, Macaque, and Mouse Testes Uncovers Conserved and Divergent Features of Mammalian Spermatogenesis. *Develop. Cel.* 54 (4), 529–547. e12. doi:10.1016/j.devcel.2020.05.010

Snyder, R. (2012). Leukemia and Benzene. *Int. J. Environ. Res. Public Health* 9 (8), 2875–2893. doi:10.3390/ijerph9082875

Sohni, A., Tan, K., Song, H.-W., Burow, D., de Rooij, D. G., Laurent, L., et al. (2019). The Neonatal and Adult Human Testis Defined at the Single-Cell Level. *Cel. Rep.* 26 (6), 1501–1517. e4. doi:10.1016/j.celrep.2019.01.045

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq Whole-Transcriptome Analysis of a Single Cell. *Nat. Methods* 6 (5), 377–382. doi:10.1038/nmeth.1315

Tanos, R., Patel, R. D., Murray, I. A., Smith, P. B., Patterson, A. D., and Perdew, G. H. (2012). Aryl Hydrocarbon Receptor Regulates the Cholesterol Biosynthetic Pathway in a Dioxin Response Element-independent Manner. *Hepatology* 55 (6), 1994–2004. doi:10.1002/hep.25571

van den Brink, S. C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C. S., et al. (2017). Single-cell Sequencing Reveals Dissociation-Induced Gene Expression in Tissue Subpopulations. *Nat. Methods* 14 (10), 935–936. doi:10.1038/nmeth.4437

Volpe, J. J., Goldberg, R. I., and Bhat, N. R. (1985). Cholesterol Biosynthesis and its Regulation in Dissociated Cell Cultures of Fetal Rat Brain: Developmental Changes and the Role of 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase. *J. Neurochem.* 45 (2), 536–543. doi:10.1111/j.1471-4159.1985.tb04021.x

Wang, J.-J., Tian, Y., Li, M.-H., Feng, Y.-Q., Kong, L., Zhang, F.-L., et al. (2021). Single-cell Transcriptome Dissection of the Toxic Impact of Di (2-ethylhexyl) Phthalate on Primordial Follicle Assembly. *Theranostics* 11 (10), 4992–5009. doi:10.7150/thno.55006

Wang, M., Liu, X., Chang, G., Chen, Y., An, G., Yan, L., et al. (2018). Single-Cell RNA Sequencing Analysis Reveals Sequential Cell Fate Transition during Human Spermatogenesis. *Cell Stem Cell* 23 (4), 599–614. e4. doi:10.1016/j.stem.2018.08.007

Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019). Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference. *Nat. Commun.* 10 (1), 380. doi:10.1038/s41467-018-08023-x

Warren, T. K. (2000). Exposure to 2,3,7,8-Tetrachlorodibenzo-P-Dioxin (TCDD) Suppresses the Humoral and Cell-Mediated Immune Responses to Influenza A Virus without Affecting Cytolytic Activity in the Lung. *Toxicol. Sci.* 56 (1), 114–123. doi:10.1093/toxsci/56.1.114

Westerfield, M. (2000). *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio Rerio)*. 4th ed. Eugene: University of Oregon Press.

Wild, C. P. (2005). Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiol. Biomarkers Prev.* 14 (8), 1847–1850. doi:10.1158/1055-9965.epi-05-0456

Wilson, C. L., and Safe, S. (1998). Mechanisms of Ligand-Induced Aryl Hydrocarbon Receptor-Mediated Biochemical and Toxic Responses. *Toxicol. Pathol.* 26 (5), 657–671. doi:10.1177/019262339802600510

Wu, H., Kirita, Y., Donnelly, E. L., and Humphreys, B. D. (2018). Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *J. Am. Soc. Nephrol.* 30 (1), 23–32. doi:10.1681/asn.2018090912

Yang, H., Ma, J., Wan, Z., Wang, Q., Wang, Z., Zhao, J., et al. (2020). Characterization of Sheep Spermatogenesis through Single-cell RNA Sequencing. *FASEB J.* 35 (2), e21187. doi:10.1096/fj.202001035rrr

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., et al. (2019). Comparative Analysis of Droplet-Based Ultra-high-throughput Single-Cell RNA-Seq Systems. *Mol. Cel.* 73 (1), 130–142. e5. doi:10.1016/j.molcel.2018.10.020

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively Parallel Digital Transcriptional Profiling of Single Cells. *Nat. Commun.* 8, 14049. doi:10.1038/ncomms14049

Zou, Q., Lv, Y., Gan, Z., Liao, S., and Liang, Z. (2021). Identification and Validation of a Malignant Cell Subset Marker-Based Polygenic Risk Score in Stomach Adenocarcinoma through Integrated Analysis of Bulk and Single-Cell RNA Sequencing Data. *Front. Cel. Dev. Biol.* 9, 720649. doi:10.3389/fcell.2021.720649

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Haimbaugh, Meyer, Akemann, Gurdziel and Baker. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

REVIEW
published: 16 May 2022
doi: 10.3389/fgene.2022.867880

Breaking the Immune Complexity of the Tumor Microenvironment Using Single-Cell Technologies

*Simone Caligola, Francesco De Sanctis, Stefania Canè and Stefano Ugel**

Immunology Section, Department of Medicine, University of Verona, Verona, Italy

Tumors are not a simple aggregate of transformed cells but rather a complicated ecosystem containing various components, including infiltrating immune cells, tumor-related stromal cells, endothelial cells, soluble factors, and extracellular matrix proteins. Profiling the immune contexture of this intricate framework is now mandatory to develop more effective cancer therapies and precise immunotherapeutic approaches by identifying exact targets or predictive biomarkers, respectively. Conventional technologies are limited in reaching this goal because they lack high resolution. Recent developments in single-cell technologies, such as single-cell RNA transcriptomics, mass cytometry, and multiparameter immunofluorescence, have revolutionized the cancer immunology field, capturing the heterogeneity of tumor-infiltrating immune cells and the dynamic complexity of tenets that regulate cell networks in the tumor microenvironment. In this review, we describe some of the current single-cell technologies and computational techniques applied for immune-profiling the cancer landscape and discuss future directions of how integrating multi-omics data can guide a new "precision oncology" advancement.

Keywords: immune system, single-cell technologies, cancer, tumor microenvironment, single-cell data analysis

OPEN ACCESS

Edited by:
Rui Henrique,
Portuguese Oncology Institute, Portugal

Reviewed by:
Eleonora Lusito,
San Raffaele Telethon Institute for Gene Therapy (SR-Tiget), Italy
Wenji Ma,
Columbia University, United States
Maria J. Oliveira,
University of Porto, Portugal

****Correspondence:***
Stefano Ugel
stefano.ugel@univr.it

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *01 February 2022*
Accepted: *27 April 2022*
Published: *16 May 2022*

Citation:
Caligola S, De Sanctis F, Canè S and Ugel S (2022) Breaking the Immune Complexity of the Tumor Microenvironment Using Single-Cell Technologies. Front. Genet. 13:867880. doi: 10.3389/fgene.2022.867880

INTRODUCTION

Cancer is one of the leading causes of death worldwide, accounting for nearly 10 million deaths in 2020 (https://www.who.int/news-room/fact-sheets/detail/cancer). Unfortunately, the pandemic COVID-19 will have consequences for cancer patients in coming years, since it has been associated with delays in diagnosis as well as interruption of therapeutic treatments and follow-up care. Hence the number of cancer victims will increase in the near future. Identifying cancer as a genetic disease characterized by a set of genomic aberrations, including in-frame insertions or deletions, missense amino acid changes, and large copy number variations, initially ingrained a "cancer cell-centric" vision in the scientific community where cancer cell-intrinsic properties exclusively drove tumorigenesis (Reddy et al., 1982; Santos et al., 1984; Lengauer et al., 1998; Futreal et al., 2004; Tomlins et al., 2005; Hanahan and Weinberg, 2011). Therefore, recognizing driver gene modifications has been a central aim of cancer research over the past 30 years, resulting in global initiatives such as The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga) and the Catalog of Somatic Mutations in Cancer (COSMIC) (Tate et al., 2019) which include a broad collection of large-scale, systematic sequencing studies that constitute comprehensive catalogs of mutational abnormalities in the major tumor types. More than 100,000 somatic mutations in cancer genomes have been identified in the quarter-century since the first somatic mutation was reported (Zack et al., 2013), opening a new age for the classification and treatment protocols of some cancers, such as breast (Cancer Genome Atlas,

2012b; Ciriello et al., 2015), cervical (Cancer Genome Atlas Research Network, 2017), colorectal (Cancer Genome Atlas, 2012a), pancreatic (Consortium, 2020), gastric (Cancer Genome Atlas Research, 2014a), prostate (Cancer Genome Atlas Research, 2015), and lung carcinomas (Cancer Genome Atlas Research, 2012; Cancer Genome Atlas Research, 2014b).

In addition to the study of genetic diversity as a source of tumor cell heterogeneity, stable epigenetic changes in cancer have been received intense interest since they derive from poised or initiated chromatin states of several genes that can modulate expression of different pathways (Brown et al., 2014). Indeed, alterations of cellular activities, such as cell growth and differentiation, can be driven by epigenetic events that involve DNA methylation, histone modification, the readout of these modifications, chromatin remodeling and the effects of noncoding RNA (Feinberg et al., 2016). These alterations are temporary or yet long-lasting and they impact on tumorigenesis. For instance, aberrant DNA methylation has been, generally, associated with cancer development by inactivating gene transcription or repressing gene transcription and affecting chromatin structure (Baylin and Jones, 2011; Hanahan and Weinberg, 2011). Indeed, some gene promoters, especially key tumor suppressor genes, are unmethylated in normal tissues and highly methylated in cancer (Baylin and Jones, 2011). Interestingly, DNA methylation in cancer has generally been associated with tumor size (Feinberg et al., 2016) as well as with drug resistance and predicting response to treatment (Brown et al., 2014). Therefore, the comprehensive genetic and epigenetic analysis of cancer genomes has been for many years the most effective way to identify causative changes involved in tumorigenesis.

Despite the achievement of these extraordinary milestones in deciphering cancer cell biology, new insights demonstrate that aberrant genetic profiles of transforming cells alone are required but insufficient to nurse tumor development and progression. Indeed, cancer cells need to alter the stromal framework of the microenvironment to manipulate diverse physiological processes, such as promoting angiogenesis and vasculogenesis, to provide adequate nutrients and oxygen, and alter immune responses to avoid activating tumor-fighting elements, such as cytotoxic T lymphocytes (CTLs) (Hanahan and Weinberg, 2011). Therefore, "cancer-cell extrinsic" factors, such as local inflammation, metabolic switch, and immunity, are critical in fueling cancer growth. At this point, the key question is whether these extrinsic factors are independent of the genetic profile of cancer cells. A cornerstone study demonstrated a remarkable difference in the composition of tumor-infiltrating leukocytes in different tumor types by analyzing data on clinical outcomes and gene expression of 18,000 human tumors (Gentles et al., 2015). Interestingly, memory $CD4^+$ T lymphocyte frequency correlated positively with a favorable outcome in lung cancer patients, whereas the same cell subset was associated with a worse outcome in patients affected with bladder tumors (Gentles et al., 2015), suggesting that cancer cell-intrinsic features can dictate the immune landscape of the tumor microenvironment (TME). Indeed, oncogene-driven modifications can alter tumor immunogenicity in a completely different way. Ongoing mutational processes generate either cancer neoantigens capable of activating tumor-eliminating immune cells (Balachandran et al., 2017; Luksza et al., 2017; Keenan et al., 2019) or produce immune soluble factors such as interleukins (e.g., IL-6, IL1-β), growth factors (e.g., granulocyte-macrophage colony-stimulating factor (GM-CSF)), and chemokines (e.g., CCL4), capable of differentiating immune cells into pro-tumor elements (Bronte et al., 2006; Pylayeva-Gupta et al., 2012; Calcinotto et al., 2018; Fiore et al., 2018; Vitale et al., 2019; Hofer et al., 2021). Host immune cells also influence cancer progression (Schreiber et al., 2011). The notion that tumors derived from immunodeficient hosts are more immunogenic than those derived from immunocompetent mice allowed us to hypothesize that the immune system actively shapes cancer cells, promoting the acquisition of genetic aberrations that can compromise cancer cell immunogenicity, favoring tumors escaping immune attacks. Several clinical and pre-clinical observations have validated the cancer immunoediting theory. For instance, two large studies demonstrated that tumor-infiltrating immune subsets in colorectal cancer were significant independent prognostic markers as well as microsatellite instability, long interspersed nucleotide element-1 (LINE-1) hypomethylation, and BRAF mutations (Ogino et al., 2009; Nosho et al., 2010). Moreover, the linear correlation between the density of effector memory CTLs at the site of the primary tumor and the survival of patients explicitly revealed the importance of TME immunity in cancer control (Galon et al., 2006; Fridman et al., 2017). Thus, the definition of the cancer genome and immune landscape of the TME is hierarchically equivalent and complementary to predict disease progression and therapeutic outcome.

Immune cell heterogeneity and the presence of various cell subsets complicate TME immune profiling (Binnewies et al., 2018). Moreover, the spatial distribution of immunity within the tumor mass is a critical parameter that significantly influences immune cells acquiring pro- or anti-tumor functions (Bindea et al., 2013; Sautes-Fridman et al., 2019). A multi-omic perspective considering genomics, transcriptomics, epigenomics and proteomics is necessary to unveil the immune complexity of the TME (**Figures 1A–D**). Unfortunately, traditional technologies used to evaluate the immune landscape of the TME, such as immunohistochemistry (IHC), flow cytometry (FC), and bulk analysis of genomic, transcriptional, and proteomic analyses, have several limitations, such as large amount of biological material requirements, fewer parameters tested simultaneously, and low analysis resolution (Peregrin et al., 1973; Bendall et al., 2012; Abel et al., 2014). Recently, high-dimensional single-cell techniques (Finotello and Eduati, 2018; Chen G. et al., 2019; Gohil et al., 2021) have revolutionized the approach to decipher the cellular diversity, cell interactions and dynamics that exists in the TME and heterogeneity across patients by resolving cell subset complexity at the single-cell level using unsupervised clustering to identify potential unknown subpopulations of cells within the populations under study (Thorsson et al., 2019; Wagner et al., 2019; Zhao J. et al., 2020; Kieffer et al., 2020; Cheng et al., 2021). Here, we review how single-cell technologies (**Figure 1E**) and related computational techniques have improved our knowledge

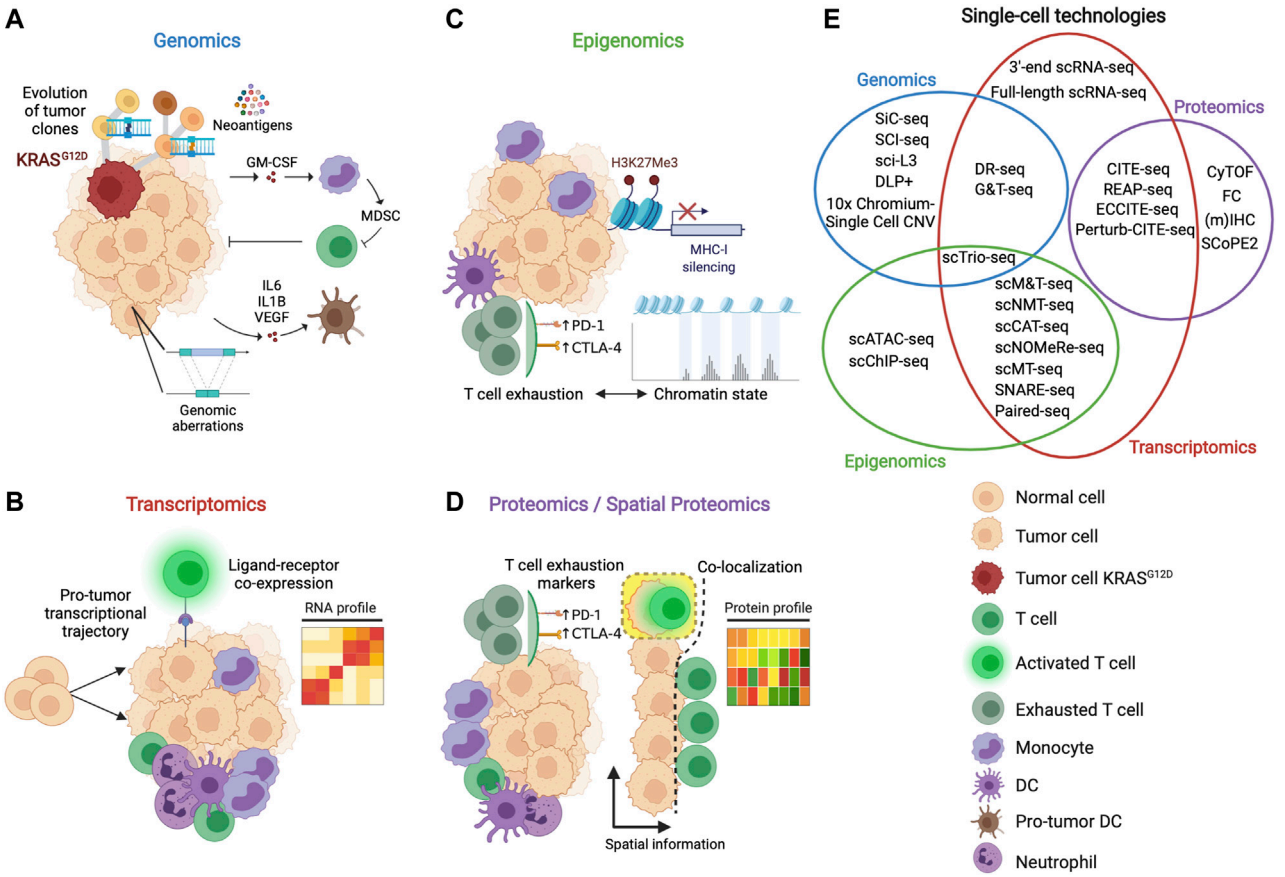


FIGURE 1 | Multi-omic perspective to study the features of the TME. **(A)** Genomics analysis informs about how the tumor mutational landscape influences the TME, favoring, for example, the production of cytokines (e.g., IL-6, IL1-β) and growth factors (e.g., GM-CSF) inducing the proliferation of suppressive myeloid cells and the pro-tumor differentiation of antigen-presenting cells. **(B)** Transcriptomic analysis enables to inspect the transcriptional machinery of the single cells of the TME, deciphering, for example, developmental trajectories, cell states and cell-cell interactions. **(C)** Epigenomic analysis reveals how specific switches such as histone methylation and chromatin dynamics regulate different mechanisms capable to interfere with anti-tumor immune recognition and effector functions. **(D)** Proteomics provides information about the state of activation of the immune cells of the TME looking at the expression of immunomodulatory proteins such as checkpoint inhibitors (e.g., PD-L1, CTLA-4). Spatial proteomics gives additional information about the localization of the cells allowing, for example, to identify cell that interact in the TME. **(E)** Venn diagram depicting single-cell technologies to study single (non-overlapping sets) or multimodal (overlapping sets) omics. Genomics, transcriptomics, epigenomics and proteomics are represented as blue, red, green, and purple sets, respectively.

about the TME and discuss future applications of these cutting-edge techniques in immune-oncology to develop more effective personalized immunotherapy.

SINGLE-CELL TECHNOLOGIES TO STUDY THE TME

Single-Cell Genomics

Several studies have demonstrated the connection between the tumor mutations and the immune composition or TME. Deciphering the complexity of the mutational landscape of tumor cell clones and sub-clones that underlies intratumor heterogeneity is necessary to understand tumor patient lethal outcomes, therapeutic failures, and drug resistance (McGranahan and Swanton, 2017). Single-cell genomics technologies can make an important contribution to this goal. However, the study of tumor clonality by analyzing single-nucleotide variations (SNVs) and copy number variations (CNVs) at the single-cell level is challenging. Most methods for creating single-cell libraries rely on whole-genome amplification (WGA) to overcome the inability of sequencing technologies to capture single-cell DNA molecules in low amounts of material. Since the first method based on degenerate oligonucleotide-primed PCR (DOP-PCR) (Telenius et al., 1992), several methods have been developed to increase the coverage and uniformity of the genome to allow both SNVs and CSVs to be studied within the same experiment (Dean et al., 2001; Zong et al., 2012), given the importance of the WGA step. More recent methods based on tagmentation (Adey et al., 2010), such as direct library preparation (DLP) (Zahn et al., 2017), and linear amplification via transposon insertion (LIANTI) (Chen et al., 2017), outperformed WGA methods in terms of

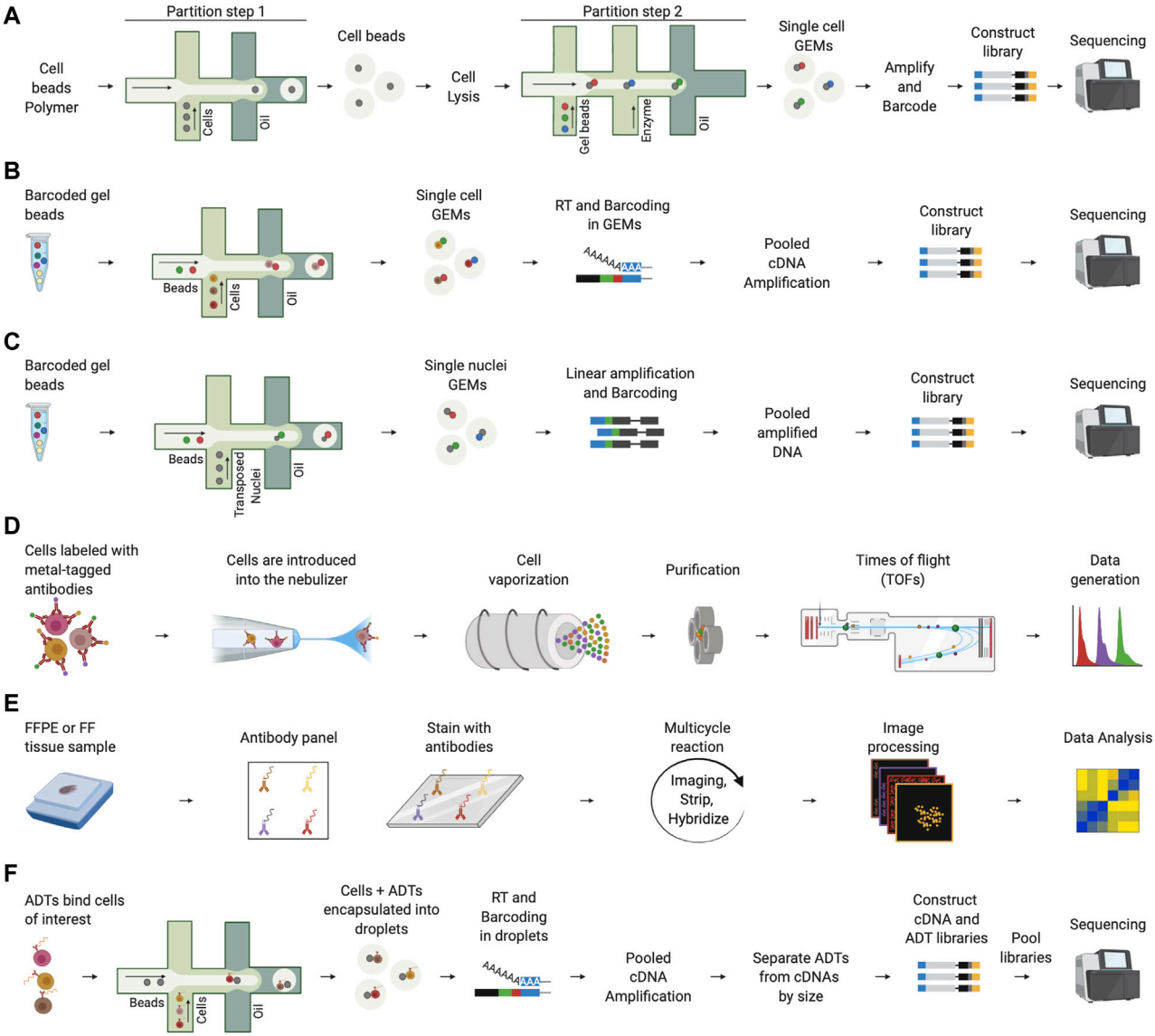


FIGURE 2 | Representative single-cell technologies to study the TME. **(A)** 10x Genomics single-cell genomics involves two steps of encapsulation using the microfluidics system. In the first step, cells are partitioned using a cell beads polymer. The obtained cell beads are lysed to denature the genomic DNA and a second step on microfluidics chip is performed to encapsulate cell beads with barcode gel beads. After collecting single cell GEMs, amplification and barcoding of fragments is performed prior to breaking the emulsion and constructing the library for sequencing. **(B)** In 10x Genomics scRNA-seq, cells are encapsulated into droplets together with barcoded beads. Next, reverse transcription (RT) is performed in the collected GEMs and barcoded cDNAs are amplified for library construction and sequencing. **(C)** In 10x Genomics scATAC-seq, nuclei are transposed and encapsulated into droplets using the microfluidics chip. Next, the collected single nuclei GEMs are linearly amplified, and barcoded accessible DNA fragments are obtained after breaking the emulsion. Finally, DNA fragments are ready for library construction and sequencing. **(D)** In CyTOF, the cells are labeled using stable heavy metals, nebulized, and vaporized to form ion clouds through an argon plasma torch. Each cloud passes through a quadrupole which performs a purification step, the remaining heavy ions are quantified by a time-of-flight (TOF) mass spectrometer that determines the value of each marker. **(E)** In co-detection by indexing (CODEX), FFPE or FF tissues samples are stained with DNA-barcoded antibodies. Next, a multicycle reaction characterized by iteratively imaging up to three antibodies and nuclear stain, stripping and hybridizing is performed. This process is performed for all antibodies. Finally, raw images are processed and analyzed. **(F)** In CITE-seq, antibody-derived tags (ADTs) are used to bind the cells of interest. Next, cells are incapsulated into droplets using a microfluidics platform and after cell lysis in droplets, mRNAs and ADTs are barcoded during the RT. After amplification, cDNAs and ADTs are separated by size, converted into two independent libraries that are, finally, pooled, and sequenced.

accuracy, time, and cost, expanding potential future application scenarios.

Finally, high-throughput methods capable of processing thousands of single cells per experiment were introduced, such as those based on microfluidics and barcoding (SiC-seq) (Lan et al., 2017), split-pool strategies (SCI-seq) (Vitak et al., 2017), high-throughput versions of linear amplification via transposon insertion (sci-L3) (Yin et al., 2019), and direct library preparation

(DLP+) (Laks et al., 2019). The aim of high-throughput methods was introducing automation to increase the number of cells analyzed while maintaining accuracy of genomic information. SiC-seq exploits droplet microfluidics to encapsulate cells into microspheres in which to perform the reactions required for cell and genome processing without compromising genomic DNA. SiC-seq performs a series of steps to lyse the cells, fragment the genomes, barcode the DNA fragments and sequence them after library preparation. SCI-seq uses the strategy of transposase-based combinatorial indexing (Adey et al., 2010) to obtain barcoded libraries without using droplet microfluidics. Sci-L3 addresses the problem of genomic artifacts due to PCR amplification and low-throughput using respectively linear amplification and a three-level combinatorial indexing. DLP + takes advantage of specialized hardware and software for imaging microscopy to capture genomic information of thousands of cells per experiment. Given the complexity in obtaining single cell genomic information, commercial solutions such as those based on the 10x Genomics droplet microfluidics system have been also proposed to simplify single-cell genomics data generation (**Figure 2A**). Single-cell genomic technologies have been used, for example, for the deconvolution of clonal cell clusters and tracing the evolutionary trajectories of clonal breast cancer cells (Wang et al., 2014), the identification of structural and mutational events of melanoma cell line clones (Velazquez-Villarreal et al., 2020), and for an in-depth view of the intratumoral copy number alteration (CNA) heterogeneity present in breast cancer genomes (Baslan et al., 2020).

All these aspects offer a greater understanding of the dynamics of the TME, especially in the cancer therapy response context. For example, identifying clonal and sub-clonal cell composition or the presence of specific mutated subsets would be useful for stratifying patients to understand whether they could benefit from immunotherapy. In the future, longitudinal studies coupled with other single-cell omics could help to construct a more detailed map of clone evolution kinetics and elucidate the time-dependent mechanisms underlying therapy and the emergence of more aggressive clones caused by selection pressures (McGranahan and Swanton, 2015). Furthermore, these studies could also suggest evolutionary time points with a favorable TME to perform more effective therapeutic interventions. In the future, the widespread use of single-cell genomics methods will depend on improvements of the actual technical limitations. Current technologies are not optimal to study all the different genomic aberrations such as copy number variations (CNVs), small indels, single-nucleotide variations and structural variations (SVs) at the same time (Fan et al., 2021). Advancements on the current protocols and bioinformatics solutions will be critical for a wider adoption of single-cell genomics by the scientific community.

Single-Cell Transcriptomics

Single-cell RNA-sequencing (scRNA-seq) is a key technique to explore the TME. scRNA-seq has expanded the scenarios opened by previous bulk RNA-seq technology to investigate the transcriptome of single cells inside a biological sample. Despite countless discoveries in the last decade due to the ability to sequence millions of RNA fragments from a "bulk" of cells, it is now mandatory to look at the transcription machinery of each single cell to understand the complexity of the TME. Indeed, the possibility of observing the expression of thousands of genes for each cell has allowed us to resolve cell subset heterogeneity of the TME in an unbiased way (Lambrechts et al., 2018; Zilionis et al., 2019). Furthermore, scientists can go beyond merely characterizing cell compositions and obtain information about cell state, differentiation trajectories, and cellular pathway activation (Qian et al., 2020; Zhang et al., 2020; Raghavan et al., 2021).

The TME is crucial for tumor cells to evade immune surveillance, necessitating detailed identities of cancer, immune, and non-immune cells. A central aspect in the fight against cancer requires an enhanced understanding of the role of immune cells in cancer therapies, especially in immunotherapy based on immune checkpoint inhibitors (ICIs). Hence, in recent years, scientists have tried to understand the reasons for the success or failure of immunotherapy in relation to the immune features of the TME. It is now clear that T cell infiltration into tumor tissues is a key feature of the immunotherapy response (Tang et al., 2016). However, the presence of T lymphocytes is a necessary but insufficient condition for an effective antitumor response. Indeed, the immunotherapeutic reactivity of a tumor largely depends on the functional state of infiltrated T cells and the expression of specific drug-targetable molecules, such as programmed cell death protein 1 (PD-1) and cytotoxic T lymphocyte-associated antigen 4 (CTLA-4), as well as and the lymphocyte-activation gene 3 (LAG-3), which negatively regulates T-cell proliferation and effector T-cell functions. Recently, a phase II/III, global, double-blind, randomized study named RELATIVITY-047, demonstrated that the dual inhibition of LAG-3 and PD-1, using relatlimab and nivolumab, had a synergistic effect on progression-free survival in melanoma patients highlighting the possibility to develop more effective therapy by targeting different T cell activation brakes (Tawbi et al., 2022). Furthermore, the expression of co-stimulatory molecules, such as programmed death-ligand 1 (PD-L1), by cancer cells and an inflamed TME are crucial elements for an effective immunotherapy response (Marigo et al., 2016; Duan et al., 2020; Ugel et al., 2021). In contrast, the presence of myeloid-infiltrating cells with immunosuppressive features, such as myeloid-derived suppressor cells (MDSCs) (De Sanctis et al., 2016) and a low frequency of infiltrating $CD8^+$ T cells in the TME, correlate with tumors averse to immunotherapy (Zhu et al., 2017; Li J. et al., 2018).

Cancer immunology discoveries have continued in parallel with advancements in scRNA-seq technologies. Among others, methods based on droplet microfluidics and cellular barcoding, such as InDrop (Klein et al., 2015), Drop-seq (Macosko et al., 2015), and the 10x Genomics platform (Zheng et al., 2017) have gained attention in recent years. Similarly to single-cell genomics technologies, their success is due to the greater scalability offered by droplet microfluidics platforms that allow to isolate thousands of cells in a short time (**Figure 2B**). A drawback of these technologies is their low transcript coverage. Indeed, these methods cannot be used to study, for example, splicing and enhancer RNAs at the single-cell level (Hayashi et al., 2018) because they sequence only the 3' end of each transcript. In contrast, full-length technologies, such as SMART-seq (Ramskold et al., 2012; Picelli et al., 2014) and MATQ-seq (Sheng et al., 2017), have a higher accuracy level in

transcript detection, allowing the study of gene variants and splicing events at the cost of lower single-cell profiling throughput.

Given the complexity of the TME in terms of cellular heterogeneity, high-throughput scRNA-seq techniques have been widely used in cancer studies because they can extract thousands of cells from a biological sample. scRNA-seq has been successfully used to decipher several mechanisms by which the tumor can shape the microenvironment towards a hostile ecosystem characterized by exhausted infiltrating T cells, suppressive myeloid cell subsets, cells with pro-tumor differentiation programs, and aberrant cell-cell interaction networks (Guo et al., 2018; Elyada et al., 2019; Davidson et al., 2020; Kim et al., 2020; Marigo et al., 2020; Wu et al., 2021; De Sanctis et al., 2022). The integration of scRNA-seq data with clinical patient information has further allowed us to define cellular and molecular signatures of the TME to explain the response to immunotherapy or survival outcome in different types of cancer (Jerby-Arnon et al., 2018; Sade-Feldman et al., 2018; Ma et al., 2019; Peng et al., 2019; Di Pilato et al., 2021; Zhang et al., 2021). These results motivate future applications of scRNA-seq as a valuable tool for clinicians to perform TME screening, stratification, and identification of new druggable targets based on the integration of different patient datasets (Giladi and Amit, 2017; Binnewies et al., 2018). Data integration is of primary importance because transcriptional-level results cannot be directly translated to the functional level, requiring additional support and experimental validation at the protein level.

Recently, scRNA-seq has been combined with the innovative spatial transcriptomics (Zhang et al., 2022). This technology has revolutionized the traditional RNA-fluorescence *in situ* hybridization (FISH) tool able to identify target messenger RNA transcripts in tissue sections by-passing the absence of selective antibodies for unknown candidates (Facciponte et al., 2014; Cui et al., 2016). Indeed, this innovative approach allows to visualize profiles of RNA molecules in identified tissue regions, including technologies based on micro-dissected gene expression, *in situ* hybridization, *in situ* capturing and *in situ* sequencing technologies (Calvanese et al., 2022; Nirmal et al., 2022; Wei et al., 2022). This technology continues to aid the development of human cellular atlases of cancer, the reclassification of the immune landscape of TME, and overall, the identification of important therapeutic targets.

Single-Cell Epigenomics

The epigenome is the complete atlas of chemical modifications that can induce changes in gene expression without modifying the DNA sequence. Such modifications involve DNA, RNA, and histone proteins and can cause chromatin remodeling that can turn genes "on" or "off". Methylation and acetylation of histones on lysine and arginine residues are the best-known epigenetic mechanisms capable of enhancing or repressing gene transcription, as exemplified by histone H3 on lysine 27 (H3K27ac) and histone H3 on lysine 9 (H3K9me) modifications. Other epigenetic mechanisms that provoke chromatin remodeling include nucleosome positioning/reorganization and DNA methylation.

Several studies have highlighted the ability of tumor cells to induce epigenetic modifications in TME-infiltrating immune cells to aid immune surveillance evasion. The epigenetic strategies implemented by tumors to avoid immune surveillance are based on the disruption of different anti-cancer immune mechanisms, such as immune recognition, signal activation, and effector functions. Common mechanisms exploited by tumors to evade immune recognition and signal triggering include the epigenetic silencing of major histocompatibility complex (MHC) genes and the inhibition of cytokines and chemokines. For example, trimethylation of H3K4 (H3K4me3) and H3K27 (H3K27me3) by polycomb repressive complex 2 (PRC2) has been linked to MHC-I repression and missed tumor recognition by $CD8^+$ T cells mediated by EED and EZH2 (Burr et al., 2019). Furthermore, the lack of tumor-infiltrating lymphocytes (TILs) in several human cancers is associated with DNA methylation-induced epigenetic silencing of *CCL5* (Dangaj et al., 2019). An illuminating example of how epigenetic modifications can result in the loss of immune effector function is the acetylation of H3K27 (H3K27ac). Indeed, high H3K27ac levels have been linked to a TNF-NFKB1 pathway capable of inducing CD47 upregulation and inhibiting macrophage phagocytosis of breast cancer cells (Betancur et al., 2017).

Our understanding of the cancer epigenome has evolved rapidly with the adoption of next-generation sequencing (NGS). In this context, chromatin immunoprecipitation followed by sequencing (ChIP-seq) and an assay for transposase-accessible chromatin using sequencing (ATAC-seq) have been widely used tools to study epigenetic regulation. ChIP-seq is a method for studying the interactions between proteins and DNA. It allows us to analyze chromatin states induced by histone modifications that alter gene transcription. ATAC-seq measures chromatin accessibility by directly deciphering its effects on gene transcription without detailed histone modification and chromatin state characterization. ChIP-seq and ATAC-seq have been used successfully, for example, to link epigenetic features capable of maintaining an immune cell-excluded TME and immunotherapy resistance (Benci et al., 2016; Yang et al., 2021).

The advancements of these techniques are their single-cell counterparts, scChIP-seq and scATAC-seq. Several cancer studies have shown the utility of scChIP-seq or scATAC-seq to study the epigenetic regulators responsible for tumor cell heterogeneity (Grosselin et al., 2019; LaFave et al., 2020; Taavitsainen et al., 2021). Other applications of single-cell epigenomics include adopting scATAC-seq to uncover chromatin regulators responsible for T cell exhaustion in the TME of patients treated with immunotherapy (Satpathy et al., 2019; Zhang et al., 2021). These examples motivate the use of single-cell technologies to study cancer epigenetics.

In the future, it will be pertinent to understand the link between epigenetic changes in the TME and cancer progression and how to obtain a more effective therapy response targeting epigenetic switches. One reason for the growing interest in targeting the epigenome is the possibility of identifying small molecules, such as proteolysis-targeting chimeras (PROTAC) (Sakamoto et al., 2001), that can indirectly interfere with aberrant gene expression, given the

difficulties in targeting oncogenic transcription factors, such as Myc and p53 (Jones et al., 2016).

The dissemination of scChIP-seq studies depends on future technological advances. Despite recent successful attempts to improve scChIP-seq in terms of cellular sensitivity and data sparsity (Kaya-Okur et al., 2019; Wang et al., 2019), the use of this technology remains limited and largely relies on being able to increase the number of sequenced reads per cell. In contrast, scATAC-seq has a simpler and more efficient experimental protocol that requires a lower number of cells. Furthermore, introducing microfluidics approaches (**Figure 2C**) has dramatically improved the throughput compared to the previous technology based on combinatorial cellular indexing (sciATAC-seq) (Cusanovich et al., 2015). However, low per-cell coverage remains a weakness, even for scATAC-seq, potentially limiting the identification of significant open chromatin sites, especially in rare cell subsets (Ma and Zhang, 2020).

Single-Cell Proteomics

Despite the wide adoption of single-cell transcriptomics technologies, single-cell proteomic approaches remain the key tools for studying the functional status of TME cell populations. Indeed, in addition to identifying proteins and related isoforms, they allow us to recognize post-translational modifications that single-cell transcriptomics cannot capture.

FC is still a fundamental multiparametric technique for identifying immune cell subsets within the TME based on morphological characteristics and the expression of certain proteins. FC also allows specific cell populations to be isolated by cell sorting before analysis with other omics techniques. The main limitation of FC is the number of parameters, usually around 20, which can be intercepted simultaneously in the same experiment because of signal overlap (i.e., spillover) between the channels. This limits the number of cell-surface proteins that can be identified and, consequently, the resolution of identifying the related cell subsets present in the experiment.

A more advanced technology is Full Spectrum Flow Cytometry (FSFC), an improvement of spectral flow cytometry (SFC) (Robinson, 2004; Nolan et al., 2013). FSFC exploits high-sensitive light detectors to measure the full spectral profile of fluorophores. This technology has a higher quality and resolution than conventional FC, allowing to design multicolor panels up to 40 parameters useful for characterizing important aspects of the immune context in cancer studies (Bonilla et al., 2020).

Another recent technology for single-cell proteomics is cytometry by time-of-flight (CyTOF) (Bandura et al., 2009) (**Figure 2D**). Compared to FC, CyTOF allows at least 40 markers per cell to be detected in a single run and is more sensitive and less prone to errors (Bandura et al., 2009; Bendall et al., 2011). The disadvantages of CyTOF compared to FC include a lower acquisition flow rate, more critical sample preparation to avoid contamination, and the inability to perform cell sorting for populations of interest due to the final vaporization of the cells (Gadalla et al., 2019). CyTOF potentiates better resolution of TME heterogeneity, which is particularly important in the context of cancer immunology. For example, CyTOF makes finding specific TIL subsets that correlate with patient survival and response to immunotherapy, respectively, in follicular lymphoma (Yang et al., 2020) and melanoma (Subrahmanyam et al., 2018) possible. Furthermore, scientists were able to localize the expression of the T cell inhibitory molecule VISTA in $CD68^+$ macrophages of human pancreatic cancer (Blando et al., 2019) and retrieve information on the composition, expansion, and activity of TILs in patients with non-small cell lung cancer (NSCLC) (Sanmamed et al., 2021). Since CyTOF dissects the cellular composition and the activation status of the immune cells that surround and infiltrate the tumor, it is an effective tool for studying the TME immune landscape. CyTOF and scRNA-seq allowed us to answer similar biological questions. However, these two technologies are not interchangeable. The main limitations of CyTOF are the resolution and bias due to prior parameter selection. These two aspects can limit the discovery of rare cell populations. An obvious advantage of CyTOF is that protein expression indicates specific functional states or activities of the cell that in scRNA-seq must be validated with other techniques. Since CyTOF is simpler than scRNA-seq and allows for better discrimination of certain immune cell subsets (Kashima et al., 2021), it represents a convenient method to monitor TME features, such as cell subsets, activation states, and immune checkpoint molecules in patient cohorts in clinical trials. It is mandatory to rapidly standardize experimental procedures, computational tools, and antibody panels to make results comparable between different institutions to accomplish this aim (Hartmann et al., 2019).

In CyTOF, protein identification strictly depends on the availability of highly specific antibodies and the quality of their interactions. Mass spectrometry-based methods, such as SCoPE-MS and the improved version SCoPE2, have been recently introduced to improve throughput and sensitivity concerning the number of proteins detected in single cells (Budnik et al., 2018; Specht et al., 2021). With SCoPE2, the authors were able to dissect cellular heterogeneity by protein expression and trace the differentiation of monocytes into macrophage-like cells in the absence of specific cytokines. This work is an important step towards future applications of single-cell technologies based on mass spectrometry for TME dissection, looking at the expression of thousands of proteins.

Single-Cell Spatial Proteomics

An important aspect to consider in the study of the TME is its spatial organization and heterogeneity. Indeed, like natural ecosystems, tumor tissues can reveal strong heterogeneity in relatively small spatial distances due to tumor cells adapting to the microenvironment or through its remodeling (Yuan, 2016). Several studies have emphasized the importance of considering spatial heterogeneity in the TME. Some of these reports have shown that the location, density, and spatial distribution of immune cells are more robust markers for predicting patient outcomes than traditional clinical parameters (Galon et al., 2006; Maley et al., 2015). These facts motivated the development and improvement of technological platforms for spatial analysis and their use in dissecting the TME.

IHC is a widely used technique for both basic research and cancer diagnosis. It is used to localize cells that express specific protein markers and study the spatial localization of cells in a tissue slide. In conventional IHC (i.e., chromogenic IHC), antibodies recognize specific antigens in the tissue and are conjugated to an enzyme to catalyze a color-producing reaction. A major problem with conventional IHC is that it only allows labeling of one marker for tissue sections. This severely limits understanding the cellular complexity of the TME. Several multiplex platforms have been introduced over the years to facilitate the analysis of cellular composition, functional states, and cell-cell interactions within the TME to address this problem. Multiplex IHC (mIHC) methods based on chromogens (Remark et al., 2016; Tsujikawa et al., 2017), fluorophores (Gerdes et al., 2013; Gorris et al., 2018; Viratham Pulsawatdi et al., 2020), metal-tagged antibodies (Angelo et al., 2014), and DNA barcodes (Goltsev et al., 2018; Manesse et al., 2020) (**Figure 2E**) increased the number of biomarkers to be used simultaneously by up to 50 in a single tissue section. The introduction of more effective mIHC technologies has been accompanied by the development of new software solutions capable of performing sophisticated analyses of digitalized images (e.g., segmentation and filtering) and support the work of pathologists in sample processing.

mIHC technologies have been used in different types of cancer to better understand the spatial architecture of the TME and how it can affect the response to therapy and clinical outcomes. For example, in breast cancer, mIHC was used to obtain detailed information on the spatial localization, cellular composition, and expression of regulatory proteins in the TME and recover clinically relevant characteristics (Keren et al., 2018; Jackson et al., 2020). In colorectal cancer, mIHC has been used to decipher the complex dynamic interplay between TME components (Schurch et al., 2020). In human pancreatic ductal adenocarcinoma (PDAC), mIHC was applied to assess the density and spatial distribution of myeloid and lymphoid cells in the TME and its correlation with the clinical outcome of patients (Liudahl et al., 2021).

Single-Cell Multimodal Omics

In cancer, single-cell omics aims to dissect all aspects of cellular machinery to understand its functional status and relationship with other TME cells. The use of single-cell omics technologies to study single-cell modalities, such as transcriptomics and proteomics, has increased our knowledge of cell biology in cancer without precedent. However, cells are dynamic entities whose states are characterized by a complex interplay of genomic, transcriptomic, epigenomic, and proteomic features that non-linearly contribute to the TME's heterogeneity. The new Frontier is the simultaneous measurement of multiple modalities of the same cells to gain a better understanding of cellular and molecular mechanisms in cancer. In recent years, this has motivated the introduction of numerous single-cell approaches capable of combining two or more modalities between genomics, transcriptomics, epigenomics, and proteomics (**Figure 1E**). Multimodal single-cell approaches have evolved from those using tubes or microwells of plates to measure single cells (i.e., low-throughput) to those that take advantage of droplet-based technologies or combinatorial DNA barcoding strategies (i.e., high-throughput), allowing increased scalability and reduced costs per run (Zhu et al., 2020). Almost all current single-cell multimodal omics technologies extract the transcriptome of each cell. Low-throughput multimodal omics with transcriptomics paired with genomics are gDNA-mRNA sequencing (DR-seq) (Dey et al., 2015) and genome and transcriptome sequencing (G&T-seq) (Macaulay et al., 2015), while transcriptomics and epigenomics are obtained with single-cell methylome and transcriptome sequencing (scM&T) (Angermueller et al., 2016), scMT-seq (Hu et al., 2016), single-cell nucleosome, methylation, and transcription sequencing (snNMT-seq) (Clark et al., 2018), single-cell nucleosome occupancy, methylome, and RNA expression sequencing (scNOMeRe-seq) (Wang et al., 2021), and single-cell chromatin accessibility and transcriptome sequencing (scCAT-seq) (Liu et al., 2019). An interesting multimodal omics technique integrating transcriptomics, genomics, and epigenomics is single-cell triple omics sequencing (scTrio-seq) (Hou et al., 2016).

In high-throughput multimodal omics, transcriptomics is paired with epigenomics in parallel analysis of individual cells for RNA expression and DNA accessibility by sequencing (Paired-seq) (Zhu et al., 2019) and single-nucleus chromatin accessibility and mRNA expression sequencing (SNARE-seq) (Chen S. et al., 2019). Furthermore, transcriptomics is paired with epitopes in RNA expression and protein sequencing (REAP-seq) (Peterson et al., 2017) and cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) (Stoeckius et al., 2017) (**Figure 2F**). CITE-seq protocol has been extended to obtain transcriptomics and proteomics of the same cells after CRISPR-Cas9 genetic perturbations in expanded CRISPR-compatible CITE-seq (ECCITE-seq) (Mimitou et al., 2019) and Perturb-CITE-sequencing (Perturb-CITE-seq) (Frangieh et al., 2021). These techniques perform genetic perturbations through single gRNAs (sgRNA) linking them to transcriptomic and proteomic profiles of the same cells to allow demultiplexing. ECCITE-seq combines the use of CRISPR libraries and cell hashtags (Stoeckius et al., 2018) to perform genetic perturbations and pool together different experimental samples. Perturb-CITE-seq uses a method based on CROP-seq (Datlinger et al., 2017) to express sgRNAs and to link them to transcripts and surface proteins of single cells. Other interesting multimodal omics techniques that do not include transcriptomics are, for example, those that provide different layers of epigenomics (Guo et al., 2017; Pott, 2017), epigenomics paired with genomics, and CRISPR-Cas9 genetic perturbations (Rubin et al., 2019; Tedesco et al., 2021).

Low-throughput multimodal omics were successfully adopted in cancer studies to better understand heterogeneity, complexity, and the evolution of cancer cells by integrating genomic, epigenomic, and transcriptomic features of the same cells (Macaulay et al., 2015; Hou et al., 2016; Bian et al., 2018; Zhu et al., 2021). However, the possibility of obtaining only tens or hundreds of cells provided by low-throughput multimodal omics and their costs have limited their application in TME studies.

Furthermore, given the importance of immunomodulatory proteins (e.g., PD-L1, CTLA-4) in response to cancer

immunotherapy, high-throughput multimodal technologies combining transcriptomics and proteomics have been applied in TME studies. In a recent report, CITE-seq successfully discovered new macrophage populations expressing PD-L1 and PD-L2 surface proteins linked to survival in breast cancer (Wu et al., 2021). In other studies, ECCITE-seq and Perturb-CITE-seq were used to define new clinically relevant resistance mechanisms to ICIs in human cancer cell lines and melanoma by exploiting CRISPR-Cas9 screens with multimodal single-cell readouts (Frangieh et al., 2021; Papalexi et al., 2021). Furthermore, the possibility of having both transcript and protein expression from the same cells also helped scientists increase the robustness of their scRNA-seq results. Indeed, they were able to validate the expression of known protein markers and identify novel proteins expressed by immune subsets in different human cancers (Leader et al., 2021; Pombo Antunes et al., 2021). The progress in single cell technologies will guarantee in the near future the possibility to explore at single cell level other omics such as glycomics, lipidomics, metabolomics and microbiomics able to highlight essential cell functions and biological proprieties of tissue components generating a more detailed map of the immune landscape of TME.

COMPUTATIONAL APPROACHES TO ANALYZE SINGLE-CELL DATA OF THE TME

With the widespread use of high-dimensional biological datasets, the scientific community designed specific computational techniques capable of extracting knowledge from complex multi-omic data. Compared to its bulk counterpart, single-cell data analysis is particularly challenging because of the high dimensionality given by the number of cells and markers and the presence of peculiar technical and biological factors that are important to keep in mind. For example, an underestimated problem in single-cell analysis is computational power. As the number of cells increases, all the data analysis steps become more computationally intensive, and some of these steps may require more scalable computational methods and architectures. Machine learning (ML) techniques, especially deep learning (DL), are an emerging class computational methods in single-cell data analysis for their capability to manage complex datasets and the possibility to be implemented into high-parallel architectures (e.g., GPU). Additionally, single-cell data is noisier and more subject to batch effects compared to the bulk counterpart due, for example, to more critical experimental procedures such as single cell isolation, the technical variability in the number of reads and cells sequenced in each sample, and the biological variability caused by heterogeneity in cell composition. TME studies particularly exemplify the latter aspect, in which the presence of cancer and immune and non-immune cells contributes to the complexity of the experiment. Potential confounding factors must be removed or included in statistical models during data analysis. Typical computational steps performed during single-cell data analysis include 1) preprocessing and harmonization steps in which multiple datasets or modalities are combined to perform an integrated analysis after the removal of outliers or low-quality cells, 2) applying dimensionality reducing techniques useful for visualization, 3) clustering, 4) cell annotation, and 5) cellular and molecular functional analysis.

Pre-Processing and Harmonization of Single-Cell Data

Pre-processing is essential before the downstream analysis of single-cell data. This phase encompasses many computational steps ranging from raw to processed data through various types and file formats. Sequencing- and mass spectrometry-based single-cell data involve several steps before quantifying the features-by-cell matrices. Sequencing-based methods have common procedures, including processing raw FASTQ files containing the reads and alignments to the reference genome or transcriptome. Next, for both sequencing-based and mass spectrometry-based approaches, feature detection and quantification (e.g., exons, peaks, and peptides) are common stages to obtain the final features for analysis. Prior to matrix processing and analysis, preliminary quality control is required to assess, for example, the quality of the reads, the percentage of valid barcodes, and reads mapped to the genome. The software suite of specific platforms (e.g., 10x Genomics Cell Ranger) often provides this information. Next, the matrices are reduced by removing low-quality and outlier cells. For example, cells with an unexpectedly high or low number of features detected or with poor quantification are removed from further analyses. In droplet-based single-cell technologies, a filtering step removes cells with a hybrid transcriptome (e.g., doublets), that is, two or more cells incapsulated in the same droplet, using specialized software (McGinnis et al., 2019; Wolock et al., 2019; DePasquale et al., 2020). A specific scRNA-seq pre-processing step involves removing cells with a high percentage of mitochondrial and ribosomal genes expressed because they are usually considered low quality. Thresholds on mitochondrial and ribosomal expression must be chosen carefully, especially when dissecting the immune complexity of the TME, to avoid removing certain cell subsets (Zilionis et al., 2019; Osorio and Cai, 2021; Subramanian et al., 2021).

In FC and CyTOF, pre-processing steps and downstream analyses are typically performed using flow cytometry standard (FCS) files. Typical FC pre-processing steps include data compensation and transformation (e.g., biexponential, generalized Box-Cox) to correct the channel spillover and the effects of outliers and distorted distributions, respectively. Like droplet-based single-cell technologies, an important FC pre-processing step involves removing doublets. This happens when the cytometer cannot discriminate between 2 cells because they pass too closely through the trigger laser. In FC, single cells are differentiated at the beginning of the gating strategy using the 2D plot with Forward Side Channel-Aria (FSC-A) and Forward Side Channel-Height (FSC-H). Cells that do not display a linear correlation of these two parameters are marked as doublets and are excluded from the analysis. In CyTOF, the discrimination of doublets is more complex

because the cell size parameters used in FC are not available. Here, DNA intercalators and event length are used to obtain single cells (Gadalla et al., 2019). In spatial proteomics, the pre-processing step involves some adjustment (e.g., illumination and contrast correction) to the images before cell identification through the segmentation process. After segmentation, the cells are classified and quantified for subsequent statistical analysis.

As part of the pre-processing phase, normalization is a fundamental procedure in single-cell data analysis to make all the cells comparable. In sequencing-based single-cell data, the variability of reads sequenced per cell and data sparsity have been carefully considered in computational pipelines to avoid technical effects confounding biological heterogeneity. In scRNA-seq, this has motivated the design of different normalization methods, for example, those based on the estimation of size factors to correct gene expression (Lun et al., 2016; Wolf et al., 2018; Stuart et al., 2019) and others based on linear regression (Bacher et al., 2017; Yip et al., 2017; Hafemeister and Satija, 2019). Text-mining techniques are used to normalize scATAC-seq datasets (Cusanovich et al., 2015; Cusanovich et al., 2018; Fang et al., 2021). Normalization is also an important step in single-cell proteomics data to remove the technical variability due, for example, to differences in instrument performance over acquisition time, particularly in mass cytometry (Rybakowska et al., 2020). Bead-based normalization was introduced to correct these technical artifacts (Finck et al., 2013). This method uses information obtained through standard calibration beads that track changes in the signal over the acquisition time to adjust the marker values. However, technical and biological differences may arise from the disparate technical and biological aspects of data generation in single-cell experiments. A key step in studying the TME through single-cell technologies is the study of changes in the cell subset composition of different individuals. Harmonization techniques have been proposed for single-cell data to facilitate the comparison of multiple samples. These techniques aim to overcome noise due to the variability among cells, individuals, species, and protocols trying to maintain the true biological signals. An important advantage of data harmonization is the possibility to aggregate multiple samples into a single dataset making faster identifying shared or sample-specific cell subsets. This approach makes the analysis simpler and less error prone because it avoids processing each dataset individually or merging them without considering potential biases. A drawback of harmonization is the possibility to lose true biological signals due to the "correction" procedure. In single-cell proteomics, methods have been introduced to correct unwanted variability in FC and mass cytometry, such as aligning the marker intensity distributions across samples or performing cell-type-specific normalization using shared controls across multiple batches (Hahne et al., 2009; Finak et al., 2014; Van Gassen et al., 2020). Several methods ranging from more conventional methods to highly scalable and fast ML approaches have been proposed to align and make cell subsets comparable to gene expression and chromatin accessibility of different datasets, considering diverse technical and biological variation sources (Luecken et al., 2022). The advent of single-cell multimodal omics has posed additional computational challenges in extrapolating useful information from the different layers measured for each cell. This has motivated the introduction of harmonization methods for performing a joint analysis, such as clustering, to exploit the power given by all available cell modalities (Argelaguet et al., 2020; Wang et al., 2020; Gayoso et al., 2021; Hao et al., 2021; Singh et al., 2021; Zuo and Chen, 2021). It is worth to remark that DL approaches (e.g., deep generative models) represent a significant part of harmonization methods for single-cell datasets. Their power is given by their capacity to learn complex mechanisms of biological systems from multiple biological datasets and modalities (Li Y. et al., 2018). This makes them important tools to manage the increasing complexity of single-cell data.

Dimensionality Reduction, Clustering, and Cell Annotation of Single-Cell Data

Single-cell datasets are intrinsically high dimensional. A common step in single-cell data analysis is projecting the data into a low-dimensional space using dimensionality reducing techniques to dissect the complexity and understand the TME's cellular composition, cell states, and trajectories. The dimensions calculated by these algorithms are useful for visualizing the cell subsets, usually using 2D scatter plots, and as inputs of other computational techniques. Methods such as principal component analysis (PCA) and singular value decomposition (SVD) are commonly used to decompose high-dimensional datasets into several important axes of variation. Owing to their linearity, these methods cannot accurately represent the complex structure of single-cell data. However, the dimensions extracted by these approaches are commonly injected into other dimensionality reducing methods, such as t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP), to improve their accuracy (Kobak and Berens, 2019; Baek and Lee, 2020; Kobak and Linderman, 2021). t-SNE and UMAP are currently the most commonly used techniques for visualizing single-cell datasets intuitively. With these methods, each cell is represented as a point in a 2D scatter plot, where the proximity to other cells indicates a similarity in expression profiles. These algorithms can position cells in a biologically meaningful way allowing the correct interpretation of the data. A vital property of these algorithms is to preserve the local and global structure of the data, namely, the distances between points (e.g., cells) within the same cluster and between different clusters. The local topology ensures that the cells in a cluster are homogeneous and represent, for example, the same immune cell subset. In contrast, the global structure of data can provide important biological insights into cell subset relationships, such as monocyte-to-macrophage or epithelial-to-malignant cell transitions, and it is the most debated feature of t-SNE and UMAP. Although previous reports have shown UMAP to be better than t-SNE in preserving the organization of cell clusters (Becht et al., 2018), recent studies suggest that more effective t-SNE parameterizations make it as good as UMAP for conserving global data geometry (Kobak and Berens, 2019; Kobak and Linderman, 2021). Considering the importance of studying the state transition and differentiation trajectories of single cells, other dimensionality reducing approaches have been proposed to better preserve the global features of the data and, consequently, provide biologically relevant information (Weinreb et al., 2018; Moon et al., 2019). After dimensionality reduction, clustering and

cell annotation are usually the next steps in single-cell data analysis. Clustering algorithms are mainly used to define homogenous cell subsets to be annotated using specific molecular measures, such as gene expression, proteins, or peak counts. In single-cell multimodal omics data, all the provided molecular measures can be combined to perform an integrated analysis, as the previous section introduced.

Among the large toolboxes of clustering techniques made available by computational sciences, several methods are more commonly used in the single-cell data analysis context. Graph-based clustering has become very popular in single-cell data analysis owing to its minimal required assumptions compared to other well-known techniques, such as k-means. These include methods based on clique detection (Xu and Su, 2015), spectral clustering (Ng et al., 2002), and community detection algorithms (Blondel et al., 2008; Waltman and Van Eck, 2013; Traag et al., 2019). The latter has been highly appreciated in recent years because they can scale the number of cells and be implemented in popular R and Python packages, such as Seurat (Stuart et al., 2019), Monocle (Qiu et al., 2017), and SCANPY (Wolf et al., 2018). In recent years, ML techniques based on self-organizing maps (SOMs) (Kohonen, 1990) have also been introduced for both mass cytometry (Van Gassen et al., 2015) and scRNA-seq data (Camp et al., 2017). These tools usually provide useful visualization features for an intuitive interpretation of expression similarity among cell clusters.

Cluster analysis is usually used for cell annotation following an iterative process in which each cluster is mapped to a biologically relevant cell type by observing the expression of multiple markers. Several clustering algorithms allow us to set a level of granularity (i.e., resolution) based on the level of detail at which the cell subsets are dissected. The use of clustering to annotate cells has several drawbacks. For example, the number of clusters may be overestimated, underestimated, or not reproducible. Furthermore, manual annotation may be error-prone or not correspond to similar annotated cells in the literature. Alternatively, several classification techniques have been introduced in recent years to perform automatic cell annotation (Zhao X. et al., 2020) based on well-annotated reference datasets containing, for example, fluorescence-activated cell sorting (FACS)-sorted cell populations. These methods typically use bulk or single-cell references from large general-purpose (Consortium, 2012; Mabbott et al., 2013; Martens and Stunnenberg, 2013), immune-specific (Heng et al., 2008), or tumor-specific atlases (Rozenblatt-Rosen et al., 2020) to infer the cell types present in dataset by correlating their expression profiles. Most of these tools make feature selections before performing cell classifications. However, the unbiased nature of these approaches makes them critical for improving the reproducibility and consistency among single-cell studies. With the spread of large-scale disease datasets, an important goal will be to integrate these datasets into harmonized and batch-corrected references to be queried efficiently. Scalable architectures based on ML will be valuable tools for integrating human references to study the TME (Lotfollahi et al., 2022).

Analysis of Cell Function and Differentiation of Single-Cell Data

Analyzing changes in the cellular composition of the TME is necessary for understanding the various mechanisms of cancer. However, a deep molecular characterization is crucial for identifying the key drivers of functional cellular changes. Differential expression (DE) analysis is the main statistical technique for detecting functional perturbations caused by changes in gene or protein expression, chromatin accessibility, and genomic aberrations. Although general-purpose statistical tests or DE methods for bulk datasets have been widely used for DE analysis of different single-cell omics datasets, several specialized techniques have been adopted over the years to deal with the heterogeneity and sparsity of scRNA-seq (Kharchenko et al., 2014; Finak et al., 2015). However, DE methods based on pseudo-bulk aggregation of biological replicates have recently gained attention for their capacity to extrapolate more robust results than general-purpose and specialized single-cell DE methods (Squair et al., 2021). After differential expression analysis, a typical task in the study of the TME is to extract a list of biological processes linked to the molecular changes induced by cancer or therapy in different cell subsets. Methods for gene set analysis that have been widely used in recent years for bulk data, such as over-representation, are commonly used in single-cell data. These approaches take into account a list of differentially expressed molecules and gene sets from Gene Ontology (GO) (Gene Ontology, 2015), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2016), Reactome (Fabregat et al., 2018), and the Molecular Signature Database (MSigDB) (Liberzon et al., 2011) to identify altered molecular pathways.

In addition to molecular changes concerning pathway regulation, characterizing molecular states between different cells is a common step in single-cell data analysis. While clustering and cell annotation techniques provide information about the cell types present in the experiment, they supply no information on the relationships between cell types and clusters. Recently, several methods have been proposed to reconstruct differentiation trajectory maps by ordering cells based on expression pattern similarity (Saelens et al., 2019) or transcriptional dynamics (La Manno et al., 2018; Bergen et al., 2020). These techniques have been successfully applied to study the developmental trajectories of different cell types and TMEs (Zhang et al., 2019; Chen et al., 2020; Qian et al., 2020; He et al., 2021; Liu et al., 2021). With the improvement in single-cell omics technology performance and their integration ability, we need to address the problem of cell number scalability and the generalizability of these methods to different omics (Stassen et al., 2021).

Pro- and anti-tumor mechanisms are governed by changes driven by complex molecular and cellular interactions within the TME (Anderson and Simon, 2020). Various methods have been introduced in single-cell data analyses to study the intra- and inter-cellular interactome through the inference of gene regulatory networks (GRNs) and ligand-receptor pairs to understand the biological processes underlying these mechanisms. Several methods for GRN inference have been proposed to decipher intracellular networks, including tools originally designed for bulk transcriptomics (Huynh-Thu and Sanguinetti, 2015; Moerman et al., 2019) and techniques specifically designed for single-cell transcriptomics that exploit additional information, such as pseudo-temporal ordering (Matsumoto et al., 2017; Specht and Li, 2017; Deshpande

et al., 2021) and information about transcription factors and their targets (Aibar et al., 2017). The first class of methods tries to learn the gene regulatory structure without prior information, and the second class uses pseudo-temporal information to better explain gene regulation during cell differentiation and development. A systematic evaluation of these techniques was recently published (Pratapa et al., 2020). In addition to the intracellular regulation state of a cell, a crucial aspect in understanding the TME involves exploring the cell-to-cell interactome induced by cancer cells (Whiteside, 2008). In spatial technologies, this task can be accomplished by observing cellular co-localization by inspecting appropriate cell markers in histological regions of interest (ROIs) of tissue sections. Without spatial information, cellular interactions must be inferred from the ligand and respective cognate receptor expression levels.

Several computational methods based on different mathematical models have been proposed to identify cell-to-cell interactions. Among others, methods based on the permutation of expression have been widely used (Armingol et al., 2021). These methods typically calculate the communication score of a list of ligand-receptor pairs obtained from curated databases and evaluate the significance of the interactions through cluster label permutation and statistical tests. Ligand-receptor interactions inferred from single-cell transcriptomics may provide interesting hypotheses that need to be further validated using other technologies.

CONCLUSION AND FUTURE PERSPECTIVES

The shift from bulk to single-cell sequencing has allowed us to move forward from a general molecular signature in which the contribution of each cell is averaged to the complete molecular fingerprint of each sequenced cell. This is particularly important when complex samples characterized by heterogeneous cell compositions, such as tumor tissues, are analyzed. Accordingly, the employment of single-cell technologies has radically improved the understanding of the cancer framework both quantitatively and qualitatively. Indeed, a single-cell platform can resolve the plasticity of tumor cells and decode tumor phenotypes (invasiveness, stemness, proliferation, and apoptosis), revealing the composition of the TME and the differentiation of immune and stromal cells towards anti- or pro-tumor phenotypes. In addition, single-cell sequencing can track the evolutionary trajectories of neoplastic clones in primary tumors, with results that challenge the original vision of gradual neoplastic evolution (Gao et al., 2016) and improve the understanding of the metastatic spreading process by profiling circulating tumor cells or metastatic lesions (Leung et al., 2017). Similarly, molecular tumor fingerprinting can predict the response to target therapy (Tirosh et al., 2016; Rambow et al., 2018). Although single-cell technologies were initially developed for research purposes and contributed significantly to dissecting cancer evolution mechanisms, clinical settings will soon use them. Investigation at the single-cell level can improve early tumor detection, prognostic biomarker identification, and patient risk stratification, thus supporting a more tumor-tailored therapy. Similarly, dissecting the complexity of the TME can help design the best immunotherapy approach, reverting local immune suppression or empowering the fitness and killing abilities of tumor-infiltrating effector cells. Finally, single-cell sequencing platforms can be employed as potent diagnostic tools for non-invasive monitoring of tumor evolution and patient relapse by profiling circulating tumor cells. This approach can detect, for instance, the appearance of clones resistant to targeted therapy, promptly driving the clinical decision towards an alternative therapeutic solution. However, some limitations need to be overcome to make this technology available for mainstream clinical purposes. First, its usage requires a more complex team, including surgeons, oncologists, pathologists, and researchers working in a fast and coordinated manner and the development of robust tissue processing protocols for primary tumors. Moreover, sample processing requires loss of tissue architecture, whereas spatial single-cell technologies combining molecular and histological information do not guarantee the same resolution of single-cell sequencing on suspension cells. Secondly, scRNA-seq data depend on their intrinsic noisy since eukaryotic transcription does not occur at a persistent basal rate but it takes place in pulses (Chubb et al., 2006). Therefore, a failure to uncover a transcript of a specific gene in a cell at a single time point is an ambiguous result since it can be considered as a result of either permanent gene inactivation or timely limitation of gene transcription detection where the gene is active but the transcript in the time window of the sampling is not present. To avoid possible serious faults, the interpretation of scRNA-seq results should directed on pathways analysis and gene-set enrichment rather than single gene expression. Finally, batch effects can take places when aggregating multiple samples. Several methods of correcting batch effects have been optimized but their use must be balanced against the risk of eclipsing true biological differences. Nonetheless, as recently exemplified by the introduction of bulk NGS, single-cell technology use will soon be extended to patients to support cancer diagnosis and treatment in the near future.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by the PRIN programs of the Italian Ministry of Education, University and Research (MIUR, CUP: B38D19000140006, CUP: B39J22001200001) and Fondazione Associazione Italiana per la Ricerca sul Cancro (AIRC, Project: 21509).

ACKNOWLEDGMENTS

Figures 1A–D were adapted from the templates "Cancer Cell Mutations Affect Tumor Microenvironment", "Epigenetic Deregulation in Cancer", "Cancer Evolution with Treatment", "ATAC Sequencing", "Immune Mechanisms Affected by COVID-19″, "TI-1 Antigens Stimulate B cells in Conjuntion

with TLR Signal", "Icon Pack–Immunology" by BioRender.com (2022), "CITE-seq (Cellular Indexing of Transcriptomes & Epitopes by Sequencing) Workflow". Retrieved from "https://app.biorender.com/biorender-templates"

REFERENCES

Abel, E. J., Bauman, T. M., Weiker, M., Shi, F., Downs, T. M., Jarrard, D. F., et al. (2014). Analysis and Validation of Tissue Biomarkers for Renal Cell Carcinoma Using Automated High-Throughput Evaluation of Protein Expression. *Hum. Pathol.* 45 (5), 1092–1099. doi:10.1016/j.humpath.2014.01.008

Adey, A., Morrison, H. G., AsanXun, A., Xun, X., Kitzman, J. O., Turner, E. H., et al. (2010). Rapid, Low-Input, Low-Bias Construction of Shotgun Fragment Libraries by High-Density *In Vitro* Transposition. *Genome Biol.* 11 (12), R119. doi:10.1186/gb-2010-11-12-r119

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). SCENIC: Single-Cell Regulatory Network Inference and Clustering. *Nat. Methods* 14 (11), 1083–1086. doi:10.1038/nmeth.4463

Anderson, N. M., and Simon, M. C. (2020). The Tumor Microenvironment. *Curr. Biol.* 30 (16), R921–R925. doi:10.1016/j.cub.2020.06.081

Angelo, M., Bendall, S. C., Finck, R., Hale, M. B., Hitzman, C., Borowsky, A. D., et al. (2014). Multiplexed Ion Beam Imaging of Human Breast Tumors. *Nat. Med.* 20 (4), 436–442. doi:10.1038/nm.3488

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., et al. (2016). Parallel Single-Cell Sequencing Links Transcriptional and Epigenetic Heterogeneity. *Nat. Methods* 13 (3), 229–232. doi:10.1038/nmeth.3728

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., et al. (2020). MOFA+: a Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data. *Genome Biol.* 21 (1), 111. doi:10.1186/s13059-020-02015-1

Armingol, E., Officer, A., Harismendy, O., and Lewis, N. E. (2021). Deciphering Cell-Cell Interactions and Communication from Gene Expression. *Nat. Rev. Genet.* 22 (2), 71–88. doi:10.1038/s41576-020-00292-x

Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., et al. (2017). SCnorm: Robust Normalization of Single-Cell RNA-Seq Data. *Nat. Methods* 14 (6), 584–586. doi:10.1038/nmeth.4263

Baek, S., and Lee, I. (2020). Single-cell ATAC Sequencing Analysis: From Data Preprocessing to Hypothesis Generation. *Comput. Struct. Biotechnol. J.* 18, 1429–1439. doi:10.1016/j.csbj.2020.06.012

Balachandran, V. P., Łuksza, M., Łuksza, M., Zhao, J. N., Makarov, V., Moral, J. A., et al. (2017). Identification of Unique Neoantigen Qualities in Long-Term Survivors of Pancreatic Cancer. *Nature* 551 (7681), 512–516. doi:10.1038/nature24462

Bandura, D. R., Baranov, V. I., Ornatsky, O. I., Antonov, A., Kinach, R., Lou, X., et al. (2009). Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-Of-Flight Mass Spectrometry. *Anal. Chem.* 81 (16), 6813–6822. doi:10.1021/ac901049w

Baslan, T., Kendall, J., Volyanskyy, K., McNamara, K., Cox, H., D'Italia, S., et al. (2020). Novel Insights into Breast Cancer Copy Number Genetic Heterogeneity Revealed by Single-Cell Genome Sequencing. *Elife* 9. doi:10.7554/eLife.51480

Baylin, S. B., and Jones, P. A. (2011). A Decade of Exploring the Cancer Epigenome - Biological and Translational Implications. *Nat. Rev. Cancer* 11 (10), 726–734. doi:10.1038/nrc3130

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., et al. (2018). Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat. Biotechnol.* 37, 38–44. doi:10.1038/nbt.4314

Benci, J. L., Xu, B., Qiu, Y., Wu, T. J., Dada, H., Twyman-Saint Victor, C., et al. (2016). Tumor Interferon Signaling Regulates a Multigenic Resistance Program to Immune Checkpoint Blockade. *Cell* 167 (6), 1540–1554. doi:10.1016/j.cell.2016.11.022

Bendall, S. C., Nolan, G. P., Roederer, M., and Chattopadhyay, P. K. (2012). A Deep Profiler's Guide to Cytometry. *Trends Immunol.* 33 (7), 323–332. doi:10.1016/j.it.2012.02.010

Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E.-a. D., Krutzik, P. O., Finck, R., et al. (2011). Single-cell Mass Cytometry of Differential Immune and Drug Responses across a Human Hematopoietic Continuum. *Science* 332 (6030), 687–696. doi:10.1126/science.1198704

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020). Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling. *Nat. Biotechnol.* 38 (12), 1408–1414. doi:10.1038/s41587-020-0591-3

Betancur, P. A., Abraham, B. J., Yiu, Y. Y., Willingham, S. B., Khameneh, F., Zarnegar, M., et al. (2017). A CD47-Associated Super-enhancer Links Pro-inflammatory Signalling to CD47 Upregulation in Breast Cancer. *Nat. Commun.* 8, 14802. doi:10.1038/ncomms14802

Bian, S., Hou, Y., Zhou, X., Li, X., Yong, J., Wang, Y., et al. (2018). Single-cell Multiomics Sequencing and Analyses of Human Colorectal Cancer. *Science* 362 (6418), 1060–1063. doi:10.1126/science.aao3791

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* 39 (4), 782–795. doi:10.1016/j.immuni.2013.10.003

Binnewies, M., Roberts, E. W., Kersten, K., Chan, V., Fearon, D. F., Merad, M., et al. (2018). Understanding the Tumor Immune Microenvironment (TIME) for Effective Therapy. *Nat. Med.* 24 (5), 541–550. doi:10.1038/s41591-018-0014-x

Blando, J., Sharma, A., Higa, M. G., Zhao, H., Vence, L., Yadav, S. S., et al. (2019). Comparison of Immune Infiltrates in Melanoma and Pancreatic Cancer Highlights VISTA as a Potential Target in Pancreatic Cancer. *Proc. Natl. Acad. Sci. U.S.A.* 116 (5), 1692–1697. doi:10.1073/pnas.1811067116

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *J. Stat. Mech.* 2008 (10), P10008–P15468. doi:10.1088/1742-5468/2008/10/P10008

Bonilla, D. L., Reinin, G., and Chua, E. (2020). Full Spectrum Flow Cytometry as a Powerful Technology for Cancer Immunotherapy Research. *Front. Mol. Biosci.* 7, 612801. doi:10.3389/fmolb.2020.612801

Bronte, V., Cingarlini, S., Marigo, I., De Santo, C., Gallina, G., Dolcetti, L., et al. (2006). Leukocyte Infiltration in Cancer Creates an Unfavorable Environment for Antitumor Immune Responses: a Novel Target for Therapeutic Intervention. *Immunol. Investig.* 35 (3-4), 327–357. doi:10.1080/08820130600754994

Brown, R., Curry, E., Magnani, L., Wilhelm-Benartzi, C. S., and Borley, J. (2014). Poised Epigenetic States and Acquired Drug Resistance in Cancer. *Nat. Rev. Cancer* 14 (11), 747–753. doi:10.1038/nrc3819

Budnik, B., Levy, E., Harmange, G., and Slavov, N. (2018). SCoPE-MS: Mass Spectrometry of Single Mammalian Cells Quantifies Proteome Heterogeneity during Cell Differentiation. *Genome Biol.* 19 (1), 161. doi:10.1186/s13059-018-1547-5

Burr, M. L., Sparbier, C. E., Chan, K. L., Chan, Y.-C., Kersbergen, A., Lam, E. Y. N., et al. (2019). An Evolutionarily Conserved Function of Polycomb Silences the MHC Class I Antigen Presentation Pathway and Enables Immune Evasion in Cancer. *Cancer Cell* 36 (4), 385–401. doi:10.1016/j.ccell.2019.08.008

Calcinotto, A., Spataro, C., Zagato, E., Di Mitri, D., Gil, V., Crespo, M., et al. (2018). IL-23 Secreted by Myeloid Cells Drives Castration-Resistant Prostate Cancer. *Nature* 559 (7714), 363–369. doi:10.1038/s41586-018-0266-0

Calvanese, V., Capellera-Garcia, S., Ma, F., Fares, I., Liebscher, S., Ng, E. S., et al. (2022). Mapping Human Haematopoietic Stem Cells from Haemogenic Endothelium to Birth. *Nature* 604, 534–540. doi:10.1038/s41586-022-04571-x

Camp, J. G., Sekine, K., Gerber, T., Loeffler-Wirth, H., Binder, H., Gac, M., et al. (2017). Multilineage Communication Regulates Human Liver Bud Development from Pluripotency. *Nature* 546 (7659), 533–538. doi:10.1038/nature22796

Cancer Genome Atlas, N. (2012a). Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature* 487 (7407), 330–337. doi:10.1038/nature11252

Cancer Genome Atlas, N. (2012b). Comprehensive Molecular Portraits of Human Breast Tumours. *Nature* 490 (7418), 61–70. doi:10.1038/nature11412

Cancer Genome Atlas Research, N. (2012). Comprehensive Genomic Characterization of Squamous Cell Lung Cancers. *Nature* 489 (7417), 519–525. doi:10.1038/nature11404

Cancer Genome Atlas Research, N. (2014a). Comprehensive Molecular Characterization of Gastric Adenocarcinoma. *Nature* 513 (7517), 202–209. doi:10.1038/nature13480

Cancer Genome Atlas Research, N. (2014b). Comprehensive Molecular Profiling of Lung Adenocarcinoma. *Nature* 511 (7511), 543–550. doi:10.1038/nature13385

Cancer Genome Atlas Research, N. (2015). The Molecular Taxonomy of Primary Prostate Cancer. *Cell* 163 (4), 1011–1025. doi:10.1016/j.cell.2015.10.025

Cancer Genome Atlas Research, Network (2017). Analytical BiologicalIntegrated Genomic and Molecular Characterization of Cervical Cancer. *Nature* 543 (7645), 378–384. doi:10.1038/nature21386

Chen, C., Xing, D., Tan, L., Li, H., Zhou, G., Huang, L., et al. (2017). Single-cell Whole-Genome Analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* 356 (6334), 189–194. doi:10.1126/science.aak9787

Chen, G., Ning, B., and Shi, T. (2019a). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.* 10, 317. doi:10.3389/fgene.2019.00317

Chen, S., Lake, B. B., and Zhang, K. (2019b). High-throughput Sequencing of the Transcriptome and Chromatin Accessibility in the Same Cell. *Nat. Biotechnol.* 37 (12), 1452–1457. doi:10.1038/s41587-019-0290-0

Chen, Z., Zhou, L., Liu, L., Hou, Y., Xiong, M., Yang, Y., et al. (2020). Single-cell RNA Sequencing Highlights the Role of Inflammatory Cancer-Associated Fibroblasts in Bladder Urothelial Carcinoma. *Nat. Commun.* 11 (1), 5077. doi:10.1038/s41467-020-18916-5

Cheng, S., Li, Z., Gao, R., Xing, B., Gao, Y., Yang, Y., et al. (2021). A Pan-Cancer Single-Cell Transcriptional Atlas of Tumor Infiltrating Myeloid Cells. *Cell* 184 (3), 792–809. doi:10.1016/j.cell.2021.01.010

Chubb, J. R., Trcek, T., Shenoy, S. M., and Singer, R. H. (2006). Transcriptional Pulsing of a Developmental Gene. *Curr. Biol.* 16 (10), 1018–1025. doi:10.1016/j.cub.2006.03.092

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., et al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* 163 (2), 506–519. doi:10.1016/j.cell.2015.09.033

Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., et al. (2018). scNMT-seq Enables Joint Profiling of Chromatin Accessibility DNA Methylation and Transcription in Single Cells. *Nat. Commun.* 9 (1), 781. doi:10.1038/s41467-018-03149-4

Consortium, E. P. (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* 489 (7414), 57–74. doi:10.1038/nature11247

Consortium, I. T. P.-C. A. o. W. G. (2020). Pan-cancer Analysis of Whole Genomes. *Nature* 578 (7793), 82–93. doi:10.1038/s41586-020-1969-6

Cui, C., Shu, W., and Li, P. (2016). Fluorescence *In Situ* Hybridization: Cell-Based Genetic Diagnostic and Research Applications. *Front. Cell Dev. Biol.* 4, 89. doi:10.3389/fcell.2016.00089

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., et al. (2015). Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing. *Science* 348 (6237), 910–914. doi:10.1126/science.aab1601

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., et al. (2018). A Single-Cell Atlas of *In Vivo* Mammalian Chromatin Accessibility. *Cell* 174 (5), 1309–1324. e1318. doi:10.1016/j.cell.2018.06.052

Dangaj, D., Bruand, M., Grimm, A. J., Ronet, C., Barras, D., Duttagupta, P. A., et al. (2019). Cooperation between Constitutive and Inducible Chemokines Enables T Cell Engraftment and Immune Attack in Solid Tumors. *Cancer Cell* 35 (6), 885–900. doi:10.1016/j.ccell.2019.05.004

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., et al. (2017). Pooled CRISPR Screening with Single-Cell Transcriptome Readout. *Nat. Methods* 14 (3), 297–301. doi:10.1038/nmeth.4177

Davidson, S., Efremova, M., Riedel, A., Mahata, B., Pramanik, J., Huuhtanen, J., et al. (2020). Single-Cell RNA Sequencing Reveals a Dynamic Stromal Niche that Supports Tumor Growth. *Cell Rep.* 31 (7), 107628. doi:10.1016/j.celrep.2020.107628

De Sanctis, F., Bronte, V., and Ugel, S. (2016). Tumor-Induced Myeloid-Derived Suppressor Cells. *Microbiol. Spectr.* 4 (3). doi:10.1128/microbiolspec.MCHD-0016-2015

De Sanctis, F., Lamolinara, A., Boschi, F., Musiu, C., Caligola, S., Trovato, R., et al. (2022). Interrupting the Nitrosative Stress Fuels Tumor-specific Cytotoxic T Lymphocytes in Pancreatic Cancer. *J. Immunother. Cancer* 10 (1), e003549. doi:10.1136/jitc-2021-003549

Dean, F. B., Nelson, J. R., Giesler, T. L., and Lasken, R. S. (2001). Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. *Genome Res.* 11 (6), 1095–1099. doi:10.1101/gr.180501

DePasquale, E. A. K., Schnell, D., Chetal, K., and Salomonis, N. (2020). Protocol for Identification and Removal of Doublets with DoubletDecon. *Star. Protoc.* 1 (2), 100085. doi:10.1016/j.xpro.2020.100085

Deshpande, A., Chu, L.-F., Stewart, R., and Gitter, A. (2021). Network Inference with Granger Causality Ensembles on Single-Cell Transcriptomic Data. *Cell Rep.* 38, 534834. doi:10.1016/j.celrep.2022.110333

Dey, S. S., Kester, L., Spanjaard, B., Bienko, M., and van Oudenaarden, A. (2015). Integrated Genome and Transcriptome Sequencing of the Same Cell. *Nat. Biotechnol.* 33 (3), 285–289. doi:10.1038/nbt.3129

Di Pilato, M., Kfuri-Rubens, R., Pruessmann, J. N., Ozga, A. J., Messemaker, M., Cadilha, B. L., et al. (2021). CXCR6 Positions Cytotoxic T Cells to Receive Critical Survival Signals in the Tumor Microenvironment. *Cell* 184 (17), 4512–4530. e4522. doi:10.1016/j.cell.2021.07.015

Duan, Q., Zhang, H., Zheng, J., and Zhang, L. (2020). Turning Cold into Hot: Firing up the Tumor Microenvironment. *Trends Cancer* 6 (7), 605–618. doi:10.1016/j.trecan.2020.02.022

Elyada, E., Bolisetty, M., Laise, P., Flynn, W. F., Courtois, E. T., Burkhart, R. A., et al. (2019). Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. *Cancer Discov.* 9 (8), 1102–1123. doi:10.1158/2159-8290.CD-19-0094

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 46 (D1), D649–D655. doi:10.1093/nar/gkx1132

Facciponte, J. G., Ugel, S., De Sanctis, F., Li, C., Wang, L., Nair, G., et al. (2014). Tumor Endothelial Marker 1-specific DNA Vaccination Targets Tumor Vasculature. *J. Clin. Invest.* 124 (4), 1497–1511. doi:10.1172/JCI67382

Fan, X., Yang, C., Li, W., Bai, X., Zhou, X., Xie, H., et al. (2021). SMOOTH-seq: Single-Cell Genome Sequencing of Human Cells on a Third-Generation Sequencing Platform. *Genome Biol.* 22 (1), 195. doi:10.1186/s13059-021-02406-y

Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., et al. (2021). Comprehensive Analysis of Single Cell ATAC-Seq Data with SnapATAC. *Nat. Commun.* 12 (1), 1337. doi:10.1038/s41467-021-21583-9

Feinberg, A. P., Koldobskiy, M. A., and Göndör, A. (2016). Epigenetic Modulators, Modifiers and Mediators in Cancer Aetiology and Progression. *Nat. Rev. Genet.* 17 (5), 284–299. doi:10.1038/nrg.2016.13

Finak, G., Frelinger, J., Jiang, W., Newell, E. W., Ramey, J., Davis, M. M., et al. (2014). OpenCyto: an Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-To-End Flow Cytometry Data Analysis. *PLoS Comput. Biol.* 10 (8), e1003806. doi:10.1371/journal.pcbi.1003806

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: a Flexible Statistical Framework for Assessing Transcriptional Changes and Characterizing Heterogeneity in Single-Cell RNA Sequencing Data. *Genome Biol.* 16, 278. doi:10.1186/s13059-015-0844-5

Finck, R., Simonds, E. F., Jager, A., Krishnaswamy, S., Sachs, K., Fantl, W., et al. (2013). Normalization of Mass Cytometry Data with Bead Standards. *Cytometry* 83A (5), 483–494. doi:10.1002/cyto.a.22271

Finotello, F., and Eduati, F. (2018). Multi-Omics Profiling of the Tumor Microenvironment: Paving the Way to Precision Immuno-Oncology. *Front. Oncol.* 8, 430. doi:10.3389/fonc.2018.00430

Fiore, A., Ugel, S., De Sanctis, F., Sandri, S., Fracasso, G., Trovato, R., et al. (2018). Induction of Immunosuppressive Functions and NF-Kb by FLIP in Monocytes. *Nat. Commun.* 9 (1), 5193. doi:10.1038/s41467-018-07654-4

Frangieh, C. J., Melms, J. C., Thakore, P. I., Geiger-Schuller, K. R., Ho, P., Luoma, A. M., et al. (2021). Multimodal Pooled Perturb-CITE-Seq Screens in Patient Models Define Mechanisms of Cancer Immune Evasion. *Nat. Genet.* 53 (3), 332–341. doi:10.1038/s41588-021-00779-1

Fridman, W. H., Zitvogel, L., Sautès–Fridman, C., and Kroemer, G. (2017). The Immune Contexture in Cancer Prognosis and Treatment. *Nat. Rev. Clin. Oncol.* 14 (12), 717–734. doi:10.1038/nrclinonc.2017.101

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A Census of Human Cancer Genes. *Nat. Rev. Cancer* 4 (3), 177–183. doi:10.1038/nrc1299

Gadalla, R., Noamani, B., MacLeod, B. L., Dickson, R. J., Guo, M., Xu, W., et al. (2019). Validation of CyTOF against Flow Cytometry for Immunological Studies and Monitoring of Human Cancer Clinical Trials. *Front. Oncol.* 9, 415. doi:10.3389/fonc.2019.00415

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., et al. (2006). Type, Density, and Location of Immune Cells within Human Colorectal Tumors Predict Clinical Outcome. *Science* 313 (5795), 1960–1964. doi:10.1126/science.1129139

Gao, R., Davis, A., McDonald, T. O., Sei, E., Shi, X., Wang, Y., et al. (2016). Punctuated Copy Number Evolution and Clonal Stasis in Triple-Negative Breast Cancer. *Nat. Genet.* 48 (10), 1119–1130. doi:10.1038/ng.3641

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., et al. (2021). Joint Probabilistic Modeling of Single-Cell Multi-Omic Data with totalVI. *Nat. Methods* 18 (3), 272–282. doi:10.1038/s41592-020-01050-x

Gene Ontology, C. (2015). Gene Ontology Consortium: Going Forward. *Nucleic Acids Res.* 43, D1049–D1056. doi:10.1093/nar/gku1179

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., et al. (2015). The Prognostic Landscape of Genes and Infiltrating Immune Cells across Human Cancers. *Nat. Med.* 21 (8), 938–945. doi:10.1038/nm.3909

Gerdes, M. J., Sevinsky, C. J., Sood, A., Adak, S., Bello, M. O., Bordwell, A., et al. (2013). Highly Multiplexed Single-Cell Analysis of Formalin-Fixed, Paraffin-Embedded Cancer Tissue. *Proc. Natl. Acad. Sci. U.S.A.* 110 (29), 11982–11987. doi:10.1073/pnas.1300136110

Giladi, A., and Amit, I. (2017). Immunology, One Cell at a Time. *Nature* 547 (7661), 27–29. doi:10.1038/547027a

Gohil, S. H., Iorgulescu, J. B., Braun, D. A., Keskin, D. B., and Livak, K. J. (2021). Applying High-Dimensional Single-Cell Technologies to the Analysis of Cancer Immunotherapy. *Nat. Rev. Clin. Oncol.* 18 (4), 244–256. doi:10.1038/s41571-020-00449-x

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., et al. (2018). Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* 174 (4), 968–981. e915. doi:10.1016/j.cell.2018.07.010

Gorris, M. A. J., Halilovic, A., Rabold, K., van Duffelen, A., Wickramasinghe, I. N., Verweij, D., et al. (2018). Eight-Color Multiplex Immunohistochemistry for Simultaneous Detection of Multiple Immune Checkpoint Molecules within the Tumor Microenvironment. *J. I.* 200 (1), 347–354. doi:10.4049/jimmunol.1701262

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., et al. (2019). High-throughput Single-Cell ChIP-Seq Identifies Heterogeneity of Chromatin States in Breast Cancer. *Nat. Genet.* 51 (6), 1060–1066. doi:10.1038/s41588-019-0424-9

Guo, F., Li, L., Li, J., Wu, X., Hu, B., Zhu, P., et al. (2017). Single-cell Multi-Omics Sequencing of Mouse Early Embryos and Embryonic Stem Cells. *Cell Res.* 27 (8), 967–988. doi:10.1038/cr.2017.82

Guo, X., Zhang, Y., Zheng, L., Zheng, C., Song, J., Zhang, Q., et al. (2018). Global Characterization of T Cells in Non-small-cell Lung Cancer by Single-Cell Sequencing. *Nat. Med.* 24 (7), 978–985. doi:10.1038/s41591-018-0045-3

Hafemeister, C., and Satija, R. (2019). Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression. *Genome Biol.* 20 (1), 296. doi:10.1186/s13059-019-1874-1

Hahne, F., LeMeur, N., Brinkman, R. R., Ellis, B., Haaland, P., Sarkar, D., et al. (2009). flowCore: a Bioconductor Package for High Throughput Flow Cytometry. *BMC Bioinforma.* 10, 106. doi:10.1186/1471-2105-10-106

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: the Next Generation. *Cell* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., et al. (2021). Integrated Analysis of Multimodal Single-Cell Data. *Cell* 184 (13), 3573–3587. doi:10.1016/j.cell.2021.04.048

Hartmann, F. J., Babdor, J., Gherardini, P. F., Amir, E.-A. D., Jones, K., Sahaf, B., et al. (2019). Comprehensive Immune Monitoring of Clinical Trials to Advance Human Immunotherapy. *Cell Rep.* 28 (3), 819–831. e814. doi:10.1016/j.celrep.2019.06.049

Hayashi, T., Ozaki, H., Sasagawa, Y., Umeda, M., Danno, H., and Nikaido, I. (2018). Single-cell Full-Length Total RNA Sequencing Uncovers Dynamics of Recursive Splicing and Enhancer RNAs. *Nat. Commun.* 9 (1), 619. doi:10.1038/s41467-018-02866-0

He, D., Wang, D., Lu, P., Yang, N., Xue, Z., Zhu, X., et al. (2021). Single-cell RNA Sequencing Reveals Heterogeneous Tumor and Immune Cell Populations in Early-Stage Lung Adenocarcinomas Harboring EGFR Mutations. *Oncogene* 40 (2), 355–368. doi:10.1038/s41388-020-01528-0

Heng, T. S. P., Painter, M. W., Painter, M. W., Elpek, K., Lukacs-Kornek, V., Mauermann, N., et al. (2008). The Immunological Genome Project: Networks of Gene Expression in Immune Cells. *Nat. Immunol.* 9 (10), 1091–1094. doi:10.1038/ni1008-1091

Hofer, F., Di Sario, G., Musiu, C., Sartoris, S., De Sanctis, F., and Ugel, S. (2021). A Complex Metabolic Network Confers Immunosuppressive Functions to Myeloid-Derived Suppressor Cells (MDSCs) within the Tumour Microenvironment. *Cells* 10 (10), 2700. doi:10.3390/cells10102700

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., et al. (2016). Single-cell Triple Omics Sequencing Reveals Genetic, Epigenetic, and Transcriptomic Heterogeneity in Hepatocellular Carcinomas. *Cell Res.* 26 (3), 304–319. doi:10.1038/cr.2016.23

Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., et al. (2016). Simultaneous Profiling of Transcriptome and DNA Methylome from a Single Cell. *Genome Biol.* 17, 88. doi:10.1186/s13059-016-0950-z

Huynh-Thu, V. A., and Sanguinetti, G. (2015). Combining Tree-Based and Dynamical Systems for the Inference of Gene Regulatory Networks. *Bioinformatics* 31 (10), 1614–1622. doi:10.1093/bioinformatics/btu863

Jackson, H. W., Fischer, J. R., Zanotelli, V. R. T., Ali, H. R., Mechera, R., Soysal, S. D., et al. (2020). The Single-Cell Pathology Landscape of Breast Cancer. *Nature* 578 (7796), 615–620. doi:10.1038/s41586-019-1876-x

Jerby-Arnon, L., Shah, P., Cuoco, M. S., Rodman, C., Su, M.-J., Melms, J. C., et al. (2018). A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell* 175 (4), 984–997. doi:10.1016/j.cell.2018.09.006

Jones, P. A., Issa, J.-P. J., and Baylin, S. (2016). Targeting the Cancer Epigenome for Therapy. *Nat. Rev. Genet.* 17 (10), 630–641. doi:10.1038/nrg.2016.93

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res.* 44 (D1), D457–D462. doi:10.1093/nar/gkv1070

Kashima, Y., Togashi, Y., Fukuoka, S., Kamada, T., Irie, T., Suzuki, A., et al. (2021). Potentiality of Multiple Modalities for Single-Cell Analyses to Evaluate the Tumor Microenvironment in Clinical Specimens. *Sci. Rep.* 11 (1), 341. doi:10.1038/s41598-020-79385-w

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., et al. (2019). CUT&Tag for Efficient Epigenomic Profiling of Small Samples and Single Cells. *Nat. Commun.* 10 (1), 1930. doi:10.1038/s41467-019-09982-5

Keenan, T. E., Burke, K. P., and Van Allen, E. M. (2019). Genomic Correlates of Response to Immune Checkpoint Blockade. *Nat. Med.* 25 (3), 389–402. doi:10.1038/s41591-019-0382-x

Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., et al. (2018). A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell* 174 (6), 1373–1387. doi:10.1016/j.cell.2018.08.039

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian Approach to Single-Cell Differential Expression Analysis. *Nat. Methods* 11 (7), 740–742. doi:10.1038/nmeth.2967

Kieffer, Y., Hocine, H. R., Gentric, G., Pelon, F., Bernard, C., Bourachot, B., et al. (2020). Single-Cell Analysis Reveals Fibroblast Clusters Linked to Immunotherapy Resistance in Cancer. *Cancer Discov.* 10 (9), 1330–1351. doi:10.1158/2159-8290.CD-19-1384

Kim, N., Kim, H. K., Lee, K., Hong, Y., Cho, J. H., Choi, J. W., et al. (2020). Single-cell RNA Sequencing Demonstrates the Molecular and Cellular Reprogramming of Metastatic Lung Adenocarcinoma. *Nat. Commun.* 11 (1), 2285. doi:10.1038/s41467-020-16164-1

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* 161 (5), 1187–1201. doi:10.1016/j.cell.2015.04.044

Kobak, D., and Berens, P. (2019). The Art of Using T-SNE for Single-Cell Transcriptomics. *Nat. Commun.* 10 (1), 5416. doi:10.1038/s41467-019-13056-x

Kobak, D., and Linderman, G. C. (2021). Initialization Is Critical for Preserving Global Data Structure in Both T-SNE and UMAP. *Nat. Biotechnol.* 39 (2), 156–157. doi:10.1038/s41587-020-00809-z

Kohonen, T. (1990). The Self-Organizing Map. *Proc. IEEE* 78 (9), 1464–1480. doi:10.1109/5.58325

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., et al. (2018). RNA Velocity of Single Cells. *Nature* 560 (7719), 494–498. doi:10.1038/s41586-018-0414-6

LaFave, L. M., Kartha, V. K., Ma, S., Meli, K., Del Priore, I., Lareau, C., et al. (2020). Epigenomic State Transitions Characterize Tumor Progression in Mouse Lung Adenocarcinoma. *Cancer Cell* 38 (2), 212–228. e213. doi:10.1016/j.ccell.2020.06.006

Laks, E., McPherson, A., Zahn, H., Lai, D., Steif, A., Brimhall, J., et al. (2019). Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell* 179 (5), 1207–1221. e1222. doi:10.1016/j.cell.2019.10.026

Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., et al. (2018). Phenotype Molding of Stromal Cells in the Lung Tumor Microenvironment. *Nat. Med.* 24 (8), 1277–1289. doi:10.1038/s41591-018-0096-5

Lan, F., Demaree, B., Ahmed, N., and Abate, A. R. (2017). Single-cell Genome Sequencing at Ultra-high-throughput with Microfluidic Droplet Barcoding. *Nat. Biotechnol.* 35 (7), 640–646. doi:10.1038/nbt.3880

Leader, A. M., Grout, J. A., Maier, B. B., Nabet, B. Y., Park, M. D., Tabachnikova, A., et al. (2021). Single-cell Analysis of Human Non-small Cell Lung Cancer Lesions Refines Tumor Classification and Patient Stratification. *Cancer Cell* 39 (12), 1594–1609. e1512. doi:10.1016/j.ccell.2021.10.009

Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1998). Genetic Instabilities in Human Cancers. *Nature* 396 (6712), 643–649. doi:10.1038/25292

Leung, M. L., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., et al. (2017). Single-cell DNA Sequencing Reveals a Late-Dissemination Model in Metastatic Colorectal Cancer. *Genome Res.* 27 (8), 1287–1299. doi:10.1101/gr.209973.116

Li, J., Byrne, K. T., Yan, F., Yamazoe, T., Chen, Z., Baslan, T., et al. (2018a). Tumor Cell-Intrinsic Factors Underlie Heterogeneity of Immune Cell Infiltration and Response to Immunotherapy. *Immunity* 49 (1), 178–193. e177. doi:10.1016/j.immuni.2018.06.006

Li, Y., Wu, F.-X., and Ngom, A. (2018b). A Review on Machine Learning Principles for Multi-View Biological Data Integration. *Brief. Bioinform* 19 (2), bbw113–340. doi:10.1093/bib/bbw113

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics* 27 (12), 1739–1740. doi:10.1093/bioinformatics/btr260

Liu, B., Hu, X., Feng, K., Gao, R., Xue, Z., Zhang, S., et al. (2021). Temporal Single-Cell Tracing Reveals Clonal Revival and Expansion of Precursor Exhausted T Cells during Anti-PD-1 Therapy in Lung Cancer. *Nat. Cancer* 3, 108–121. doi:10.1038/s43018-021-00292-8

Liu, L., Liu, C., Quintero, A., Wu, L., Yuan, Y., Wang, M., et al. (2019). Deconvolution of Single-Cell Multi-Omics Layers Reveals Regulatory Heterogeneity. *Nat. Commun.* 10 (1), 470. doi:10.1038/s41467-018-08205-7

Liudahl, S. M., Betts, C. B., Sivagnanam, S., Morales-Oyarvide, V., da Silva, A., Yuan, C., et al. (2021). Leukocyte Heterogeneity in Pancreatic Ductal Adenocarcinoma: Phenotypic and Spatial Features Associated with Clinical Outcome. *Cancer Discov.* 11 (8), 2014–2031. doi:10.1158/2159-8290.CD-20-0841

L. Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across Cells to Normalize Single-Cell RNA Sequencing Data with Many Zero Counts. *Genome Biol.* 17, 75. doi:10.1186/s13059-016-0947-7

Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Wagenstetter, M., et al. (2022). Mapping Single-Cell Data to Reference Atlases by Transfer Learning. *Nat. Biotechnol.* 40 (1), 121–130. doi:10.1038/s41587-021-01001-7

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., et al. (2022). Benchmarking Atlas-Level Data Integration in Single-Cell Genomics. *Nat. Methods* 19 (1), 41–50. doi:10.1038/s41592-021-01336-8

Łuksza, M., Riaz, N., Makarov, V., Balachandran, V. P., Hellmann, M. D., Solovyov, A., et al. (2017). A Neoantigen Fitness Model Predicts Tumour Response to Checkpoint Blockade Immunotherapy. *Nature* 551 (7681), 517–520. doi:10.1038/nature24473

Ma, L., Hernandez, M. O., Zhao, Y., Mehta, M., Tran, B., Kelly, M., et al. (2019). Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer. *Cancer Cell* 36 (4), 418–430. e416. doi:10.1016/j.ccell.2019.08.007

Ma, S., and Zhang, Y. (2020). Profiling Chromatin Regulatory Landscape: Insights into the Development of ChIP-Seq and ATAC-Seq. *Mol. Biomed.* 1 (1), 9. doi:10.1186/s43556-020-00009-w

Mabbott, N. A., Baillie, J., Brown, H., Freeman, T. C., and Hume, D. A. (2013). An Expression Atlas of Human Primary Cells: Inference of Gene Function from Coexpression Networks. *BMC Genomics* 14, 632. doi:10.1186/1471-2164-14-632

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., et al. (2015). G&T-seq: Parallel Sequencing of Single-Cell Genomes and Transcriptomes. *Nat. Methods* 12 (6), 519–522. doi:10.1038/nmeth.3370

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161 (5), 1202–1214. doi:10.1016/j.cell.2015.05.002

Maley, C. C., Koelble, K., Natrajan, R., Aktipis, A., and Yuan, Y. (2015). An Ecological Measure of Immune-Cancer Colocalization as a Prognostic Factor for Breast Cancer. *Breast Cancer Res.* 17 (1), 131. doi:10.1186/s13058-015-0638-4

Manesse, M., Patel, K. K., Bobrow, M., and Downing, S. R. (2020). The InSituPlex Staining Method for Multiplexed Immunofluorescence Cell Phenotyping and Spatial Profiling of Tumor FFPE Samples. *Methods Mol. Biol.* 2055, 585–592. doi:10.1007/978-1-4939-9773-2_26

Marigo, I., Trovato, R., Hofer, F., Ingangi, V., Desantis, G., Leone, K., et al. (2020). Disabled Homolog 2 Controls Prometastatic Activity of Tumor-Associated Macrophages. *Cancer Discov.* 10 (11), 1758–1773. doi:10.1158/2159-8290.CD-20-0036

Marigo, I., Zilio, S., Desantis, G., Mlecnik, B., Agnellini, A. H. R., Ugel, S., et al. (2016). T Cell Cancer Therapy Requires CD40-Cd40l Activation of Tumor Necrosis Factor and Inducible Nitric-Oxide-Synthase-Producing Dendritic Cells. *Cancer Cell* 30 (4), 651. doi:10.1016/j.ccell.2016.09.009

Martens, J. H. A., and Stunnenberg, H. G. (2013). BLUEPRINT: Mapping Human Blood Cell Epigenomes. *Haematologica* 98 (10), 1487–1489. doi:10.3324/haematol.2013.094243

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S. H., Ko, S. B. H., Gouda, N., et al. (2017). SCODE: an Efficient Regulatory Network Inference Algorithm from Single-Cell RNA-Seq during Differentiation. *Bioinformatics* 33 (15), 2314–2321. doi:10.1093/bioinformatics/btx194

McGinnis, C. S., Murrow, L. M., and Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* 8 (4), 329–337. doi:10.1016/j.cels.2019.03.003

McGranahan, N., and Swanton, C. (2015). Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell* 27 (1), 15–26. doi:10.1016/j.ccell.2014.12.001

McGranahan, N., and Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* 168 (4), 613–628. doi:10.1016/j.cell.2017.01.018

Mimitou, E. P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., et al. (2019). Multiplexed Detection of Proteins, Transcriptomes, Clonotypes and CRISPR Perturbations in Single Cells. *Nat. Methods* 16 (5), 409–412. doi:10.1038/s41592-019-0392-0

Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., et al. (2019). GRNBoost2 and Arboreto: Efficient and Scalable Inference of Gene Regulatory Networks. *Bioinformatics* 35 (12), 2159–2161. doi:10.1093/bioinformatics/bty916

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., et al. (2019). Visualizing Structure and Transitions in High-Dimensional Biological Data. *Nat. Biotechnol.* 37 (12), 1482–1492. doi:10.1038/s41587-019-0336-3

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). *On Spectral Clustering: Analysis and an Algorithm*, 849–856.

Nirmal, A. J., Maliga, Z., Vallius, T., Quattrochi, B., Chen, A. A., Jacobson, C. A., et al. (2022). The Spatial Landscape of Progression and Immunoediting in Primary Melanoma at Single Cell Resolution. *Cancer Discov.* doi:10.1158/2159-8290.CD-21-1357

Nolan, J. P., Condello, D., Duggan, E., Naivar, M., and Novo, D. (2013). Visible and Near Infrared Fluorescence Spectral Flow Cytometry. *Cytometry* 83A (3), 253–264. doi:10.1002/cyto.a.22241

Nosho, K., Baba, Y., Tanaka, N., Shima, K., Hayashi, M., Meyerhardt, J. A., et al. (2010). Tumour-infiltrating T-Cell Subsets, Molecular Changes in Colorectal Cancer, and Prognosis: Cohort Study and Literature Review. *J. Pathol.* 222 (4), 350–366. doi:10.1002/path.2774

Ogino, S., Nosho, K., Irahara, N., Meyerhardt, J. A., Baba, Y., Shima, K., et al. (2009). Lymphocytic Reaction to Colorectal Cancer Is Associated with Longer Survival, Independent of Lymph Node Count, Microsatellite Instability, and CpG Island Methylator Phenotype. *Clin. Cancer Res.* 15 (20), 6412–6420. doi:10.1158/1078-0432.CCR-09-1438

Osorio, D., and Cai, J. J. (2021). Systematic Determination of the Mitochondrial Proportion in Human and Mice Tissues for Single-Cell RNA-Sequencing Data Quality Control. *Bioinformatics* 37 (7), 963–967. doi:10.1093/bioinformatics/btaa751

Papalexi, E., Mimitou, E. P., Butler, A. W., Foster, S., Bracken, B., Mauck, W. M., 3rd, et al. (2021). Characterizing the Molecular Regulation of Inhibitory Immune Checkpoints with Multimodal Single-Cell Screens. *Nat. Genet.* 53 (3), 322–331. doi:10.1038/s41588-021-00778-2

Peng, J., Sun, B.-F., Chen, C.-Y., Zhou, J.-Y., Chen, Y.-S., Chen, H., et al. (2019). Single-cell RNA-Seq Highlights Intra-tumoral Heterogeneity and Malignant Progression in Pancreatic Ductal Adenocarcinoma. *Cell Res.* 29 (9), 725–738. doi:10.1038/s41422-019-0195-y

Peregrin, J., Svěrák, J., and Doskocil, Z. (1973). Photostimulator for the Study of Visual Functions. *Sb. Ved. P. R. Lek. Fak. Karlovy Univerzity Hradci Kralove Suppl.* 16 (3), 203–207.

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., et al. (2017). Multiplexed Quantification of Proteins and Transcripts in Single Cells. *Nat. Biotechnol.* 35 (10), 936–939. doi:10.1038/nbt.3973

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-Seq from Single Cells Using Smart-Seq2. *Nat. Protoc.* 9 (1), 171–181. doi:10.1038/nprot.2014.006

Pombo Antunes, A. R., Scheyltjens, I., Lodi, F., Messiaen, J., Antoranz, A., Duerinck, J., et al. (2021). Single-cell Profiling of Myeloid Cells in Glioblastoma across Species and Disease Stage Reveals Macrophage Competition and Specialization. *Nat. Neurosci.* 24 (4), 595–610. doi:10.1038/s41593-020-00789-y

Pott, S. (2017). Simultaneous Measurement of Chromatin Accessibility, DNA Methylation, and Nucleosome Phasing in Single Cells. *Elife* 6. doi:10.7554/eLife.23203

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking Algorithms for Gene Regulatory Network Inference from Single-Cell Transcriptomic Data. *Nat. Methods* 17 (2), 147–154. doi:10.1038/s41592-019-0690-6

Pylayeva-Gupta, Y., Lee, K. E., Hajdu, C. H., Miller, G., and Bar-Sagi, D. (2012). Oncogenic Kras-Induced GM-CSF Production Promotes the Development of Pancreatic Neoplasia. *Cancer Cell* 21 (6), 836–847. doi:10.1016/j.ccr.2012.04.024

Qian, J., Olbrecht, S., Boeckx, B., Vos, H., Laoui, D., Etlioglu, E., et al. (2020). A Pan-Cancer Blueprint of the Heterogeneous Tumor Microenvironment Revealed by Single-Cell Profiling. *Cell Res.* 30 (9), 745–762. doi:10.1038/s41422-020-0355-0

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., et al. (2017). Reversed Graph Embedding Resolves Complex Single-Cell Trajectories. *Nat. Methods* 14 (10), 979–982. doi:10.1038/nmeth.4402

Raghavan, S., Winter, P. S., Navia, A. W., Williams, H. L., DenAdel, A., Lowder, K. E., et al. (2021). Microenvironment Drives Cell State, Plasticity, and Drug Response in Pancreatic Cancer. *Cell* 184 (25), 6119–6137. e6126. doi:10.1016/j.cell.2021.11.017

Rambow, F., Rogiers, A., Marin-Bejar, O., Aibar, S., Femel, J., Dewaele, M., et al. (2018). Toward Minimal Residual Disease-Directed Therapy in Melanoma. *Cell* 174 (4), 843–855. doi:10.1016/j.cell.2018.06.025

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length mRNA-Seq from Single-Cell Levels of RNA and Individual Circulating Tumor Cells. *Nat. Biotechnol.* 30 (8), 777–782. doi:10.1038/nbt.2282

Reddy, E. P., Reynolds, R. K., Santos, E., and Barbacid, M. (1982). A Point Mutation Is Responsible for the Acquisition of Transforming Properties by the T24 Human Bladder Carcinoma Oncogene. *Nature* 300 (5888), 149–152. doi:10.1038/300149a0

Remark, R., Merghoub, T., Grabe, N., Litjens, G., Damotte, D., Wolchok, J. D., et al. (2016). In-depth Tissue Profiling Using Multiplexed Immunohistochemical Consecutive Staining on Single Slide. *Sci. Immunol.* 1 (1), aaf6925. doi:10.1126/sciimmunol.aaf6925

Robinson, J. P. (2004). Multispectral Cytometry: the Next Generation. *Biophot. Int.* 11, 36–41.

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J. E., et al. (2020). The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* 181 (2), 236–249. doi:10.1016/j.cell.2020.03.053

Rubin, A. J., Parker, K. R., Satpathy, A. T., Qi, Y., Wu, B., Ong, A. J., et al. (2019). Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* 176 (1-2), 361–376. doi:10.1016/j.cell.2018.11.022

Rybakowska, P., Alarcón-Riquelme, M. E., and Marañón, C. (2020). Key Steps and Methods in the Experimental Design and Data Analysis of Highly Multi-Parametric Flow and Mass Cytometry. *Comput. Struct. Biotechnol. J.* 18, 874–886. doi:10.1016/j.csbj.2020.03.024

Sade-Feldman, M., Yizhak, K., Bjorgaard, S. L., Ray, J. P., de Boer, C. G., Jenkins, R. W., et al. (2018). Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* 175 (4), 998–1013. doi:10.1016/j.cell.2018.10.038

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A Comparison of Single-Cell Trajectory Inference Methods. *Nat. Biotechnol.* 37 (5), 547–554. doi:10.1038/s41587-019-0071-9

Sakamoto, K. M., Kim, K. B., Kumagai, A., Mercurio, F., Crews, C. M., and Deshaies, R. J. (2001). Protacs: Chimeric Molecules that Target Proteins to the Skp1-Cullin-F Box Complex for Ubiquitination and Degradation. *Proc. Natl. Acad. Sci. U.S.A.* 98 (15), 8554–8559. doi:10.1073/pnas.141230798

Sanmamed, M. F., Nie, X., Desai, S. S., Villaroel-Espindola, F., Badri, T., Zhao, D., et al. (2021). A Burned-Out CD8+ T-Cell Subset Expands in the Tumor Microenvironment and Curbs Cancer Immunotherapy. *Cancer Discov.* 11 (7), 1700–1715. doi:10.1158/2159-8290.CD-20-0962

Santos, E., Martin-Zanca, D., Reddy, E. P., Pierotti, M. A., Della Porta, G., and Barbacid, M. (1984). Malignant Activation of a K- Ras Oncogene in Lung Carcinoma but Not in Normal Tissue of the Same Patient. *Science* 223 (4637), 661–664. doi:10.1126/science.6695174

Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., et al. (2019). Massively Parallel Single-Cell Chromatin Landscapes of Human Immune Cell Development and Intratumoral T Cell Exhaustion. *Nat. Biotechnol.* 37 (8), 925–936. doi:10.1038/s41587-019-0206-z

Sautès-Fridman, C., Petitprez, F., Calderaro, J., and Fridman, W. H. (2019). Tertiary Lymphoid Structures in the Era of Cancer Immunotherapy. *Nat. Rev. Cancer* 19 (6), 307–325. doi:10.1038/s41568-019-0144-6

Schreiber, R. D., Old, L. J., and Smyth, M. J. (2011). Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion. *Science* 331 (6024), 1565–1570. doi:10.1126/science.1203486

Schürch, C. M., Bhate, S. S., Barlow, G. L., Phillips, D. J., Noti, L., Zlobec, I., et al. (2020). Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell* 182 (5), 1341–1359. doi:10.1016/j.cell.2020.07.005

Sheng, K., Cao, W., Niu, Y., Deng, Q., and Zong, C. (2017). Effective Detection of Variation in Single-Cell Transcriptomes Using MATQ-Seq. *Nat. Methods* 14 (3), 267–270. doi:10.1038/nmeth.4145

Singh, R., Hie, B. L., Narayan, A., and Berger, B. (2021). Schema: Metric Learning Enables Interpretable Synthesis of Heterogeneous Single-Cell Modalities. *Genome Biol.* 22 (1), 131. doi:10.1186/s13059-021-02313-2

Specht, A. T., and Li, J. (2017). LEAP: Constructing Gene Co-expression Networks for Single-Cell RNA-Sequencing Data Using Pseudotime Ordering. *Bioinformatics* 33 (5), btw729–766. doi:10.1093/bioinformatics/btw729

Specht, H., Emmott, E., Petelski, A. A., Huffman, R. G., Perlman, D. H., Serra, M., et al. (2021). Single-cell Proteomic and Transcriptomic Analysis of Macrophage Heterogeneity Using SCoPE2. *Genome Biol.* 22 (1), 50. doi:10.1186/s13059-021-02267-5

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., et al. (2021). Confronting False Discoveries in Single-Cell Differential Expression. *Nat. Commun.* 12 (1), 5692. doi:10.1038/s41467-021-25960-2

Stassen, S. V., Yip, G. G. K., Wong, K. K. Y., Ho, J. W. K., and Tsia, K. K. (2021). Generalized and Scalable Trajectory Inference in Single-Cell Omics Data with VIA. *Nat. Commun.* 12 (1), 5528. doi:10.1038/s41467-021-25773-3

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., et al. (2017). Simultaneous Epitope and Transcriptome Measurement in Single Cells. *Nat. Methods* 14 (9), 865–868. doi:10.1038/nmeth.4380

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B. Z., Mauck, W. M., 3rd, et al. (2018). Cell Hashing with Barcoded Antibodies Enables Multiplexing and Doublet Detection for Single Cell Genomics. *Genome Biol.* 19 (1), 224. doi:10.1186/s13059-018-1603-1

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, et al. (2019). Comprehensive Integration of Single-Cell Data. *Cell* 177 (7), 1888–1902. doi:10.1016/j.cell.2019.05.031

Subrahmanyam, P. B., Dong, Z., Gusenleitner, D., Giobbie-Hurder, A., Severgnini, M., Zhou, J., et al. (2018). Distinct Predictive Biomarker Candidates for Response to Anti-CTLA-4 and Anti-PD-1 Immunotherapy in Melanoma Patients. *J. Immunother. cancer* 6 (1), 18. doi:10.1186/s40425-018-0328-8

Subramanian, A., Alperovich, M., Yang, Y., and Li, B. (2021). Biology-inspired Data-Driven Quality Control for Scientific Discovery in Single-Cell Transcriptomics. *bioRxiv*. doi:10.1101/2021.10.27.466176

Taavitsainen, S., Engedal, N., Cao, S., Handle, F., Erickson, A., Prekovic, S., et al. (2021). Single-cell ATAC and RNA Sequencing Reveal Pre-existing and Persistent Cells Associated with Prostate Cancer Relapse. *Nat. Commun.* 12 (1), 5307. doi:10.1038/s41467-021-25624-1

Tang, H., Wang, Y., Chlewicki, L. K., Zhang, Y., Guo, J., Liang, W., et al. (2016). Facilitating T Cell Infiltration in Tumor Microenvironment Overcomes Resistance to PD-L1 Blockade. *Cancer Cell* 29 (3), 285–296. doi:10.1016/j.ccell.2016.02.004

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 47 (D1), D941–D947. doi:10.1093/nar/gky1015

Tawbi, H. A., Schadendorf, D., Lipson, E. J., Ascierto, P. A., Matamala, L., Castillo Gutiérrez, E., et al. (2022). Relatlimab and Nivolumab versus Nivolumab in Untreated Advanced Melanoma. *N. Engl. J. Med.* 386 (1), 24–34. doi:10.1056/NEJMoa2109970

Tedesco, M., Giannese, F., Lazarević, D., Giansanti, V., Rosano, D., Monzani, S., et al. (2021). Chromatin Velocity Reveals Epigenetic Dynamics by Single-Cell Profiling of Heterochromatin and Euchromatin. *Nat. Biotechnol.* 40, 235–244. doi:10.1038/s41587-021-01031-1

Telenius, H. k., Carter, N. P., Bebb, C. E., Nordenskjo¨ld, M., Ponder, B. A. J., and Tunnacliffe, A. (1992). Degenerate Oligonucleotide-Primed PCR: General Amplification of Target DNA by a Single Degenerate Primer. *Genomics* 13 (3), 718–725. doi:10.1016/0888-7543(92)90147-k

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T. H., et al. (2019). The Immune Landscape of Cancer. *Immunity* 51 (2), 411–412. doi:10.1016/j.immuni.2019.08.004

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., 2nd, Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq. *Science* 352 (6282), 189–196. doi:10.1126/science.aad0501

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., et al. (2005). Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. *Science* 310 (5748), 644–648. doi:10.1126/science.1117679

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci. Rep.* 9 (1), 5233. doi:10.1038/s41598-019-41695-z

Tsujikawa, T., Kumar, S., Borkar, R. N., Azimi, V., Thibault, G., Chang, Y. H., et al. (2017). Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. *Cell Rep.* 19 (1), 203–217. doi:10.1016/j.celrep.2017.03.037

Ugel, S., Canè, S., De Sanctis, F., and Bronte, V. (2021). Monocytes in the Tumor Microenvironment. *Annu. Rev. Pathol. Mech. Dis.* 16, 93–122. doi:10.1146/annurev-pathmechdis-012418-013058

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., et al. (2015). FlowSOM: Using Self-Organizing Maps for Visualization and Interpretation of Cytometry Data. *Cytometry* 87 (7), 636–645. doi:10.1002/cyto.a.22625

Van Gassen, S., Gaudilliere, B., Angst, M. S., Saeys, Y., and Aghaeepour, N. (2020). CytoNorm: A Normalization Algorithm for Cytometry Data. *Cytometry* 97 (3), 268–278. doi:10.1002/cyto.a.23904

Velazquez-Villarreal, E. I., Maheshwari, S., Sorenson, J., Fiddes, I. T., Kumar, V., Yin, Y., et al. (2020). Single-cell Sequencing of Genomic DNA Resolves Sub-clonal Heterogeneity in a Melanoma Cell Line. *Commun. Biol.* 3 (1), 318. doi:10.1038/s42003-020-1044-8

Viratham Pulsawatdi, A., Craig, S. G., Bingham, V., McCombe, K., Humphries, M. P., Senevirathne, S., et al. (2020). A Robust Multiplex Immunofluorescence and Digital Pathology Workflow for the Characterisation of the Tumour Immune Microenvironment. *Mol. Oncol.* 14 (10), 2384–2402. doi:10.1002/1878-0261.12764

Vitak, S. A., Torkenczy, K. A., Rosenkrantz, J. L., Fields, A. J., Christiansen, L., Wong, M. H., et al. (2017). Sequencing Thousands of Single-Cell Genomes with Combinatorial Indexing. *Nat. Methods* 14 (3), 302–308. doi:10.1038/nmeth.4154

Vitale, I., Manic, G., Coussens, L. M., Kroemer, G., and Galluzzi, L. (2019). Macrophages and Metabolism in the Tumor Microenvironment. *Cell Metab.* 30 (1), 36–50. doi:10.1016/j.cmet.2019.06.001

Wagner, J., Rapsomaniki, M. A., Chevrier, S., Anzeneder, T., Langwieder, C., Dykgers, A., et al. (2019). A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell* 177 (5), 1330–1345. e1318. doi:10.1016/j.cell.2019.03.005

Waltman, L., and Van Eck, N. J. (2013). A Smart Local Moving Algorithm for Large-Scale Modularity-Based Community Detection. *Eur. Phys. J. B* 86 (11), 11434–146036. doi:10.1140/epjb/e2013-40829-0

Wang, Q., Xiong, H., Ai, S., Yu, X., Liu, Y., Zhang, J., et al. (2019). CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. *Mol. Cell* 76 (1), 206–216. doi:10.1016/j.molcel.2019.07.015

Wang, X., Sun, Z., Zhang, Y., Xu, Z., Xin, H., Huang, H., et al. (2020). BREM-SC: a Bayesian Random Effects Mixture Model for Joint Clustering Single Cell Multi-Omics Data. *Nucleic Acids Res.* 48 (11), 5814–5824. doi:10.1093/nar/gkaa314

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., et al. (2014). Clonal Evolution in Breast Cancer Revealed by Single Nucleus Genome Sequencing. *Nature* 512 (7513), 155–160. doi:10.1038/nature13600

Wang, Y., Yuan, P., Yan, Z., Yang, M., Huo, Y., Nie, Y., et al. (2021). Single-cell Multiomics Sequencing Reveals the Functional Regulatory Landscape of Early Embryos. *Nat. Commun.* 12 (1), 1247. doi:10.1038/s41467-021-21409-8

Wei, R., He, S., Bai, S., Sei, E., Hu, M., Thompson, A., et al. (2022). Spatial Charting of Single-Cell Transcriptomes in Tissues. *Nat. Biotechnol.* doi:10.1038/s41587-022-01233-1

Weinreb, C., Wolock, S., and Klein, A. M. (2018). SPRING: a Kinetic Interface for Visualizing High Dimensional Single-Cell Expression Data. *Bioinformatics* 34 (7), 1246–1248. doi:10.1093/bioinformatics/btx792

Whiteside, T. L. (2008). The Tumor Microenvironment and its Role in Promoting Tumor Growth. *Oncogene* 27 (45), 5904–5912. doi:10.1038/onc.2008.271

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biol.* 19 (1), 15. doi:10.1186/s13059-017-1382-0

Wolock, S. L., Lopez, R., and Klein, A. M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* 8 (4), 281–291. doi:10.1016/j.cels.2018.11.005

Wu, S. Z., Al-Eryani, G., Roden, D. L., Junankar, S., Harvey, K., Andersson, A., et al. (2021). A Single-Cell and Spatially Resolved Atlas of Human Breast Cancers. *Nat. Genet.* 53 (9), 1334–1347. doi:10.1038/s41588-021-00911-1

Xu, C., and Su, Z. (2015). Identification of Cell Types from Single-Cell Transcriptomes Using a Novel Clustering Method. *Bioinformatics* 31 (12), 1974–1980. doi:10.1093/bioinformatics/btv088

Yang, W., Feng, Y., Zhou, J., Cheung, O. K.-W., Cao, J., Wang, J., et al. (2021). A Selective HDAC8 Inhibitor Potentiates Antitumor Immunity and Efficacy of Immune Checkpoint Blockade in Hepatocellular Carcinoma. *Sci. Transl. Med.* 13 (588). doi:10.1126/scitranslmed.aaz6804

Yang, Z.-Z., Kim, H. J., Wu, H., Jalali, S., Tang, X., Krull, J. E., et al. (2020). TIGIT Expression Is Associated with T-Cell Suppression and Exhaustion and Predicts

Clinical Outcome and Anti-PD-1 Response in Follicular Lymphoma. *Clin. Cancer Res.* 26 (19), 5217–5231. doi:10.1158/1078-0432.CCR-20-0558

Yin, Y., Jiang, Y., Lam, K.-W. G., Berletch, J. B., Disteche, C. M., Noble, W. S., et al. (2019). High-Throughput Single-Cell Sequencing with Linear Amplification. *Mol. Cell* 76 (4), 676–690. e610. doi:10.1016/j.molcel.2019.08.002

Yip, S. H., Wang, P., Kocher, J.-P. A., Sham, P. C., and Wang, J. (2017). Linnorm: Improved Statistical Analysis for Single Cell RNA-Seq Expression Data. *Nucleic Acids Res.* 45 (22), e179. doi:10.1093/nar/gkx828

Yuan, Y. (2016). Spatial Heterogeneity in the Tumor Microenvironment. *Cold Spring Harb. Perspect. Med.* 6 (8), a026583. doi:10.1101/cshperspect.a026583

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., et al. (2013). Pan-cancer Patterns of Somatic Copy Number Alteration. *Nat. Genet.* 45 (10), 1134–1140. doi:10.1038/ng.2760

Zahn, H., Steif, A., Laks, E., Eirew, P., VanInsberghe, M., Shah, S. P., et al. (2017). Scalable Whole-Genome Single-Cell Library Preparation without Preamplification. *Nat. Methods* 14 (2), 167–173. doi:10.1038/nmeth.4140

Zhang, L., Chen, D., Song, D., Liu, X., Zhang, Y., Xu, X., et al. (2022). Clinical and Translational Values of Spatial Transcriptomics. *Sig Transduct. Target Ther.* 7 (1), 111. doi:10.1038/s41392-022-00960-w

Zhang, L., Li, Z., Skrzypczynska, K. M., Fang, Q., Zhang, W., O'Brien, S. A., et al. (2020). Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. *Cell* 181 (2), 442–459. doi:10.1016/j.cell.2020.03.048

Zhang, Q., He, Y., Luo, N., Patel, S. J., Han, Y., Gao, R., et al. (2019). Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. *Cell* 179 (4), 829–845. doi:10.1016/j.cell.2019.10.003

Zhang, Y., Chen, H., Mo, H., Hu, X., Gao, R., Zhao, Y., et al. (2021). Single-cell Analyses Reveal Key Immune Cell Subsets Associated with Response to PD-L1 Blockade in Triple-Negative Breast Cancer. *Cancer Cell* 39 (12), 1578–1593. e1578. doi:10.1016/j.ccell.2021.09.010

Zhao, J., Zhang, S., Liu, Y., He, X., Qu, M., Xu, G., et al. (2020a). Single-cell RNA Sequencing Reveals the Heterogeneity of Liver-Resident Immune Cells in Human. *Cell Discov.* 6, 22. doi:10.1038/s41421-020-0157-z

Zhao, X., Wu, S., Fang, N., Sun, X., and Fan, J. (2020b). Evaluation of Single-Cell Classifiers for Single-Cell RNA Sequencing Data Sets. *Brief. Bioinform* 21 (5), 1581–1595. doi:10.1093/bib/bbz096

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively Parallel Digital Transcriptional Profiling of Single Cells. *Nat. Commun.* 8, 14049. doi:10.1038/ncomms14049

Zhu, C., Preissl, S., and Ren, B. (2020). Single-cell Multimodal Omics: the Power of Many. *Nat. Methods* 17 (1), 11–14. doi:10.1038/s41592-019-0691-5

Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., et al. (2019). An Ultra High-Throughput Method for Single-Cell Joint Analysis of Open Chromatin and Transcriptome. *Nat. Struct. Mol. Biol.* 26 (11), 1063–1070. doi:10.1038/s41594-019-0323-x

Zhu, J., Powis de Tenbossche, C. G., Cané, S., Colau, D., van Baren, N., Lurquin, C., et al. (2017). Resistance to Cancer Immunotherapy Mediated by Apoptosis of Tumor-Infiltrating Lymphocytes. *Nat. Commun.* 8 (1), 1404. doi:10.1038/s41467-017-00784-1

Zhu, Z., Wang, W., Lin, F., Jordan, T., Li, G., Silverman, S., et al. (2021). Genome Profiles of Pathologist-Defined Cell Clusters by Multiregional LCM and G&T-seq in One Triple-Negative Breast Cancer Patient. *Cell Rep. Med.* 2 (10), 100404. doi:10.1016/j.xcrm.2021.100404

Zilionis, R., Engblom, C., Pfirschke, C., Savova, V., Zemmour, D., Saatcioglu, H. D., et al. (2019). Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. *Immunity* 50 (5), 1317–1334. doi:10.1016/j.immuni.2019.03.009

Zong, C., Lu, S., Chapman, A. R., and Xie, X. S. (2012). Genome-wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science* 338 (6114), 1622–1626. doi:10.1126/science.1229164

Zuo, C., and Chen, L. (2021). Deep-joint-learning Analysis Model of Single Cell Transcriptome and Open Chromatin Accessibility Data. *Brief. Bioinform* 22 (4). doi:10.1093/bib/bbaa287

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Caligola, De Sanctis, Canè and Ugel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 26 May 2022
doi: 10.3389/fevo.2022.850183

Genetic Diversity, Habitat Relevance and Conservation Strategies of the Silver Carp in the Yangtze River by Simple Sequence Repeat

Yu-Ting Luo[1†], Di-An Fang[1,2†], Yan-Feng Zhou[2], Dong-Po Xu[2], Yun-Xin Peng[3], Min-Ying Zhang[2], Cheng-Cheng Mao[1], Xue-Mei Tang[1], Jun Xu[4] and Yang You[1,2]*

[1] Wuxi Fisheries College, Nanjing Agricultural University, Wuxi, China, [2] Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi, China, [3] National Demonstration Center for Experimental Fisheries Science Education, Shanghai Ocean University, Shanghai, China, [4] Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, China

The spatial distribution of fish individuals is affected by habitat conditions and species ecological characteristics, and it also reflects the longtime adaptation to habitat at the phenotypic and genotypic level. As a typical river-lake migratory fish species, the silver carp habitat selection was determined by its migration preference and genetic features. In this study, 15 microsatellite fluorescent markers combined with capillary electrophoresis were used to analyze the genetic diversity, genetic differentiation, and structure of nature silver carp populations in the Sanzhou (SZ), Hukou (HK), Anqing (AQ), Zhenjiang (ZJ), and Rugao (RG) sections of the Yangtze River. The results showed that 15 microsatellite loci exhibited medium to high polymorphisms. The overall genetic diversity in the Yangtze River was high, with the average value of Shannon's information index ranging from 1.559 to 1.668. The numbers of alleles (Ne) ranged from 1.630 to 10.100. The expected heterozygosity (He, 0.690–0.721) was higher than observed heterozygosity (Ho, 0.598–0.646), and the genetic variation mainly originated from within the population (94.69%). However, the entire population was in the state of heterozygous deletion, and HK, RG populations encountered the risk of inbreeding risk ($F > 1$). Interestingly, there was a distinct genetic structure for the population in the HK section, which indicated that local population has occurred to the silver carp in this river section, and they may also possess aggregation characteristics specific to the river-lake-connected (RLC) habitat. The results mostly support the conclusion that the RLC habitat is essential for geographic population formation. The potential impact of special habitats on natural populations should be considered, and continuous surveys on population dynamics should be performed.

Keywords: silver carp, the Yangtze River, habitat selection, SSR, stock enhancement

OPEN ACCESS

Edited by:
Peng Xu,
Xiamen University, China

Reviewed by:
Hong Wei Liang,
Chinese Academy of Fishery Sciences (CAFS), China
Jian Xu,
Chinese Academy of Fishery Sciences (CAFS), China

***Correspondence:**
Yan-Feng Zhou
zhouyf@ffrc.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

Received: *07 January 2022*
Accepted: *22 March 2022*
Published: *26 May 2022*

Citation:
Luo YT, Fang DA, Zhou YF, Xu DP, Peng YX, Zhang MY, Mao CC, Tang XM, Xu J and You Y (2022) Genetic Diversity, Habitat Relevance and Conservation Strategies of the Silver Carp in the Yangtze River by Simple Sequence Repeat. Front. Ecol. Evol. 10:850183. doi: 10.3389/fevo.2022.850183

INTRODUCTION

The silver carp (*Hypophthalmichthys molitrix*) belongs to the Cypriniformes order, Cyprinidae family, and *Hypophthalmichthys* genus, and it is one of four major Chinese carp (Ni and Wu, 2006). Paleontological analyses demonstrated that the silver carp originated in the Yangtze-Huanghe River basin (Li and Fang, 1990). Currently, silver carp is widely distributed and is found in the basin of the Red River, the Pearl River, and the Heilongjiang River in China (Lu et al., 2020).

TABLE 1 | Sampling of silver carp in the middle and lower Yangtze River.

Sampling sites	Coordinates	Code	Sampling time	Sample size
Sanzhou, Hubei	112°55′55″E, 29°32′33″N	SZ	2019	91
Hukou, Jiangxi	116°15′32″E, 29°47′02″N	HK	2019–2020	60
Anqing, Anhui	117°0′18″E, 30°29′28″N	AQ	2019–2021	70
Zhenjiang, Jiangsu	119°20′45″E, 32°11′42″N	ZJ	2018	51
Rugao, Jiangsu	120°31′23″E, 32°3′27″N	RG	2017	63
Total				335

Artificial cultivation technology of silver carp was founded in 1958, which got rid of the passive situation of relying on catching and promoted the development of aquaculture (Mao et al., 2010). Then silver carp were transplanted to rivers in Europe, the United States, and Africa (Pinter, 1980), and even became an invasive fish species (Kolar and Lodge, 2002; Conover et al., 2007). The overseas population of silver carp continued to expand, while the Chinese indigenous population exhibited a declining trend (Li, 1996). Because of the construction of Three Gorges Dam, water pollution, and overfishing, fishery output was 427,000 tons in 1954. From 1956 to 1960, the fishing volume decreased to 260,000 tons, 200,000 tons in the 1980s, and reduced to about 100,000 tons 2000s, less than 1/4 of the maximum annual output (Mai, 2003). The four major carps accounted for 46.15 percent of the catch weight, but that figure decreased to only 10 percent in 2001–2003. The proportion of silver carp and bighead Carp among the four major fish decreased significantly (Zeng, 1990; Liu et al., 2005; Li and Xu, 2008). With a 10-year ban on fishing in the Yangtze River initiated in January 2020 and the continuous implementation of stock enhancement, the populations of the four major Chinese carp are in the process of recovery. However, stock enhancement and unscientific artificial release may result in negative ecological impacts such as impaired growth, disease spread, and decrement of genetic diversity (Liu et al., 1997; Bell et al., 2006; Fang et al., 2021).

The Yangtze River is the longest and largest river in China, and is the main germplasm resource area for silver carp. Studies have reported that there are 11 spawning grounds for the four major Chinese carp in the middle and lower reaches of the Yangtze River (Xu et al., 2017), and silver carp spawning grounds are still being found (Tang et al., 2010; He et al., 2021). As a fish species that migrates between rivers and lakes, the silver carp spawns in the main stream of the Yangtze River every breeding season (from April to July). The postpartum parent fish and the young fish enter the lake connected to the Yangtze River for feeding (Xu et al., 2017).

Moving away from an unsuitable environment is one of the most important adaptive strategies for fish (Matter et al., 1989). During the process of migration, adult fish are able to be selective regarding their habitat, and they would choose to remain in habitats with appropriate environmental conditions, especially if there are abundant food sources and little interference (Bonte and Maelfait, 2004). For example, studies on the manini (*Acanthurus triostegus* and *A. vicensis*) fish habitat found that a covered shallow water area was its preferred habitat (Sale, 1968). Similarly, silver carp requires a great deal of energy input for the process of reproduction and fattening. When the environment lacks sufficient resources, this will lead to its continuous exploration to find a suitable habitat (McMahon and Matter, 2006). There is abundant plankton bait in river-lake-connected (RLC) habitats, which are ideal living places for silver carp (Liu et al., 2019). During the growth period for silver carp, adults will choose a habitat with rich bait for growth and fattening.

The Hukou section connects Poyang Lake with the Yangtze River, which is the largest lake connected to the Yangtze River in the lower reaches (He et al., 2021). The Anqing section connects the Wan River, which also forms a unique complex ecosystem with RLC characteristics. Because these bodies of water are rich in aquatic biological resources, they provide a key channel for fish migration behavior (Liu et al., 2019). These similar habitats provide perfect feeding and reproduction grounds for the silver carp in the Yangtze River. This habitat connectivity is vital for ecosystem function and the distribution of biota (Lindenmayer et al., 2008). However, studies on silver carp in the Yangtze River have mainly concentrated on germplasm genetic diversity, and it has been proven that there is genetic differentiation in fish from different river sections (Zhu et al., 2007; Chen et al., 2018; Sha et al., 2018). The value of connectivity in conservation has been poorly understood, and few genetic studies have linked population dynamics to habitat selection.

The simple sequence repeat (SSR) is a powerful genetic marker that has been widely used in fishery assessment. It has the advantages of large number of markers, wide distribution in gene sequence, neutrality, and co-dominance. It is one of the ideal molecular markers for genetic evaluation (Sun et al., 2008). It was previously reported that five microsatellite loci were used to investigate the population structure of *Culterery thropterus* from seven lakes in the middle and lower reaches of the Yangtze River (Wang et al., 2007). Additionally, sixteen SSR polymorphic loci were used to construct a partial genomic library of bighead carp (*Aristichthys nobilis*) (Cheng et al., 2008). In this study, fifteen microsatellite loci were used to analyze the genetic diversity and genetic structure of silver carp populations in the Sanzhou (SZ), Hukou (HK), Anqing (AQ), Zhenjiang (ZJ), and Rugao (RG) sections of the Yangtze River. The population's genetic structure difference between sampled sections in the Yangtze River was clarified, so that we can deduce potential relation between the fish distribution pattern and habitat differences.

MATERIALS AND METHODS

Sample Collection

Five nature populations of adult silver carp in the Yangtze River were collected by a traditional net trap from 2017 to 2021 avoiding the stock enhancement period. Totally 335 tail fins of silver carp were non-destructively sampled and stored in absolute ethanol for further study. The sampling sites, coordinates, time and size were listed in **Table 1**, and the distribution of sampling sites was shown in **Figure 1**. SZ belongs to the middle reaches of the Yangtze River, HK, AQ, ZJ, RG belongs to the lower reaches of the Yangtze River.

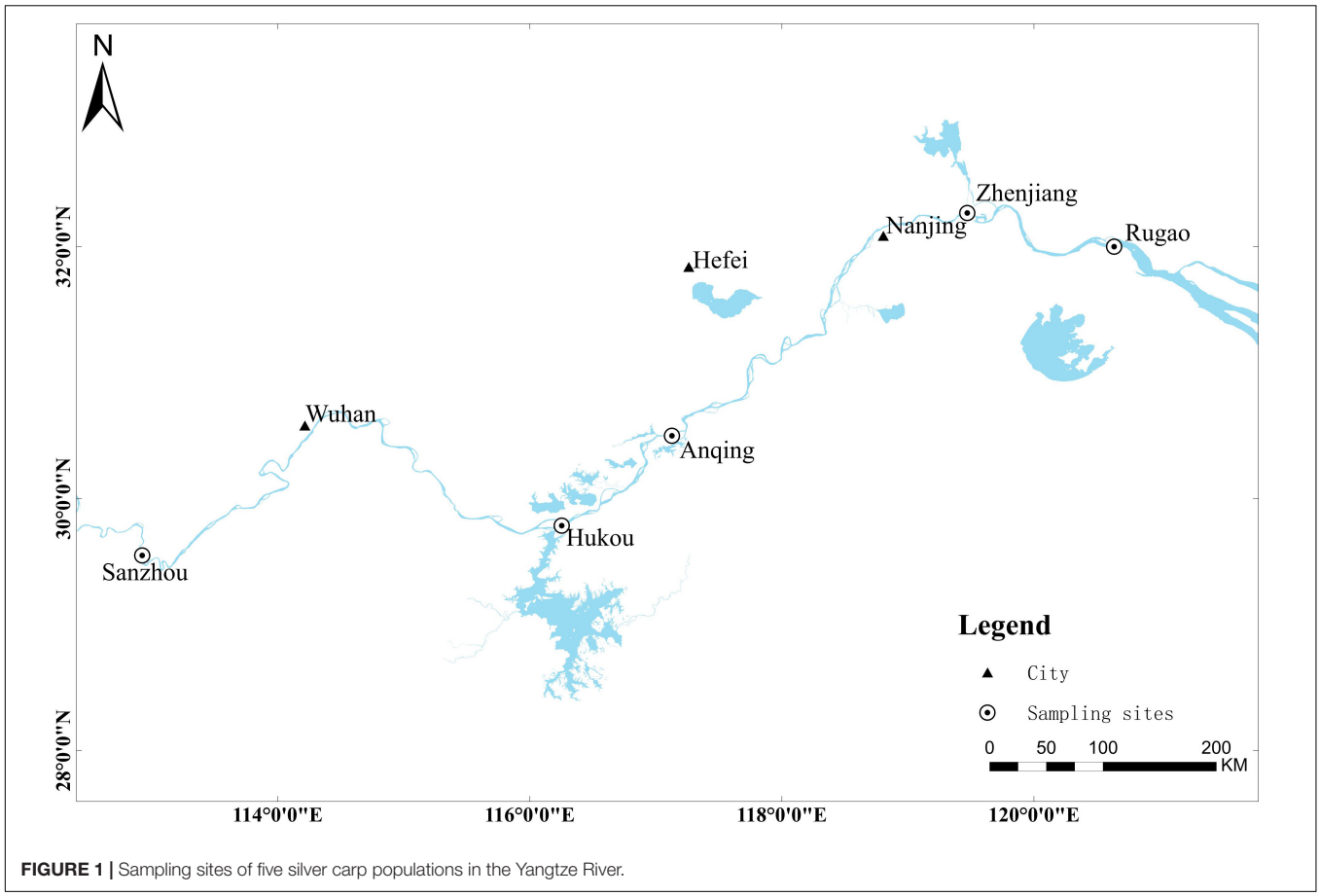


FIGURE 1 | Sampling sites of five silver carp populations in the Yangtze River.

DNA Extraction and Microsatellite Genotyping

Total genomic DNA was extracted using a TIANamp Marine Animals DNA Kit (TIANGEN Biotech Co., Ltd., Beijing, China) following the manufacturer's instructions. The quality of the extracted DNA was assessed by 1.2% agarose gel electrophoresis to ensure that successful amplification could be accomplished for all DNA strands, and then, the DNA was stored at −20°C for further study.

Fifteen polymorphic microsatellites were proactively identified and also used in another study (Tan et al., 2011). Each primer was synthesized by Sangon Biotech (Shanghai Co., Ltd., Shanghai, China) and fluorescently labeled at the 5′ terminus (**Table 2**). The PCR application of SSR markers was carried out in a total volume of 10 μL and consisted of 5 μL Taq Premix (TaKaRaTaq^{TM} Version 2.0 plus dye, Takara, Dalian, China), 0.2 μL (10 μM/OD) primer pairs, 1 μL (200 ng/μL) genomic DNA, and 3.8 μL ddH_2O. PCRs were performed in a 96-well thermal cycler (Gene Co., Ltd., Shanghai) using the following conditions: 94°C for 120 s; 30 cycles of 94°C for 20 s, 59°C for 20 s, 72°C for 20 s; extension at 72°C for 600 s, and finally, storage at 4°C. The PCR products were qualified by 2% agarose gel electrophoresis and the G:BOX automatic gel imaging system. Qualified PCR products were sent to Sangon Biotech (Shanghai Co., Ltd., Shanghai, China) for capillary electrophoresis sequencing and typing, which was performed using an ABI Prism 3730 XL automated sequencer (Rox-500 standard).

Genetic Analysis

After genotyping based on capillary electrophoresis technology, we manually confirmed the accuracy of genotypes. Pedigree analysis software Cervus 3.0.7 was used to analyze the polymorphic information content (PIC) (Kalinowski et al., 2007). GenAlEx 6.503 software was used to calculate sample sizes (n), the number of alleles (Na), effective numbers of alleles (Ne), numbers of private alleles (Ar), expected heterozygosity (He), observed heterozygosity (Ho), unbiased expected heterozygosity (uHe), Shannon's information index (I), and inbreeding index (F) (Peakall and Smouse, 2006). Bootstrapping analysis (1,000 repeated samplings) was used to evaluate the paired F-statistics values (Fst) and gene flow (Nm) among populations.

Analysis of molecular variance (AMOVA) was performed on the genetic variation of samples using the software Arlequin 3.1 to obtain the variation level difference between and within populations (Excoffier and Lischer, 2010), and the paired genetic distance (GD) was calculated at the same time. GenAlEx 6.503 was used to verify the principal coordinates analysis (PCoA) with GD as a parameter. Mega 5.0 software used the unweighted pair group method with the arithmetical mean (UPGMA)

TABLE 2 | Parameters of 15 pairs of fluorescent microsatellite markers for silver carp.

Loci	Primer sequence (5′–3′)	Repeat unit	Size range (bp)	Tm (°C)	Fluorescence labels
HLJBL164	F:cgcaccaaggacaaacctaa R:cctgcagaagctacgagacc	$(CAG)_6$	170–187	59	F:6-FAM
HLJBL165	F:ttagaggaaacactggatgacc R:tgctgtttctacagagttttgg	$(TGC)_6$	151–163	59	F:6-FAM
HLJBL167	F:ccaccggatagagaaactcg R:ttatggtgcggctcatacag	$(GTC)_6$	151-163	59	F:6-FAM
HLJBL168	F:ggcggaaatgttgactgact R:atttatgggccgtgtctcaa	$(CTG)_6$	179–189	59	F:HEX
HLJBL169	F:cgacgatcagaggagagtcc R:ggcccagaagcattctcttt	$(TCA)_{11}$	159–177	59	F:TAMRA
HLJBL170	F:tgggttcagcctttaaatgaa R:gaggagggccacctaaagac	$(CTG)_8$	146–160	59	F:Cy-3
HLJBL174	F:gtcgacgatcctgagtccat R:ggaatgagatgtgggcctaa	$(GATA)_{10}$	208–253	59	F:6-FAM
HLJBL176	F:atccgacccttaacgctaca R:tcgttcctttccttctgtcc	$(ATCT)_6N\ (ATCC)_7$	138–184	59	F:HEX
HLJBL181	F:tcgacgatctctccctgttt R:cagctgatcgatagacacac	$(TCAT)_9$	118–166	59	F:HEX
HLJBL184	F:ctgctatgctgcaccacact R:ggcatggtttcactgctgta	$(CTAT)_5N\ (TCTG)_5$	213–235	59	F:6-FAM
HLJBL202	F:ttacctggccagagactgct R:acaagcaggcgagagttttg	$(CTG)_6N\ (TGT)_6N\ (TGC)_8$	139–154	59	F:HEX
HLJBL203	F:gcaatcgctcgatacagaca R:gtgctctctgtgaggctgaa	$(GCA)_{11}$	131–146	59	F:HEX
HLJBL216	F:tatgcaggtcagtggaacga R:aacgacgacagaacgacaga	$(TCTA)_7$	189–244	59	F:6-FAM
HLJBL217	F:gggggtacattccacttcaa R:acgatctggccaacgatatg	$(GATA)_{10}$	176–249	59	F:HEX
HLJBL220	F:tcaatccggccatctatcag R:ttgctgccattccataaaga	$(TCTA)_7N\ (TCCA)_6$	196–222	59	F:TAMRA

to construct a phylogenetic tree based on genetic distance (Tamura et al., 2011).

Structure 2.3.4 software was used to divide the population genetic structure (Smouse and Peakall, 1999). Based on the Bayesian model, the number of possible genetic cluster *K* values was set at 1–10, the length of the burn in-period for Markov chain Monte Carlo (MCMC) was set at 50,000 times, and each *K* value was repeated five times. The method of Evanno was used to calculate Delta *K*, and the most optimal *K* value was the number of clusters for the population (Evanno et al., 2005). Through the repeated sampling analysis of Clumpp 1.1.2 (Jakobsson and Rosenberg, 2007), a diagram was drawn by the software GraphPad Prism 8.0.2 to illustrate the population's genetic structure (Swift, 1997).

RESULTS

Genetic Diversity

Fifteen microsatellites fluorescently labeled PCR products were genotyped by capillary electrophoresis. Cervus 3.0.7 analysis showed that the PIC (polymorphism information content) ranged from 0.380 (ZJ-HLJBL167) ∼ 0.919 (ZJ-HLJBL174). Every locus showed high polymorphism (PIC ≥ 0.5) in one or more silver carp populations, which indicated that the 15 SSR loci are suitable for the evaluation of genetic diversity of silver carp.

The genetic diversity results for different silver carp populations were based on the GenAlEx 6.503 analysis (**Table 3**). Na in every locus of each population ranged from 3 (HK, HLJBL167) to 22 (ZJ, HLJBL217). Ne ranged from 1.630 (HK, HLJBL169) to 10.100 (SZ, HLJBL217). Ar was the highest at the HLJBL165 loci in the SZ population (12). There were more private alleles in the SZ population (21), and fewer private alleles in the HK (8) and ZJ (3) populations, which indicated that SZ population has the highest gene abundance. The mean He of all populations was higher than the mean of Ho, which indicated that the proportion of homozygotes was larger than that of heterozygotes at the average level of these 15 loci. The mean value of uHe (i.e., gene diversity index) ranged from 0.697 (ZJ) to 0.727 (RG), which indicated that there was a similar level to the genetic diversity of all populations. The average value range of I was 1.559 (HK) to 1.668 (SZ). The mean of F was higher than zero and ranged from 0.073 (ZJ) to 0.169 (RG).

Genetic Differentiation

The results of AMOVA analysis showed that genetic variation mainly existed among different individuals within the population (94.69%), and a small part was contributed from different populations (5.31%) (**Table 4**). The pair-wise matrix of Fst and

TABLE 3 | Genetic diversity of silver carp in 5 sections of the Yangtze River.

Loci	Parameters	SZ	HK	AQ	ZJ	RG
HLJBL164	n	91	56	71	51	63
	Na	6	8	5	5	7
	Ne	2.033	3.268	2.421	2.113	4.573
	Ar	0	1	0	0	2
	Ho	0.473	0.661	0.521	0.529	0.746
	He	0.508	0.694	0.587	0.527	0.781
	uHe	0.511	0.700	0.591	0.532	0.788
	I	1.038	1.508	1.136	0.980	1.668
	F	0.070	0.048	0.112	−0.005	0.045
	PIC	0.473	0.735	0.541	0.471	0.749
HLJBL165	n	88	60	71	51	63
	Na	17	7	6	5	5
	Ne	6.639	3.828	2.692	2.677	2.879
	Ar	12	1	1	0	0
	Ho	0.750	0.667	0.606	0.569	0.460
	He	0.849	0.739	0.629	0.626	0.653
	uHe	0.854	0.745	0.633	0.633	0.658
	I	2.236	1.520	1.253	1.208	1.178
	F	0.117	0.098	0.036	0.092	0.295
	PIC	0.844	0.699	0.584	0.576	0.587
HLJBL167	n	91	60	71	51	62
	Na	5	3	5	4	4
	Ne	2.027	2.030	1.986	1.861	2.441
	Ar	0	0	0	1	0
	Ho	0.527	0.450	0.535	0.549	0.548
	He	0.507	0.507	0.497	0.463	0.590
	uHe	0.510	0.512	0.500	0.467	0.595
	I	0.899	0.757	0.841	0.755	1.000
	F	−0.041	0.113	−0.078	−0.187	0.071
	PIC	0.436	0.394	0.414	0.380	0.522
HLJBL168	n	91	58	68	51	62
	Na	5	7	4	4	4
	Ne	3.105	3.998	2.978	2.898	2.913
	Ar	0	2	0	0	0
	Ho	0.549	0.724	0.647	0.647	0.613
	He	0.668	0.750	0.664	0.655	0.657
	uHe	0.672	0.756	0.669	0.661	0.662
	I	1.198	1.536	1.191	1.150	1.189
	F	0.178	0.034	0.026	0.012	0.067
	PIC	0.602	0.730	0.637	0.588	0.613
HLJBL169	n	91	53	70	50	58
	Na	7	6	7	5	8
	Ne	1.902	1.630	2.193	2.491	1.664
	Ar	0	0	0	0	1
	Ho	0.396	0.189	0.571	0.540	0.224
	He	0.474	0.386	0.544	0.599	0.399
	uHe	0.477	0.390	0.548	0.605	0.403
	I	1.028	0.851	1.166	1.156	0.929
	F	0.166	0.512	−0.050	0.098	0.438
	PIC	0.451	0.484	0.529	0.571	0.468
HLJBL170	n	87	57	71	51	63
	Na	5	5	8	6	6
	Ne	2.585	3.466	3.279	2.591	3.010
	Ar	0	0	2	0	0
	Ho	0.460	0.509	0.577	0.608	0.683
	He	0.613	0.711	0.695	0.614	0.668
	uHe	0.617	0.718	0.700	0.620	0.673
	I	1.127	1.376	1.478	1.121	1.263
	F	0.250	0.285	0.169	0.010	−0.022
	PIC	0.594	0.696	0.662	0.551	0.611
HLJBL174	n	91	60	66	51	63
	Na	18	15	20	18	17
	Ne	11.771	8.491	10.809	13.103	11.488
	Ar	0	0	2	0	0
	Ho	0.956	0.900	0.879	1.000	0.889
	He	0.915	0.882	0.907	0.924	0.913
	uHe	0.920	0.890	0.914	0.933	0.920
	I	2.630	2.327	2.667	2.707	2.600
	F	−0.045	−0.020	0.032	−0.083	0.026
	PIC	0.909	0.871	0.910	0.919	0.907
HLJBL176	n	91	59	67	50	63
	Na	17	12	17	14	17
	Ne	9.294	8.846	8.567	7.386	8.909
	Ar	3	1	2	0	2
	Ho	0.890	0.966	0.925	0.860	0.905
	He	0.892	0.887	0.883	0.865	0.888
	uHe	0.897	0.895	0.890	0.873	0.895
	I	2.436	2.270	2.395	2.226	2.463
	F	0.003	−0.089	−0.048	0.005	−0.019
	PIC	0.883	0.880	0.884	0.857	0.879
HLJBL181	n	85	55	66	47	52
	Na	10	9	11	10	15
	Ne	6.855	5.926	6.188	6.077	7.941
	Ar	0	1	0	0	3
	Ho	0.471	0.691	0.576	0.362	0.596
	He	0.854	0.831	0.838	0.835	0.874
	uHe	0.859	0.839	0.845	0.844	0.883
	I	2.057	1.929	1.991	1.959	2.268
	F	0.449	0.169	0.313	0.567	0.318
	PIC	0.855	0.834	0.839	0.838	0.873
HLJBL184	n	91	57	71	51	63
	Na	9	7	10	9	9
	Ne	2.424	2.632	2.986	2.840	2.806
	Ar	2	0	1	1	1
	Ho	0.604	0.649	0.620	0.686	0.683
	He	0.587	0.620	0.665	0.648	0.644
	uHe	0.591	0.626	0.670	0.654	0.649
	I	1.340	1.367	1.566	1.435	1.458
	F	−0.029	−0.047	0.068	−0.059	−0.060
	PIC	0.566	0.633	0.644	0.616	0.618
HLJBL202	n	91	59	69	51	63
	Na	7	7	7	6	8
	Ne	2.042	2.505	2.003	1.787	1.904
	Ar	1	0	0	0	1

(Continued)

TABLE 3 | (Continued)

Loci	Parameters	SZ	HK	AQ	ZJ	RG
	Ho	0.516	0.678	0.478	0.490	0.460
	He	0.510	0.601	0.501	0.440	0.475
	uHe	0.513	0.606	0.504	0.445	0.479
	I	1.063	1.304	1.030	0.951	1.024
	F	−0.012	−0.128	0.045	−0.113	0.031
	PIC	0.480	0.588	0.490	0.420	0.448
HLJBL203	n	91	56	71	51	63
	Na	9	7	11	8	7
	Ne	3.458	4.681	4.254	3.753	4.447
	Ar	0	0	3	0	0
	Ho	0.725	0.750	0.592	0.608	0.714
	He	0.711	0.786	0.765	0.734	0.775
	uHe	0.715	0.793	0.770	0.741	0.781
	I	1.504	1.662	1.731	1.529	1.643
	F	−0.020	0.046	0.227	0.171	0.078
	PIC	0.669	0.783	0.734	0.691	0.743
HLJBL216	n	90	56	71	50	56
	Na	12	9	12	9	14
	Ne	4.820	3.329	3.468	3.477	5.473
	Ar	2	2	1	0	3
	Ho	0.633	0.393	0.620	0.520	0.386
	He	0.793	0.700	0.712	0.712	0.817
	uHe	0.797	0.706	0.717	0.720	0.825
	I	1.852	1.507	1.633	1.583	2.050
	F	0.201	0.438	0.129	0.270	0.650
	PIC	0.771	0.699	0.678	0.691	0.827
HLJBL217	n	90	59	70	50	63
	Na	19	12	20	22	17
	Ne	10.100	5.219	8.551	9.766	7.722
	Ar	1	0	2	1	1
	Ho	0.867	0.576	0.786	0.780	0.587
	He	0.901	0.808	0.883	0.898	0.870
	uHe	0.906	0.815	0.889	0.907	0.877
	I	2.606	1.927	2.548	2.626	2.296
	F	0.038	0.287	0.110	0.131	0.325
	PIC	0.896	0.797	0.876	0.896	0.857
HLJBL220	n	90	60	71	51	57
	Na	11	9	11	11	12
	Ne	6.120	3.803	5.210	5.493	5.447
	Ar	0	0	0	0	0
	Ho	0.811	0.550	0.761	0.667	0.579
	He	0.837	0.737	0.808	0.818	0.816
	uHe	0.841	0.743	0.814	0.826	0.824
	I	2.012	1.545	1.881	1.982	1.940
	F	0.030	0.254	0.059	0.185	0.291
	PIC	0.822	0.704	0.784	0.801	0.823
Mean	n	89.933	57.667	69.600	50.467	60.933
	Na	10.467	8.200	10.267	9.067	10.000
	Ne	5.006	4.243	4.506	4.554	4.908
	Ar (Total)	21	8	14	3	14
	Ho	0.642	0.623	0.646	0.628	0.598
	He	0.708	0.709	0.705	0.690	0.721
	uHe	0.721	0.716	0.710	0.697	0.727
	I	1.668	1.559	1.634	1.558	1.664
	F	0.090	0.133	0.077	0.073	0.169
	PIC	0.683	0.702	0.680	0.658	0.702

(Continued)

N, number of effective analysis samples; Na, number of alleles; Ne, number of effective alleles; Ar, number of private alleles; Ho, observed heterozygosity; He, expected heterozygosity; uHe, unbiased expected heterozygosity; I, Shannon's information index; F, fixation index; PIC, polymorphic information content.

TABLE 4 | Molecular variance (AMOVA) for 5 silver carp populations.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation
Among populations	4	182.017	0.305	5.31%
Within populations	667	3541.474	5.436	94.69%
Total variation	671	3723.491	5.741	100%

TABLE 5 | F-statistics values (Fst) (below diagonal) and gene flow (Nm) (above diagonal) matrix of 5 silver carp populations.

	SZ	HK	AQ	ZJ	RG
SZ	–	6.365	15.850	7.176	5.334
HK	0.039	–	7.104	11.676	11.105
AQ	0.016	0.034	–	8.842	6.365
ZJ	0.034	0.021	0.027	–	9.007
RG	0.045	0.022	0.038	0.027	–

TABLE 6 | Genetic distance (GD) (below diagonal) of 5 silver carp populations.

	SZ	HK	AQ	ZJ	RG
SZ	–				
HK	0.071	–			
AQ	0.029	0.058	–		
ZJ	0.063	0.030	0.047	–	
RG	0.087	0.038	0.069	0.046	–

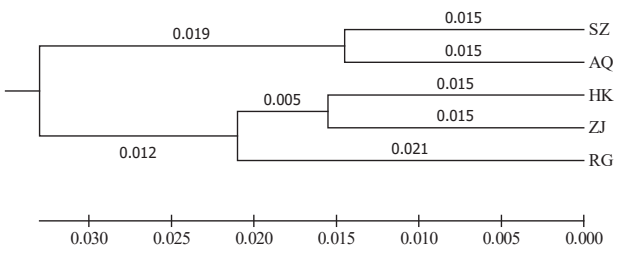


FIGURE 2 | UPGMA tree for five silver carp populations based on genetic distance.

Nm showed that the Fst ranged from 0.016 (SZ and AQ) to 0.045 (SZ and RG), Nm was 6.365 (SZ and HK) to 15.850 (SZ and AQ) (**Table 5**). The genetic distance between populations ranged from 0.029 (SZ and AQ) to 0.087 (SZ and RG) (**Table 6**). Among them, the genetic distance between SZ and HK, SZ and ZJ, SZ and RG, HK and AQ, and AQ and RG was

greater than 0.05. The genetic distance between the SZ and RG population was the largest (0.087), which was consistent with the geographical distance of the two populations. Based on the genetic distance, the population phylogenetic tree (**Figure 2**) showed that the SZ and AQ populations were first clustered, the HK and ZJ populations were also clustered, and then RG joined them together. The genetic distance between individuals was analyzed by PCoA (**Figure 3**). When the four sections were divided by the abscissa and ordinate as different habitats, HK, ZJ, and RG formed one group of habitats, and SZ and AQ formed another group of habitats. It was observed that some groups of AQ integrated into the first environment, and a small portion of ZJ individuals entered the second group of environments.

Genetic Structure

The most optimal *K* value was 2 when Delta *K* was the largest (**Figure 4**), indicating that 335 individuals can be divided into two potential groups (**Figure 5**). The results show that cluster 1 of the SZ population accounts for more than 90%. The HK and RG populations are nearly composed of cluster 2, which indicates that one potential group occupies most of the components in SZ and a part of AQ and ZJ. The other potential group occupies most of the components in HK and RG, and a part of AQ and ZJ.

DISCUSSION

Genetic Diversity

Genetic diversity is the basis for long-term survival of a species as it adapts to the environment and maintains evolution. It provides an important basis for population resource assessment (Yuan et al., 2017). Heterozygosity of 0.500–0.800 indicates that the genetic diversity of this population was high (Takezaki and Nei, 1996). From the uHe of these five populations, the genetic diversity of all populations was at a middle to high level. From the comparison of Ar and I, the genetic diversity level of the SZ population was the highest, while the ZJ population was the lowest. Compared with the research of five silver carp populations (Shishou, Jianli, Jiujiang, Xiangjiang, and Anqing) in the middle and lower Yangtze River (Zhu et al., 2007) and two populations (Wanzhou and Jianli) in the middle and upper Yangtze River (Wang et al., 2008), it was found that the genetic diversity of silver carp in the middle and upper Yangtze River was generally higher than that in the lower reaches, which was consistent with the results of this study. This is due to the sharp decrease in silver carp biomass caused by overfishing. The decrease in biomass is closely related to the decline of genetic diversity (Xuan et al., 2021).

Na and Ne reflect the difference in population genetic variation. The more evenly alleles distributed in the population, the closer the Ne value to Na (Sun et al., 2014). In this study, alleles distributed in the population unevenly. The value of Ho and He reflect the excess or deletion of heterozygotes in the population (Li et al., 2006). From the average heterozygosity of the five populations, He was higher than Ho, which indicated that the entire population was in a state of heterozygous deletion. The F value of most loci ($F > 0$) in each population indicated that there was no bottleneck effect in the silver carp population of the Yangtze River. From the mean value of the population inbreeding coefficient, the HK and RG populations have encountered the risk of inbreeding inhibition ($F > 0.1$) (Moss et al., 2007).

Large numbers of young silver carp have been introduced into the Yangtze River every year to supplement the nature resources, which has a significant negative impact on the natural population genetics (Araki and Schmid, 2010). Genetic studies on the silver carp in four sections of the Yangtze River in Jiangsu also showed that large-scale stock enhancement would increase the gene flow among populations in the Yangtze River and increase the risk of inbreeding (Fang et al., 2021). It has been proved that stock enhancement in the lower Yangtze River can result in heterozygous deletion (Feng et al., 2020), which bring potential genetic risk and lead to a decrease in desirable traits.

Genetic Differentiation and Genetic Structure

In this study, Nm among the five populations was higher than 4 (6.365–15.850), which indicated that gene exchange can be carried out among the five populations. The higher the value, the greater the degree of gene exchange, which prevented the generation of genetic differentiation to a certain extent. Moreover, Fst among 5 populations (0.016–0.045) less than 0.15 indicates low genetic differentiation ($0.05 < Fst < 0.15$, low genetic differentiation, $0.15 < Fst < 0.25$, moderate genetic differentiation) (Weight, 1978). The influence of genetic drift can be ignored, and each population can mate randomly (Hu et al., 2020), that is, the variation mainly came from within the population, which was consistent with the results of the AMOVA analysis. The high Nm was also obtained in a 2016 study among the three populations of the Ganjiang River, Poyang Lake, and the Yangtze River. This occurred because the three sampling points were close to each other, and the silver carp populations in the Ganjiang River and Yangtze River exchanged genes through Poyang Lake (Yu et al., 2016). However, the distance between the five populations in the current study was far, and therefore, it could be speculated that stock enhancement perhaps increased the gene exchange between populations in the Yangtze River. The genetic distance of fish at the three levels of genus, species, and population are 0.90, 0.30, and 0.05 respectively (Shaklee et al., 1982). There was population differentiation in SZ and HK, SZ and ZJ, SZ and RG, HK and AQ, and AQ and RG. The genetic distance between the SZ and RG populations was the largest, and it was positively correlated with the distance of the geographical location.

The results of the UPGMA phylogenetic tree and PCoA analysis divided the five silver carp populations into two groups. Among them, SZ and AQ were grouped together, HK and ZJ were grouped together, and RG was alone. It has been proven that the SZ and HK sections are the spawning grounds of silver carp (Xu et al., 2017; He et al., 2021). In the process of local fertilized eggs drifting downstream with the water, they develop into larvae and juveniles (Ren et al., 2016). When the juvenile fish have the ability to independently swim, they also have the ability to select adaptive habitats for feeding.

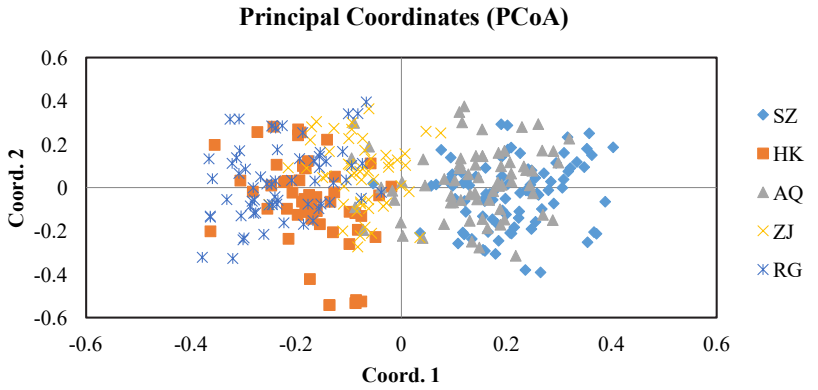


FIGURE 3 | Principal coordinate analysis (PCoA) for 335 individuals.

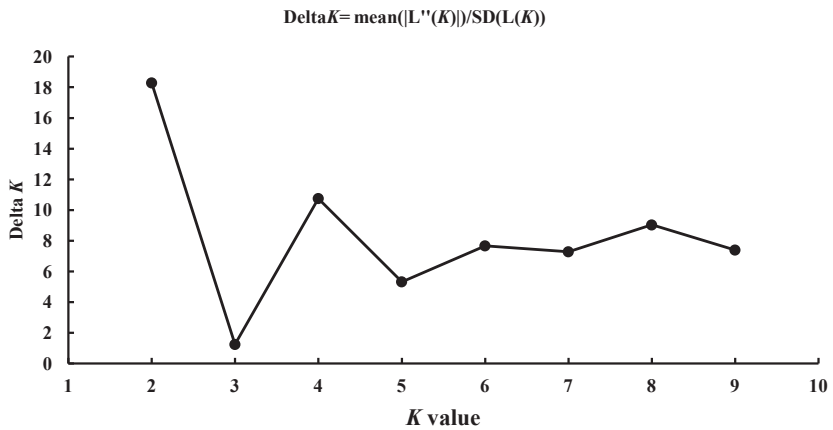


FIGURE 4 | The curve of change of Delta *K* with the changing *K* value.

Furthermore, the geographical distance between the SZ and AQ sections is similar to that from HK to ZJ (both approximately 600 km). Some supplementary populations in the AQ and ZJ sections were probably derived from the offspring of spawning populations in SZ and HK, respectively, which then locally formed corresponding populations.

HK is the dividing section between the middle and lower reaches. There was a single genetic cluster for the HK section in the genetic structure diagram. Poyang Lake is the largest lake connected to the Yangtze River. There were complex ecological conditions and abandoned bait resources (Wang et al., 2016). Silver carp likely entered Poyang Lake through the HK section for fattening after spawning. Some larvae and juveniles that developed from eggs may subsequently complete the life history of feeding and rearing in the lake until they return to the main stream of the Yangtze River for reproduction and spawning. In this case, the HK population possibly formed a distinct population composition that is different from the downstream populations. Another study that investigated the fish assemblage structure of Chinese carp in the Yangtze River indicated that the larvae and juveniles of the four major Chinese carp will passively enter the lake when the water flows backward and will grow for 3–4 years in the lake (Ru and Liu, 2013). Research on the habitat between coral reefs and mangroves indicated that the connectivity promoted fish abundance. The study also

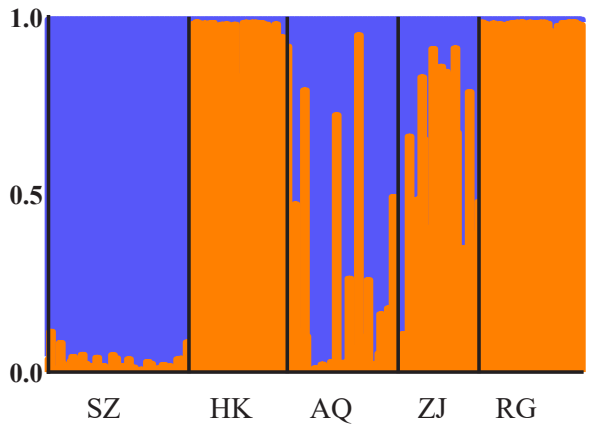


FIGURE 5 | Structural diagram of 5 populations of silver carp (*K* = 2).

recommended that connected habitats should be considered as a high priority for conservation (Olds et al., 2012). Therefore, only by protecting the habitat and fishery resources of the HK section can we ensure the natural connectivity of the river and lake and thus successfully enable the silver carp to complete its reproduction and fattening process.

The AQ section was also one of the most important fishery waters and the key habitat for aquatic animals in the lower reaches of the Yangtze River (Tian et al., 2020). The silver carp migrate to the lower reaches of the Yangtze River for overwintering after finishing the reproductive migration in the middle reaches. Therefore, silver carp populations in AQ and ZJ contain more genetic clusters. The RG population showed the same single genetic structure in the structural diagram, while the UPGMA tree showed that it was clustered into one single class, which indicated that silver carp populations in this section had a further genetic relationship with other populations. The RG section is located at the estuary of the Yangtze River and far from the spawning ground in the middle and upper reaches. Besides, the silver carp in Rugao section experienced overfishing in the process during upward migration, so it is difficult to form a surplus group of multiple reproduction. Similar studies have been found in the biological investigation of *Coilia nasus* in the Yangtze River (Luo et al., 2021). Overfishing made it difficult for large-scale *Coilia nasus* to reach the Anqing section of the Yangtze River. Most of the silver carp in this section came from stock enhancement, and thus, the genetic structure was more singular than that of other river sections in the lower reaches.

Habitat Relevance and Conservation Strategies

Experiments on fish and other animals showed that local residents readily become immigrants when the resources required by a habitat are limited. Similarly, when the resources are sufficient, immigrant individuals become residents (Matter et al., 1989; Nelson et al., 2002). The HK section is located between the SZ and AQ section, but there is closer genetic distance between SZ and AQ with more frequent gene exchange, indicating that when migratory silver carp pass the HK section with rich bait resources after spawning and breeding in the middle reaches, some silver carp populations enter the lake for fattening and become residents of local Poyang Lake (Liu et al., 2019). Although seasonal migrations are easily confused with real migration, we have largely avoided this error through years of sampling (McMahon and Matter, 2006). Therefore, the populations preferred to select the RLC habitats until sexual maturity, and then formed a local geographical population. Research on Dongting Lake also showed that the larvae and juveniles of the four major Chinese carp will grow and mature in the lake for 3–4 years (Ru and Liu, 2013).

For the silver carp populations with distinct genetic structure and inbreeding inhibition such as HK and RG, excellent silver carp larvae and juveniles should be introduced from different sections of the Yangtze River to avoid sib or half-sib mating, and reduce inbreeding and avoid heterozygote loss. In addition, in order to increase the genetic diversity of each section and maintain the ecological balance of fish species reproduction, individuals from the RG section should be encouraged to enter the HK section. In this case, fish between these two populations can exchange their migratory habits. For instance, fish from the RG section may lead fish in the HK section to migrate downward, so as to increase the information exchange of the silver carp population in the Yangtze River and maintain the overall ecological balance. Populations with additional genetic structures can better adapt to a changing environment (Fang et al., 2021). For AQ, ZJ, and other river sections with diversified genetic structures, the genetic balance in the population is in a dynamic process of regulation. The parents of released fish should be derived from the natural population in the corresponding river basin to avoid potential genetic risk caused by gene mixing (Sha et al., 2021). Furthermore, maintaining distinctive habitats of the silver carp is vital importance to protect their species diversity radically.

CONCLUSION AND FUTURE PROSPECTS

Based on 15 microsatellite loci, the genetic diversity, and genetic structure of five silver carp populations in the middle and lower Yangtze River were analyzed. The results showed that the genetic diversity level in the middle was higher than that of the lower reaches. There are potential genetic risks in distinct geographic populations. The RLC habitat characteristics perhaps have greatly contributed to forming local residents and the geographical population. Future stock enhancement of silver carp from different river sections should be considered genetic structure analysis. Furthermore, more valuable molecular markers should be developed to continuously evaluate the population dynamics of the silver carp in the Yangtze River as well as fishery resources assessment.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The animal study was reviewed and approved by the Regulations for the Administration of Affairs Concerning Experimental Animals.

AUTHOR CONTRIBUTIONS

Y-TL and Y-FZ conceived and designed the experiments. Y-TL and D-AF were responsible for data scoring and analyses, and

writing the manuscript. D-PX, YY, Y-XP, and JX helped selecting the samples. C-CM, X-MT, and M-YZ helped DNA extraction and data analysis during manuscript preparation. All authors have read and approved the final manuscript.

FUNDING

This study was supported by the National Key Research and Development Program of China (No. 2018YFD0900905), the National Infrastructure of Fishery Germplasm Resources (No. 2021DKA3047003), and Central Public-interest Scientific Institution Basal Research Fund, CAFS (No. 2020TD61).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.850183/full#supplementary-material

Supplementary Table S1 | Microsatellite loci data for the 335 silver carps.

REFERENCES

Araki, H., and Schmid, C. (2010). Is hatchery stocking a help or harm? : evidence, limitations and future directions in ecological and genetic surveys. *Aquaculture* 308, S2–S11.

Bell, J. D., Bartley, D. M., Lorenzen, K., and Loneragan, N. R. (2006). Restocking and stock enhancement of coastal fisheries: potential, problems and progress. *Fisheries Res.* 80, 1–8. doi: 10.1016/j.fishres.2006.03.008

Bonte, D., and Maelfait, L. (2004). Lack of homeward orientation and increased mobility result in high emigration rates from low-quality fragments in a dune wolf spider. *J. Animal Ecol.* 73, 643–650. doi: 10.1111/j.0021-8790.2004.00838.x

Chen, H. J., Liu, M. D., Wang, D. Q., Dong, W. W., Lv, H., Chen, D. Q., et al. (2018). Analysis of genetic diversity for 4 *Hypophthalmichthys molitrix* populations from middle and upper Yangtze River. *Freshwater Fish.* 48, 20–25.

Cheng, L., Liu, L., Yu, X., and Tong, J. (2008). Sixteen polymorphic microsatellites in bighead carp (*Aristichthys nobilis*) and cross-amplification in silver carp (*Hypophthalmichthys molitrix*). *Mol. Ecol. Resources* 8, 656–658. doi: 10.1111/j.1471-8286.2007.02037.x

Conover, G., Simmons, R., and Whalen, N. (2007). *Management and Control Plan for Bighead, Black, Grass, and Silver Carps in the United States*. Washington, D.C: Aquatic Nuisance Species Task Force, 29–30.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resources* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Fang, D. A., Luo, Y. T., Xu, D. P., Yang, X. W., and Wang, X. H. (2021). Relationship between genetic risk and stock enhancement of the silver carp (*Hypophthalmichthys molitrix*) in the Yangtze River. *Fisheries Res.* 235, 1–9.

Feng, X. T., Zhang, G. N., Xue, X. P., Wang, X. Y., Zhou, Y. F., Fang, D. A., et al. (2020). Current germplasm situation of bighead carp (*Aristichthys nobilis*) candidate parent and parent from hatchery in the lower reaches of Changjiang River based on SSR markers. *J. Fishery Sci. China* 27, 589–597.

He, X. H., Tan, L. F., Peng, Y. X., Fang, D.-A., and Xu, D. P. (2021). Four major Chinese carps eggs resources and spawning grounds distribution at Hukou section of the Yangtze River. *J. Fishery Sci. China* 28, 420–430.

Hu, Y. T., Jiang, H., Duan, G. Q., Zhou, H. X., Ling, J., and Wang, H. (2020). Genetic diversity analysis of Pelteobagrusfulvidraco from two major drainage systems in Anhui Province based on microsatellite markers. *South China Fisheries Sci.* 16, 33–41.

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233

Kalinowski, S. T., Taper, M. L., and Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol. Ecol.* 16, 1099–1106. doi: 10.1111/j.1365-294X.2007.03089.x

Kolar, C. S., and Lodge, D. M. (2002). Ecological predictions and risk assessment for alien fishes in North America. *Science* 298, 1233–1236. doi: 10.1126/science.1075753

Li, H. B., and Xu, D. P. (2008). Analysis on the change characteristics and causes of "four major Chinese carp" resources in Dongting Lake. *J. Inland Fisheries* 33, 34–36.

Li, P. F., Liu, P., Liu, X. Z., Gao, T., and Wang, Q. Y. (2006). Isozyme polymorphism in paralichthys, *Paralichthys lethostigma*. *Fishery Sci. China* 13, 13–19.

Li, S. F. (1996). *Germplasm Resources and Protection of Freshwater Fish in China*. Beijing: Freshwater fish germplasm resources and conservation in China.

Li, S. Z., and Fang, F. (1990). On the geographical distribution of the four kinds of pond-cultured carps in China. *Acta ZoologicaSinica* 36, 244–250.

Lindenmayer, D., Hobbs, R. J., Montague-Drake, R., and Alexandra, J. (2008). A checklist for ecological management of landscapes for conservation. *Ecol. Lett.* 11, 78–91. doi: 10.1111/j.1461-0248.2007.01114.x

Liu, S. P., Duan, X. B., Chen, D. Q., and Liao, M. J. (2005). Studies on status of fishery resources in the Yangtze River. *Acta Hydro Biol. Sinica* 29, 708–711.

Liu, S. P., Qiu, S. L., Chen, D. Q., and Huang, M. J. (1997). Protection andrational utilization of the germplasm resources of the four majorChinese carps in the Yangtze River system. *Resources Environ. Yangtze* 6, 127–131.

Liu, Y., Yang, X. W., Ren, P., Li, X. F., Fang, D. A., and Xu, D. P. (2019). Community characteristics of larvae and juvenile fish inHukou section of the Yangtze river in spring and summer. *Acta Hydrobiol. Sinica* 43, 142–154.

Lu, G. Q., Wang, C. H., Zhao, J. L., Liao, X. L., Wang, J., Luo, M. K., et al. (2020). Evolution and genetics of bighead and silver carps: native population conservation versus invasive species control. *Evol. Appl.* 13, 1351–1362. doi: 10.1111/eva.12982

Luo, Y. T., Dai, P., Liu, S. L., Ma, F. J., You, Y., and Liu, K. (2021). Age structure and growth characteristics of coilia nasus in anqing section of yangtze river during fishing season in 2018. *J. Guangdong Ocean University* 41, 1–8.

Mai, L. F. (2003). The fishery resources of the Yangtze River have declined seriously. *South China Fisheries Sci.* 5:4.

Mao, R. X., Zhang, Y. B., Zheng, W., Du, X. Y., and Sun, X. W. (2010). The progress of germplasm resources of four major chinese farmed carps. *Chinese J. Fish.* 23, 52–59.

Matter, W. J., Mannan, R. W., Bianchi, E. W., Mcmahon, T. E., Menke, J. H., and Tash, J. C. (1989). A laboratory approach for studying emigration. *Ecology* 70, 1543–1546. doi: 10.2307/1938212

McMahon, T. E., and Matter, W. J. (2006). Linking habitat selection, emigration and population dynamics of freshwater fishes: a synthesis of ideas and approaches. *Ecol. Freshwater Fish* 15, 200–210.

Moss, D. R., Arce, S. M., Otoshi, C. A., Doyle, R. W., and Moss, S. M. (2007). Effects of inbreeding on survival and growth of Pacific white shrimp Penaeus (*Litopenaeus vannamei*. *Aquaculture* 272, S30–S37.

Nelson, A. R., Johnson, C. L., Matter, W. J., and Mannan, R. W. (2002). Tests of emigration in small mammals under experimental conditions. *Canadian J. Zool.* 80, 2056–2060.

Ni, Y., and Wu, H. L. (2006). *Fishes of Jiangsu Province*. Beijing: China Agriculture Press.

Olds, A. D., Connolly, R. M., Pitt, K. A., and Maxwell, P. S. (2012). Habitat connectivity improves reserve performance. *Conservation Lett.* 5, 56–63. doi: 10.1111/j.1755-263x.2011.00204.x

Peakall, R., and Smouse, P. E. (2006). GENALEX 6: genetic analysis in Excel. population genetic software for teaching and research. *Mol. Ecol. Notes* 6, 288–295. doi: 10.1093/bioinformatics/bts460

Pinter, K. (1980). Exotic fishes in hungarian waters: their importance in fishery utilization of natural water bodies and fish farming. *Aquaculture Res.* 11, 163–167. doi: 10.1111/j.1365-2109.1980.tb00401.x

Ren, J., Peng, Q. D., Lin, J. Q., and Dong, G. D. (2016). An instantaneous point-source diffusion model for pelagic eggs for estimating early life stage fish resources. *J. Hydrometeorol.* 37, 82–87.

Ru, H. J., and Liu, X. Q. (2013). River-lake migration of fishes in the Dongting Lake area of the Yangtze floodplain. *J. Appl. Ichthyol.* 29, 594–601. doi: 10.1111/jai.12116

Sale, P. F. (1968). An experimental study of habitat selection in the juvenile manini, *Acanthurus triostegus* sandvicensis. *Ecology* 54, 616–623.

Sha, H., Luo, X. Z., Li, Z., Zhou, G. W., and Liang, H. W. (2018). Genetic diversity of six silver carp (*Hypophthalmichthys molitrix*) geographical populations based on mitochondrial CO? sequences. *J. Fishery Sci. China* 25, 783–792. doi: 10.3724/sp.j.1118.2018.18058

Sha, H., Luo, X. Z., Wang, D. Q., Li, X. H., Zou, G. W., and Liang, H. W. (2021). New insights to protection and utilization of silver carp (*Hypophthalmichthys molitrix*) in Yangtze River based on microsatellite analysis. *Fish. Res.* 241, 1–7.

Shaklee, J. B., Tamaru, C. S., and Waples, R. S. (1982). Speciation and evolution of marine fishes studied by the electrophoretic analysis of proteins. *Pacific J. Optimization* 36, 141–157.

Smouse, P. E., and Peakall, R. (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* 82, 561–573. doi: 10.1038/sj.hdy.6885180

Sun, L. Y., Guo, H. Y., Zhu, C. Y., Ma, Z. H., Jiang, S. G., and Zhang, D. C. (2014). Genetic polymorphism of breeding populations of goldenpompano (*Trachinotusovatus*). *South China Fisheries Sci.* 10, 67–71.

Sun, X. W., Zhang, X. F., Zhao, Y. Y., Zhang, Y., Jia, Z. Y., Chang, Y. M., et al. (2008). Development and application of microsatellite markers in aquatic species. *J. Fishery Sci. China* 15, 689–703.

Swift, M. L. (1997). GraphPad prism, data analysis, and scientific graphing. *J. Chem. Inform. Model.* 37, 411–412. doi: 10.1021/ci960402j

Takezaki, N., and Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* 144, 389–399. doi: 10.1093/genetics/144.1.389

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tan, Z. J., Zhang, T. Q., Lu, C. Y., Li, C., and Sun, X. W. (2011). Development and characteristics of tri- and tetra-nucleotide microsatellites for silver carp (*Hypophthalmichthysmolitrix*). *J. Shanghai Ocean University* 20, 328–335.

Tang, X. L., Chen, D. Q., Wang, K., Liu, S. P., Luo, H. W., and Duan, X. B. (2010). Spatial and temporal distribution of larval resources of fishes in the upper reaches of Yangtze River. *Freshwater Aquaculture* 40, 27–31.

Tian, J. L., Dai, P., Ren, P., Guo, W. J., Zhu, Z. B., Li, X. F., et al. (2020). Assemblage of larvae and juvenile fish in AnqingXinzhou section of the Yangtze River. *J. Fishery Sci. China* 27, 916–926.

Wang, C. Z., Liang, H. W., Zhou, G. W., Luo, X. Z., Li, Z., Tian, H., et al. (2008). Genetic variation analysis of two silver carp populations in the middle and upper Yangtze River by microsatellite. *Hereditas(Beijing)* 30, 1341–1348. doi: 10.3724/sp.j.1005.2008.01341

Wang, C., Yu, X., and Tong, J. G. (2007). Microsatellite diversity and population genetic structure of redfin culter (*Culter erythropterus*) in fragmented lakes of the Yangtze River. *Hydrobiologia* 586, 321–329. doi: 10.1007/s10750-007-0702-x

Wang, S., Duan, X. B., Chen, W. J., Zhang, H. X., Fu, P. F., He, G., et al. (2016). Status and changes of fish resources in the Hukou area of Poyang Lake. *Freshwater Fisheries* 46, 50–55.

Weight, S. (1978). *Evolution and the Genetics of Population Variability Within and Among Natural Populations*. Chicago, IL: University of Chicago Press.

Xu, C. S., Ai, Z. Q., and Xiao, M. (2017). A review of influencing factors on natural reproduction of four major Chinese carps in Yangtze River. *J. China Three Gorges University (Natural Sciences)* 39, 27–31.

Xuan, Z. Y., Jiang, T., Liu, H. B., Chen, X., and Yang, J. (2021). Mitochondrial DNA and microsatellite analyses reveal strong genetic differentiation between two types of estuarine tapertail anchovies (Coilia) in Yangtze River Basin, China. *Hydrobiologia* 848, 1409–1431. doi: 10.1007/s10750-021-04541-w

Yu, Y., Pang, M. X., Yu, X. M., Tong, J. G., and Shen, J. Z. (2016). Population genetic structure of silver carp from Yangtze River, Ganjiang River and Poyang Lake based on microsatellite markers. *J. Huazhong Agricultural University* 35, 104–110.

Yuan, J. G., Li, J. Q., Liu, L., Chen, Z. X., and Zhang, Y. P. (2017). Genetic diversity of GIFT strains of oreochromis niloticus based on 18S-ITS1-5.8S sequence. *J. Guangdong Ocean University* 37, 7–10.

Zeng, X. C. (1990). *Fishery Resources of the Yangtze River System*. Beijing: Ocean Publishing House.

Zhu, X. D., Geng, B., Li, J., and Sun, X. W. (2007). Analysis of genetic diversity among silver carp populations in the middle and lower Yangtse River using thirty microsatellite markers. *Hereditas* 29, 705–713. doi: 10.1360/yc-007-0705

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Luo, Fang, Zhou, Xu, Peng, Zhang, Mao, Tang, Xu and You. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Pharmacology

ORIGINAL RESEARCH
published: 01 June 2022
doi: 10.3389/fphar.2022.827215

CLP1 is a Prognosis-Related Biomarker and Correlates With Immune Infiltrates in Rheumatoid Arthritis

Zhenyu Zhao[1†], Shaojie He[1†], Sheng Tang[1†], Xiaofeng Lai[1], Jie Ren[2], XinCheng Yu[1], Jinhua Lin[1], Mohan Wang[3], Mariya M. El Akkawi[4], Shan Zeng[2] and Dingsheng Zha[1]**

[1]Department of Orthopaedics, The First Affiliated Hospital, Jinan University, Guangzhou, China, [2]Department of Rheumatology, The First Affiliated Hospital, Jinan University, Guangzhou, China, [3]School of Basic Medicine and Public Health, Jinan University, Guangzhou, China, [4]Department of Plastic and Reconstructive Surgery, Zhujiang Hospital of Southern Medical University, Guangzhou, China

OPEN ACCESS

Edited by:
William C Cho,
QEH, Hong Kong SAR, China

Reviewed by:
Hang Xue,
Huazhong University of Science and Technology, China
Hui Jiang,
The First Affiliated Hospital of Anhui University of Chinese Medicine, China

***Correspondence:**
Dingsheng Zha
zdsuser@126.com
Shan Zeng
zs870103@126.com

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Pharmacology

Received: *01 December 2021*
Accepted: *21 April 2022*
Published: *01 June 2022*

Citation:
Zhao Z, He S, Tang S, Lai X, Ren J, Yu X, Lin J, Wang M, El Akkawi MM, Zeng S and Zha D (2022) CLP1 is a Prognosis-Related Biomarker and Correlates With Immune Infiltrates in Rheumatoid Arthritis. Front. Pharmacol. 13:827215. doi: 10.3389/fphar.2022.827215

Rheumatoid arthritis (RA) is a chronic, heterogeneous autoimmune disease with a high disability rate that seriously affects society and individuals. However, there is a lack of effective and reliable diagnostic markers and therapeutic targets. In this study, we identified diagnostic markers of RA based on RNA modification and explored its role as well as degree of immune cell infiltration. We used the gene expression profile data of three synovial tissues (GSE55235, GSE55457, GSE77298) from the Gene Expression Omnibus (GEO) database and the gene of 5 RNA modification genes (including m6A, m1A, m5C, APA, A-1), combined with cluster analysis, identified four RNA modifiers closely related to RA (YTHDC1, LRPPRC, NOP2, and CLP1) and five immune cells namely T cell CD8, CD4 memory resting, T cells regulatory (Tregs) Macrophages M0, and Neutrophils. Based on the LASSO regression algorithm, hub genes and immune cell prediction models were established respectively in RA and a nomogram based on the immune cell model was built. Around 4 key RNA modification regulator genes, miRNA-mRNA, mRNA-TF networks have been established, and GSEA-GO, KEGG-GSEA enrichment analysis has been carried out. Finally, CLP1 was established as an effective RA diagnostic marker, and was highly positively correlated with T cells follicular helper (Tfh) infiltration. On the other hand, highly negatively correlated with the expression of mast cells. In short, CLP1 may play a non-negligible role in the onset and development of RA by altering immune cell infiltration, and it is predicted to represent a novel target for RA clinical diagnosis and therapy.

Keywords: rheumatoid arthritis, RNA modification, diagnostic biomarkers, gene expression profile data, CLP1

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic autoimmune disease characterized by inflammation of the connective tissue. Statistics show that the incidence of RA in the population is about 1%, which is a common chronic disease (Firestein and McInnes 2017). RA mainly affects joints and joint synovial cartilage. As the disease progresses, the damage of both will become more severe (Littlejohn and Monrad 2018). Cartilage destruction, bone erosion, and other symptoms will gradually occur, eventually leading to irreversible limb deformity and disability, bringing a serious burden to society

or individuals (Smolen et al., 2016). It is worth noting that there is a huge gap between early and late symptoms of RA. In the early stage, it is usually accompanied by mild systemic symptoms, including fatigue and morning stiffness, while in the late stage, it is often accompanied by severe multi-system autoimmune diseases, such as pulmonary interstitial disease, multiple vasculitis, etc. (Brzustewicz et al., 2017). Therefore, timely and accurate early diagnosis of RA is particularly crucial. In the current clinical management, rheumatoid factor (RF) and anti-citrullinated protein antibody (ACPA) have become important serum markers for the diagnosis of RA. However, both are often negative in early RA, and the diagnostic validity cannot meet clinical needs (Wu C. Y. et al., 2021). Some new biomarkers such as Fibrinogen-Like Protein 1 (FGL1) (Liu et al., 2020), collagen triple helix repeat containing 1 (CTHRC1) (Myngbay et al., 2019), etc. have potential diagnostic markers, but large-scale clinical validation has not yet been carried out, and there is still a long way to go in clinical services. As a result, finding novel biomarkers that can accurately identify the diagnosis and prognosis of RA is critical.

As gene sequencing technology advances, the importance of RNA modification in the onset and progression of many illnesses has been increasingly elucidated. (Djebali et al., 2012). Cancer, cardiovascular disorders, genetic birth abnormalities, metabolic diseases, neurological diseases, and mitochondrial-related defects are all linked to RNA-modifying enzyme mutations. (Jonkhout et al., 2017). N6-methyladenosine (m6A), N1-methyladenosine (m1A), 5-methylcytidine (5mC), and other RNA modifications are common. Previous research has found that certain RNA changes play a vital role in the etiology and development of RA, and that they might be used as a therapeutic target for the disease. (Wu S. et al., 2021), but similar studies are insufficient, and further in-depth exploration is still needed.

In addition, as an autoimmune disease, the infiltration of immune cells is a key part of the pathogenesis and progression of RA. Studies have suggested that the interactions between a variety of immune cells, inflammatory factors, and various cellular chemokines, including T cells, B cells, macrophages, and NK cells, are involved in the pathological process of RA (Weyand and Goronzy 2021). In addition, it has been suggested that the infiltration of immune cells may be related to the modification of RNA (Gao et al., 2020).

tIn this study, We used gene expression profiling data for three synovial tissues (GSE55235, GSE55457, GSE77298) and five RNA-modifying genes (including m6A, m1A, m5C, APA, A-1) in the Gene Expression Omnibus (GEO) database to obtain the original data information. Combined with cluster analysis, four RNA modifiers (YTHDC1, LRPPRC, NOP2, and CLP1) and five immune cells (T cells CD8, CD4 memory resting, T cell regulatory (Tregs) macrophages M0 and medium neutrophils) were closely related to the process of RA. Among them, T cell subsets, macrophages, and mast cells are considered the three most critical immune infiltrating cells in the process of RA. Hub genes/immune cell prediction models in RA based on the LASSO regression algorithm were established respectively. Simultaneously a nomogram based on the immune cell model was established. Then, we perform molecular typing based on the expression patterns of RNA modification regulators and analyzed protein interaction networks to construct miRNA/transcription factor (TF)-mRNA modifier interaction networks. Finally, we consider CLP1 the most potent RA diagnostic marker among the four key RNA modifiers, and immune correlation analysis suggests that it is highly positively correlated with T-cell follicular helper (Tfh) infiltration. CLP1 is positively correlated with the infiltration of immune cells such as B cells naïve, eosinophils, monocytes, dendritic cells activated, plasma cells, macrophages, T cells $CD8^+$ and Tfh, while negatively correlated with the degree of infiltration of mast cells, NK cells activated and T cells CD4 memory.

ARTICLE TYPES

The article types is Original Reasearch.

MATERALS AND METHODS

Data and Differentially Expressed Genes Acquirement

The gene expression profile data of synovial tissues GSE55235 (Woetzel et al., 2014), GSE55457 (Woetzel et al., 2014), and GSE77298 (Broeren et al., 2016) were retrieved using the R package in GEOquery (Davis and Meltzer 2007) by accessing the GEO database (Barrett et al., 2007), and the gene expression groups were merged and split into 39 synovial tissue of rheumatoid arthritis and 27 normal synovial tissues. We preprocessed the downloaded expression matrix, including data background adjustment, normalization, and summarization. The 5 gene sets of RNA modification writer genes (including m6A,m1A,m5C, APA, A-1) were obtained by Chen H et al.(Chen et al., 2021) and Cong P et al.(Cong et al., 2021): In order to analyze the changes in the expression values of 5 RNA modification writer genes in the synovial tissues of rheumatoid arthritis relative to the normal tissues, we further screened the differentially expressed genes in RA and normal synovium. DEG is recognized and integrated by the Limma package (Ritchie et al., 2015) and FunRich software (Pathan et al., 2015) in R, we set the genes with logFC>1 and adjPvalue<0.05 as up-regulated genes. The genes with logFC<1 and adjPvalue<0.05 as down-regulated genes. Visualization of chromosomal localization of RNA modification writer genes was done using the R circos package (An et al., 2015).

Immune Infiltration Analysis

CIBERSORT is a deconvolution technique based on the premise of linear support vector regression for immune cell subtype expression matrices, CIBERSORT was initially used for the analysis of tumor microenvironment (TME) and is now being increasingly applied in the characteristic analysis of immune infiltration in non-tumor tissues (Ge et al., 2021). The synovial tissue of rheumatoid arthritis is composed of various immune and inflammatory cells, interstitial tissue, cytokines, and chemokines, which is an integrated loading system. The infiltration analysis of immune cells has an important guiding role in disease research and prognosis prediction. RNA-Seq data were used to assess the

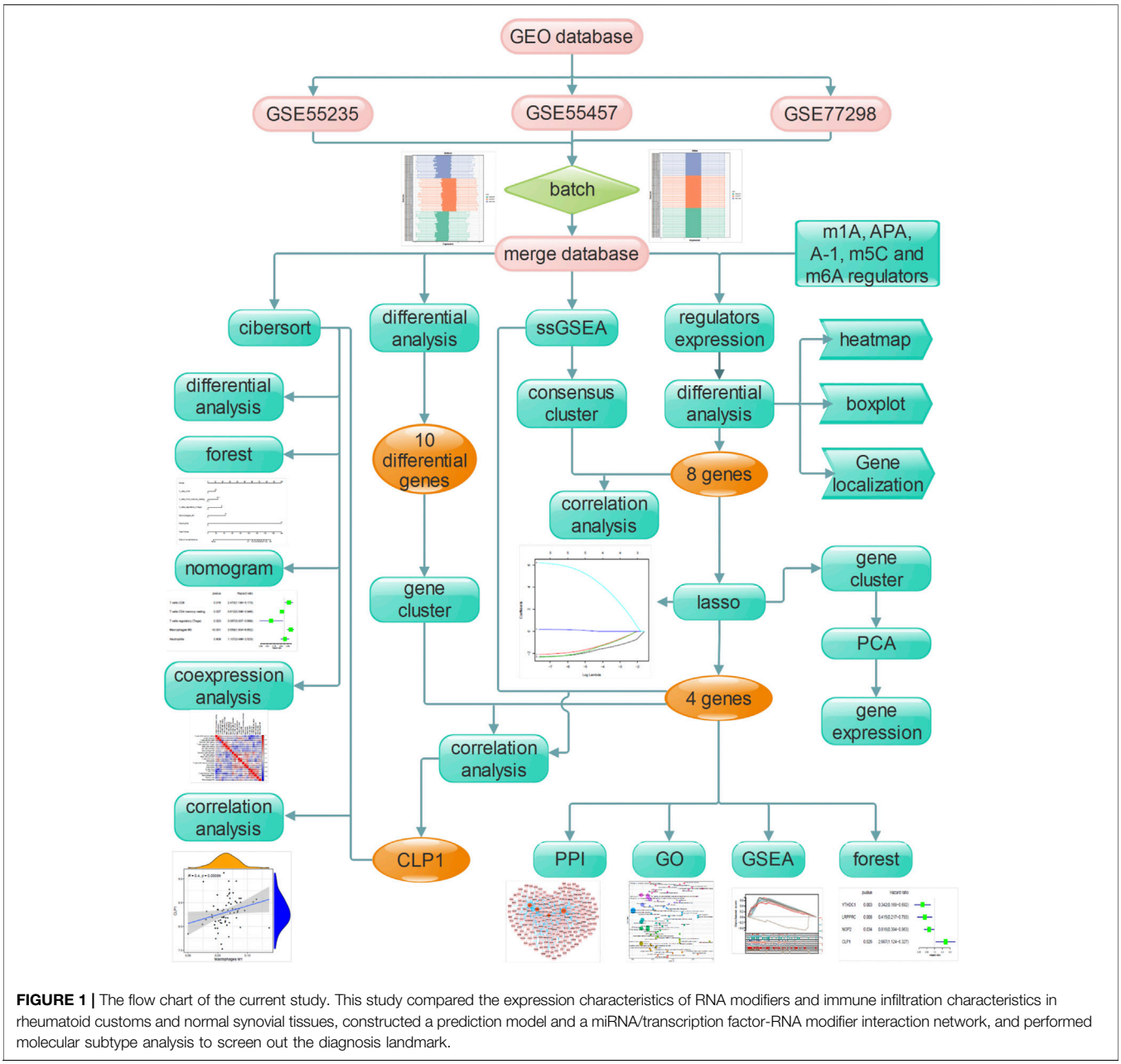


FIGURE 1 | The flow chart of the current study. This study compared the expression characteristics of RNA modifiers and immune infiltration characteristics in rheumatoid customs and normal synovial tissues, constructed a prediction model and a miRNA/transcription factor-RNA modifier interaction network, and performed molecular subtype analysis to screen out the diagnosis landmark.

infiltration of immune cells in synovial tissues of rheumatoid arthritis and normal tissues (Newman et al., 2019). CIBERSORT algorithm was used to analyze the immune infiltration between the rheumatoid and normal tissues, to identify the immune cells that were differentially enriched between the diseased and normal tissues in the two sets of data, to calculate the Pearson correlation coefficient between the expression level of key genes and immune cells, and to evaluate the relationship between the key genes and the immune infiltration level.

Construction of Prediction Model

Immune cells of significantly different infiltration levels and significantly different genes between the rheumatoid and normal synovial tissues were used to construct predictive models. LASSO regression analysis in the R glmnet package (Friedman et al., 2010) was performed on the training set. The LASSO approach can reduce the dimensionality of high-latitude data, allowing a model with fewer variables to explain the data's features (Gui and Li 2005). To avoid overfitting the model design of the training set, tenfold cross validation is performed. Finally, the regression coefficients generated using LASSO regression analysis are used to create a scoring system. We split all RA patients into high-risk and low-risk groups using the R package "SurvMiner" threshold, and then used principal component analysis (PCA) to see how well the model could identify overall survival outcome events.

RNA Modification Factor Molecular Typing

We use the R-packet "ConsensusClusterPlus" (Wilkerson and Hayes 2010) for clustering classification, and the samples were divided into different groups by expression of RNA modification factor. The parameter is set to repeat 50 times (reps = 50) and the resampling rate is 80% (pItem = 0.8). Principal component analysis (PCA) was done on the expression levels of all genes to determine the success of grouping, and the results were shown using the "pheatmap" program.

Construction of Protein-Protein Interaction (PPI) Network

Individual proteins interact with one another in protein interaction networks, which involves various aspects of biological signal transmission, gene expression management, energy and chemical metabolism, and cell cycle control. A systematic analysis of the interaction between a large number of proteins in biological systems is required to understand the working principle of proteins in biological systems, the reaction mechanism of biological signals, and energy metabolism under special physiological states such as diseases, as well as the functional relationships between proteins. The STRING database (Szklarczyk et al., 2019) is a database that searches for interactions between known and predicted proteins. The database containins 9.6 million proteins and 1.38 million protein-protein interactions from 2,031 species. It contains results obtained from experimental data, results mined from PubMed abstract text, and combining results from other databases as well as results predicted using bioinformatics methods. We constructed protein-protein interaction networks for RA prognosis-related differentially expressed genes and glioblastoma-related differentially expressed prognostic genes respectively using the STRING database.

Construction of Hub Gene-miRNA and Hub Gene-Transcription Factor Interaction Networks

In the post-transcriptional stage, the interaction with target genes by miRNA or TF control gene expression under disease-limiting conditions was analyzed (Baldwin 2001; Soifer et al., 2007). We obtained miRNAs and transcription factors of the differentially expressed genes associated with rheumatoid arthritis. The differential expression gene-miRNA network and differential expression prognostic gene-TF network associated with rheumatoid arthritis were visualized using Cytoscape software.

Functional Enrichment Analysis

Gene ontology (GO) (2015) functional annotation analysis, which includes biological process (BP), molecular function (MF), and Cellular Component (CC), is a common method for large-scale gene functional enrichment. The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2017) is a polular database that stores information about genomes, biological pathways, diseases, and medications. The ClusterProfiler package (Yu et al., 2012) was implemented in R for GO functional annotation analysis and KEGG pathway enrichment analysis of differentially expressed glioblastoma genes. In this study, pvalue<0.05 was considered statistically significant.

GSEA Enrichment Analysis

GSEA (Gene Set Enrichment Analysis) was used to evaluate the distribution trend of genes in a predefined Gene Set in the Gene list sorted by phenotypic relevance, and thus to judge its contribution to phenotype (Subramanian et al., 2005). We obtained "C2. kegg.v7.4. symbols" and "c5. go.v7.4. symbols" gene sets from the MSigDB database for GSEA analysis of the two data sets respectively, and used the "ClusterProfiler" R package for GSEA analysis (Yu et al., 2012). Pvalue<0.05 was considered to be statistically significant (Liberzon et al., 2015).

Statistical Analysis

R programming (https://www.r-project.org/, version 4.0.2) was used to execute all data processing and statistical. To compare two continuous variable groups, the statistical significance of the normally distributed variables was evaluated using the independent Student t-test, and the differences between the non-normally distributed variables were analyzed using the Mann-Whitney U test (i.e. Wilcoxon rank-sum test). p <0.05 was considered statistically significant when all statistical p values were bilateral.

RESULTS

Overall Process of Experimental Design

The flow chart designed in this study is shown in **Figure 1**. In short, we compared the expression characteristics of RNA modification factors in rheumatoid arthritis and normal synovial tissues by screening the expression matrix of samples from the GEO database. Then CIBERSORT was used to identify rheumatoid arthritis immune cell infiltration. Next, LASSO was used to identify key genes and immune cells for functional analysis. Molecular typing was performed according to the expression pattern of RNA modification factors, and protein interaction network analysis was performed to construct miRNA/transcription factor-RNA modification factor interaction network. Finally, correlation analysis was conducted between screened diagnostic markers and molecular subtypes and immune cells.

Analysis of Overall Expression Characteristics of RNA Modification Factors in Synovial Tissues of Rheumatoid Joints

To analyze the influence of gene expression values on the synovial tissues of rheumatoid joints relative to normal tissues, we first performed differential gene expression analysis on the integrated gene expression matrix using the limma package (**Figure 2**). 950 differential genes were identified in the comparison between the synovial tissues of rheumatoid joints and normal synovial tissues. There were 427 up-regulated genes and 523 down-regulated genes. We further observed the influence of RNA modification

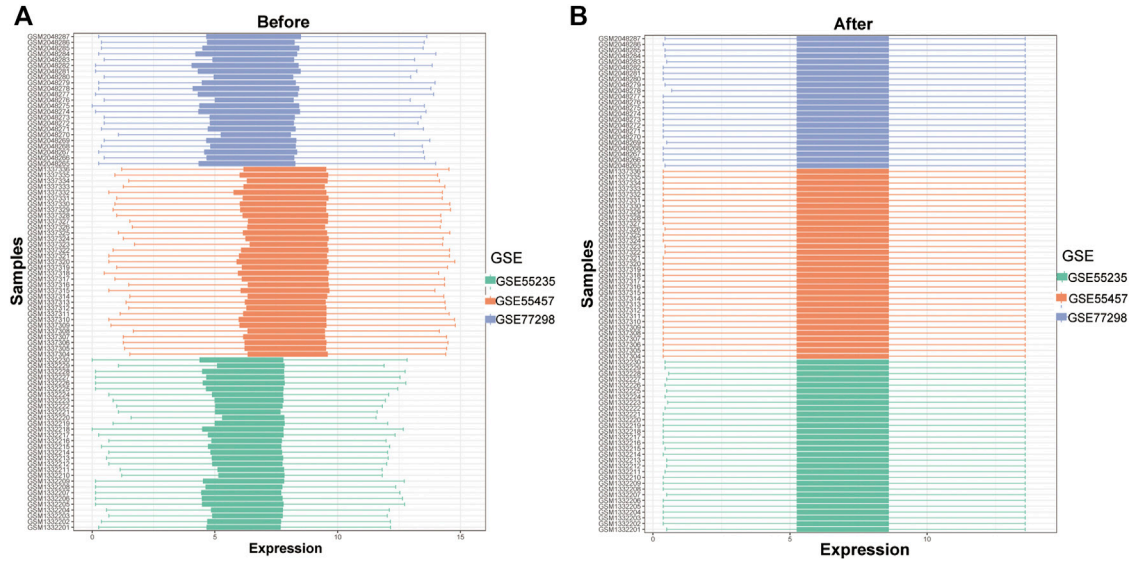


FIGURE 2 | Data integration of typhoon expression matrix. **(A)**: The overall gene expression values of the three GEO data sets before correction, **(B)**: The overall gene expression values of the three GEO data sets after correction.

factors on synovial tissues of rheumatoid arthritis. The heat map showed that the expression of most RNA modification factors was relatively low in rheumatoid joint tissues (**Figure 3A**), and a total of 8 modification factors were significantly different from normal tissues: The expression of CPSF4, IGFBP2, LRPPRC, METTL3, NOP2, TRMT61A, and YTHDC1 was significantly decreased in rheumatoid joints, while the expression of CLP1 was significantly increased (**Figure 3B**). The expression of a significant number of RNA modification factors was correlated, according to the correlation analysis results. For example, YTHDC1 expression was significantly positively correlated with YBX1 and HNRNPC expression, while CPSF4 expression was significantly positively correlated with YTHDF3 expression (**Figure 3C**). Chromosome location showed that LRPPRC and IGFBP2 genes were located on chromosome 2. YTHDC1 is located on chromosome 4, CPSF4 on chromosome 7, CLP1 on chromosome 11, NOP2 on chromosome 12, METTL3, and TRMT61A on chromosome 14 (**Figure 3D**).

Immune Infiltration in Rheumatoid Arthritis

To analyze the difference in the degree of immune infiltration between rheumatoid arthritis synovial tissue and normal tissue, we used the cibersort algorithm to calculate the degree of infiltration of 22 kinds of immune cells in rheumatoid arthritis tissue and normal tissue. Correlation analysis showed that there was a correlation between the degree of cell infiltration. For example, there was a positive correlation between B cells naïve and CD8+T cell infiltration (correlation coefficient was 0.48, **Figure 4A**). Comparing RA samples and normal tissue samples, we found that the average infiltration levels of dendritic cells resting, mast cells, T cells CD4 memory resting, Tregs and other cells in RA tissues were significantly lower than those in the healthy group, while M0/M1 macrophages, plasma cells, CD8 T cells, T cells follicular helper and T cells gamma delta were significantly higher than normal samples (**Figure 4B–W**) using the Wilcox. test algorithm. It is not difficult to see that these differentially infiltrated cells have more T cell subsets, and include three different subsets of macrophages, indicating that there may be a complex interaction between them.

MOLECULAR CLUSTERS OF RNA MODIFICATION FACTORS

To further explore the biological characteristics of the expression of RNA modification factors in different synovial tissues, we used 8 RNA modification factors expression pairs to perform unsupervised consensus clustering of synovial tissues. The optimal separation was obtained by dividing all samples into two different subtypes (A: n = 45; B: n = 21, **Figures 5A–C**). Principal component analysis (PCA) results showed high separation quality (**Figure 5D**). Further differential analysis revealed that the expression of YTHDC1, LRPPRC, NOP2, and CLP1 in various subgroups differed significantly (**Figures 5E–H**). Therefore, we defined these four genes as hub genes.

CONSTRUCTION OF RHEUMATOID ARTHRITIS PREDICTION MODEL

We used the LASSO regression algorithm to construct a prediction model of rheumatoid arthritis based on RNA modification factors and immune cell infiltration levels (**Figures 6A,B**). The results showed that we constructed predictive models for the expression of YTHDC1, LRPPRC, NOP2, and CLP1 respectively and the prediction models of five immune cells: T cells CD8, T cells CD4 memory resting,

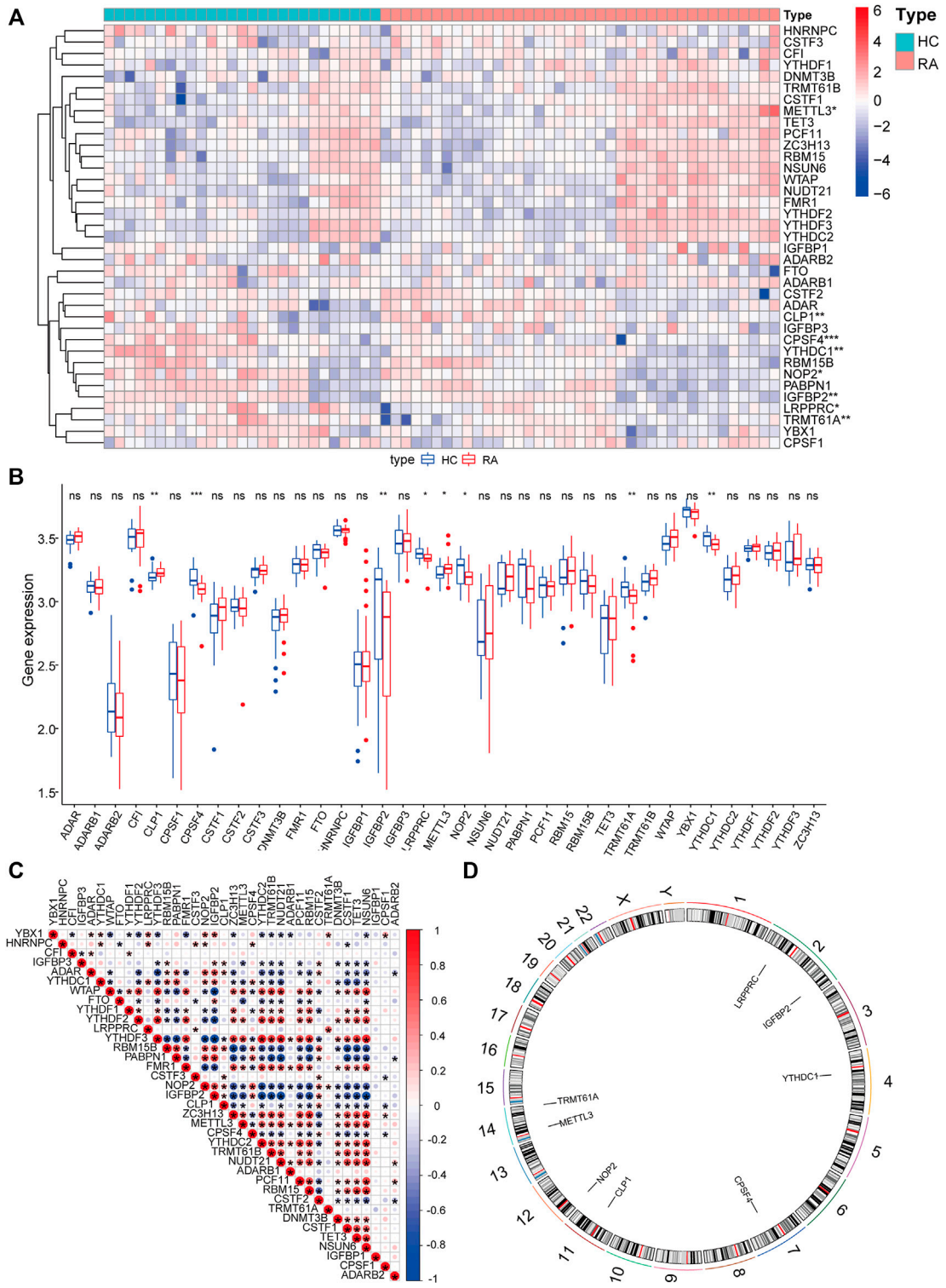


FIGURE 3 | Expression characteristics and gene location of RNA modifiers. **(A)**: Heat map shows the expression characteristics of RNA modifiers in rheumatoid synovial tissues and normal tissues. Red stands for high expression level, and blue for low expression level; **(B)**: box plot shows the difference in the expression of RNA modifiers in synovial tissues and normal tissues, with significant differences in the expression of 8 genes. **(C)**: Correlation analysis of RNA modifiers, positive correlation is represented by red while negative correlation is represented by blue. **(D)**: The position of a differential gene on the chromosome (All figures * represents p <0.05, ** represents p <0.01, *** represents p <0.001)

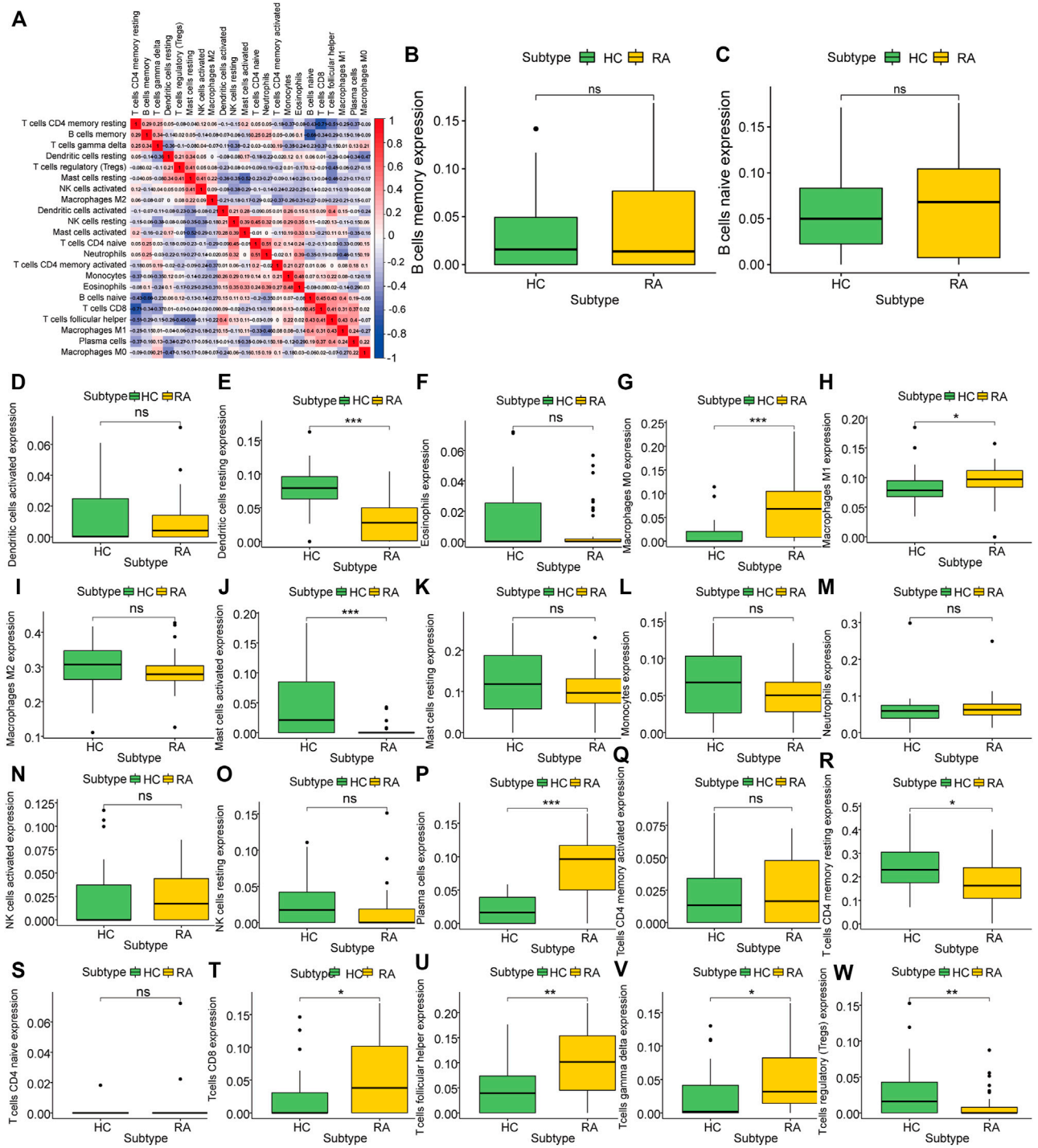


FIGURE 4 | Characteristics of Rheumatoid Arthritis Immune Infiltration. **(A)**: Correlation of infiltration degree of 22 kinds of immune cell in synovial tissue, **(B–W)**: Difference analysis of immune cell infiltration degree between synovial tissue and normal tissue in rheumatoid arthritis.

T cells regulatory (Tregs), Macrophages M0, and Neutrophils (**Figures 6C,D**). Among them, CLP1 gene and Tregs immune cells were the most influential factors in the model. The results of the immune cell prediction model are presented in the form of a nomogram (**Figure 7**).

RELATED GENES AND FUNCTIONAL ANALYSIS OF HUB RNA MODIFIERS

We constructed an mRNA-TF network of differentially expressed genes related to rheumatoid arthritis, which

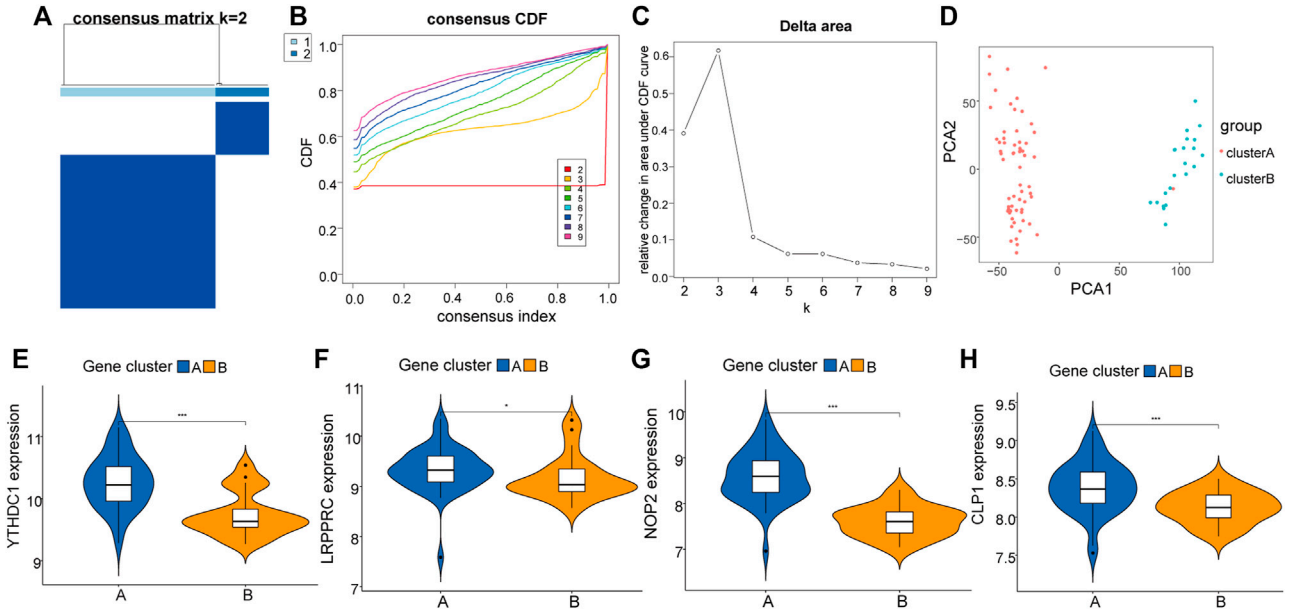


FIGURE 5 | Molecular cluster of RNA modifiers. **(A–C)**: Clustering of synovial samples based on RNA modifiers. **(D)**: PCA analysis under different groups, where red is cluster A and blue is cluster B. **(E–H)**: Differences in the expression of hub genes under different groups of.

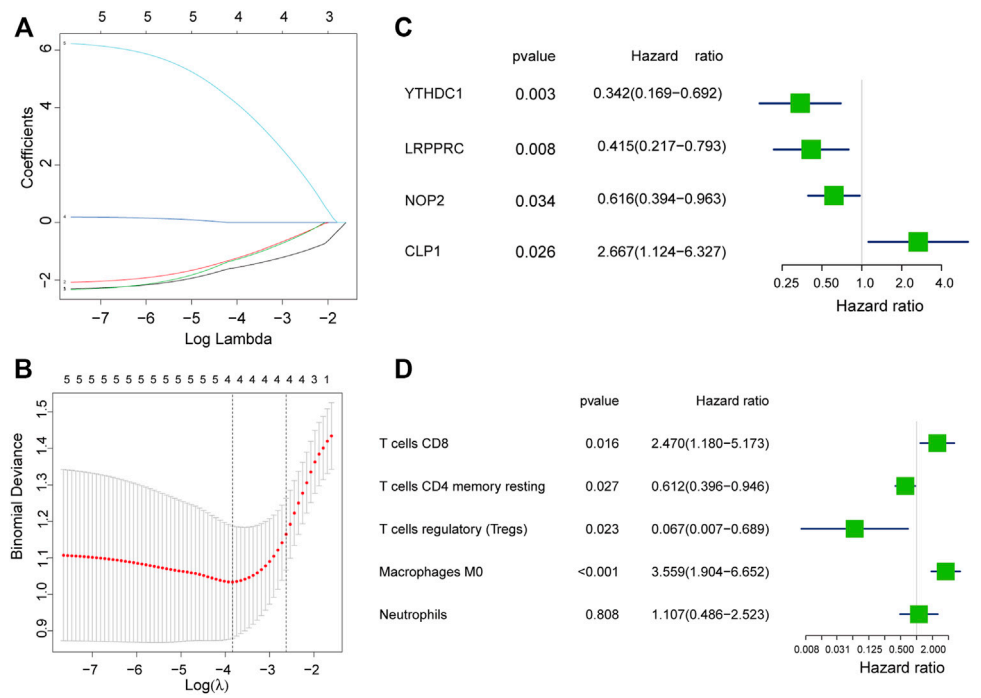


FIGURE 6 | Model construction of hub gene and immune cell. **(A,B)**: Determine the best penalty value in the LASSO regression algorithm, and screen the RNA modifiers and immune cells most related to rheumatoid arthritis. **(C,D)**: Uses forest plots to display the screened RNA modifiers and immune cells.

contained 4 mRNAs and 218 TFS (**Figure 8A**). Among them, single TF targets at most 3 RNA modification factors at the same time, and there are 17 such TF, including AFF2, CAMTA1, CSRNP3, E2F4, ELK4, HOXB3, HSF1, ID1, KLF13, NFATC3, SOX5, ZBTB33, ZNF124, ZNF205, ZNF280A, ZNF654, and ZSCAN2, additionally, the modifier NOP2 was regulated by the most TF (177 in total), followed by CLP1 (86 in total).

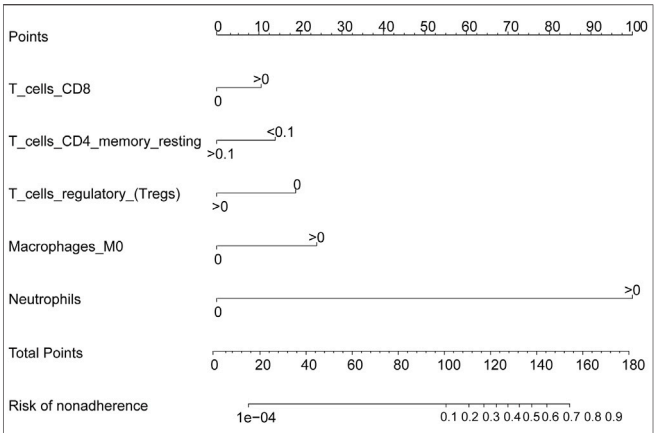


FIGURE 7 | Immune cell prediction model nomogram. The nomogram was constructed using the immune cell prediction model.

Related Genes and Functional Analysis of Hub RNA Modifiers

We constructed an mRNA-miRNA network with differential expression of RNA modifiers associated with rheumatoid arthritis, which contained 4 mRNA and 299 miRNAs (**Figure 8B**). Among them, the first seven miRNAs that simultaneously control multiple rheumatoid arthritis-related differentially expressed prognostic genes are: hsa-miR-494-3p controlling 9, hsa-miR-381-3p controlling 8, and hsa-miR-300 controlling 8 while Hsa-mir-376a-3p, HSA-Mir-3681-3p, HSA-Mir-432-5p, and HSA-Mir-543 respectively controlling 1 differentially expressed prognostic genes associated with RA.

Notably, in the mRNA-TF interaction network, TFs associated with key mRNAs are related to the progression of RA in several studies. For example, Ling-Hua Chang et al. found that after knocking out CEBPD, a protein commonly highly expressed in RA, joint damage in mice with collagen-induced arthritis was significantly lower than that in wild-type mice (Chang et al., 2012). Similar studies by Takeo Isozaki et al. pointed out that the expression of LD1 is highly correlated with CXCL16 in RA and is an essential factor affecting the inflammatory response of RA and the formation of synovial pannus (Isozaki et al., 2014). In addition, some miRNAs interacting with key mRNAs, such as miR-410-3p (Wang et al., 2020), miR-140-3p (Luo et al., 2021), etc., also constitute a vital part of the regulation of the RA process. These findings and our conclusions mutually support each other, revealing new possibilities for the treatment of RA.

To analyze the relationship between biological process, molecular function, cellular component, biological pathway and disease of the differentially expressed RNA modifiers related to rheumatoid arthritis, functional enrichment analysis of differentially expressed genes was first performed (Table 2). Further GSEA-GO analysis (**Figure 9**) showed that the differentially expressed genes associated with rheumatoid arthritis were also enriched in the regulation of macrophage apoptotic process, protein sialylation, corticosteroid receptor signaling pathway, post-Golgi vesicle mediated transport and other biological processes, phosphorylase kinase complex, eukaryotic 48s preinitiation complex, translation preinitiation complex, polysome, Golgi apparatus subcompartment, molecular functions such as NFAT protein binding, sialyltransferase activity, ligase activity forming carbon, carbon bonds, protein folding chaperone, s100 protein binding, interferon receptor activity in cell components. Next, GSEA-KEGG analysis was performed on DEGs, and the results are summarized in **Figure 10A**.

The results showed that it was enriched in arginine and proline metabolism, systemic lupus erythematosus, graft versus host disease, and other diseases (**Figures 10B–F**).

We further analyzed the correlation between key genes and the immune microenvironment. The expression of key genes CLP1 and NOP2 was significantly positively correlated with multiple immune indexes such as APC_co_stimulation,HLA,Parainflammation and the infiltration degree of various innate and acquired immune cells including $CD8^+$ T cells,T helper cells, and mast cells, according to the correlation analysis of various immune indicators calculated by ssGSEA and hub gene expression (**Figures 11A–Y**.

A–Q: Correlation analysis between NOP2 and immune cells: the horizontal axis is correlation size, the vertical axis is significantly correlated immune cells, node size represents correlation strength, node color represents significance level. R-Y: Correlation analysis between CLP1 and immune cells. The horizontal axis is correlation size, the vertical axis is significantly correlated immune cells, node size represents correlation strength, and node color represents significance level.

Molecular Clusters of Differential Genes in Rheumatoid Arthritis

To further explore the biological characteristics of gene expression in the synovial tissues of rheumatoid arthritis, we used the expression of differential genes in rheumatoid arthritis (**Figures 12A,B**) to conduct unsupervised consistent clustering again on synovial tissues. Optimal separation was achieved when all samples were divided into two different subtypes (I: $n = 27$; II: $n = 39$, **Figures 12C–E**). The expression of CLP1 was significantly different among different groups (**Figures 13A–H**), Therefore, we defined these genes as a diagnostic marker.

Diagnostic Marker Analysis

The immune correlation of diagnostic marker CLP1 was investigated, and the result demonstrated a significant positive correlation between the infiltration degree of B cells naïve, Dendritic cells activated, Eosinophils,Macrophages, Monocytes, Plasma cells,T cells CD8, and T cells follicular helper. of which, T cells follicular helper had the highest positive correlation (0.48), while mast cells, NK cells activated, and T cells CD4 memory infiltration revealed a significant negative correlation, with Mast cells having the highest negative correlation (-0.39) (**Figure 14**).

DISCUSSION

RA is one of the most common autoimmune diseases. Current research suggests that the destruction of synovial cartilage and bone caused by immune cell infiltration plays a pivotal role in the occurrence and development of RA (Coutant and Miossec,2020).

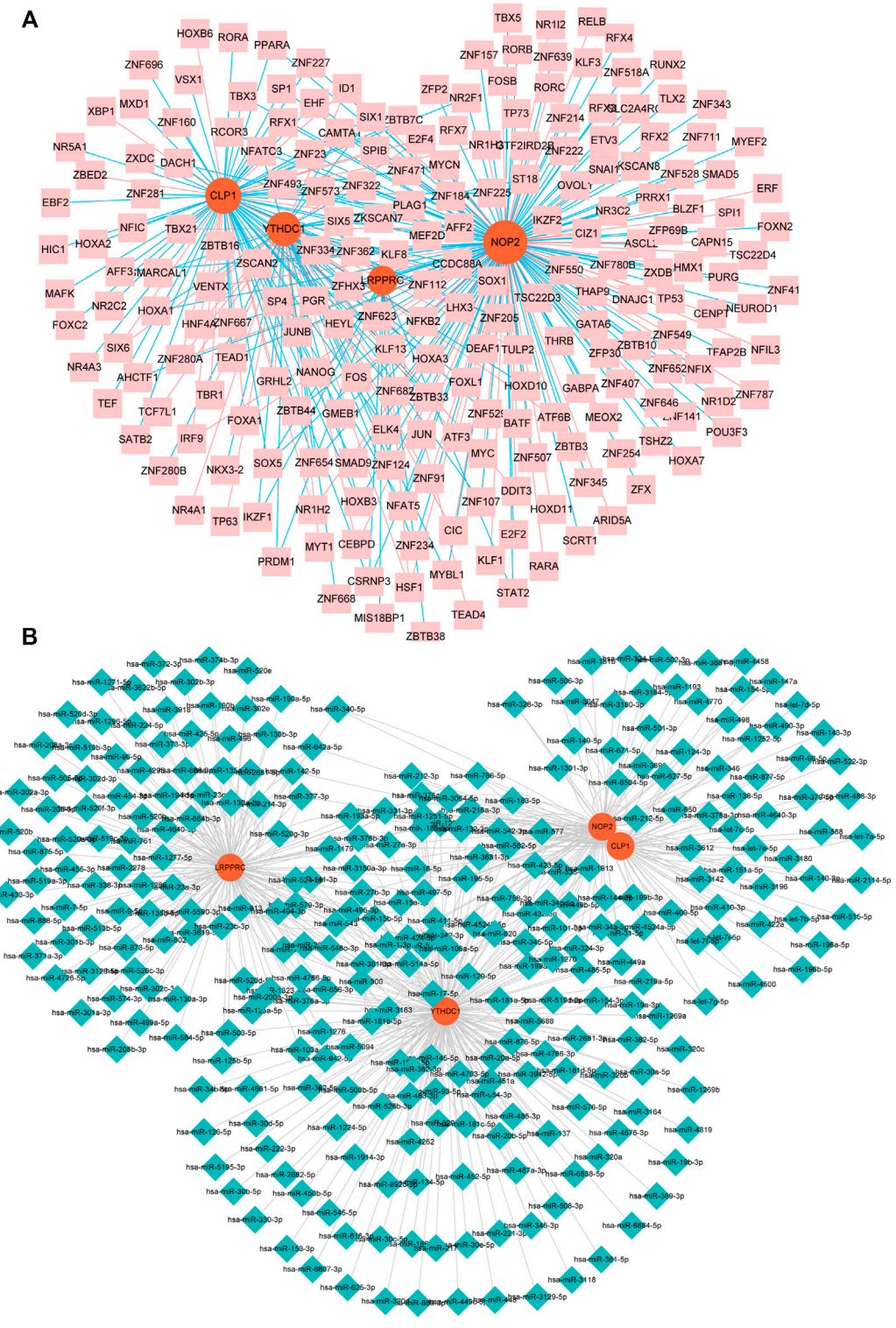


FIGURE 8 | miRNA- and TF-RNA modifier network construction. **(A)**: mRNA-TF network of hub genes related to RA, pink nodes indicate TF, red nodes indicate key genes related to RA. **(B)**: Keys related to RA in the mRNA-miRNA network of genes, the blue nodes represent miRNAs, and the red nodes represent key genes related to RA.

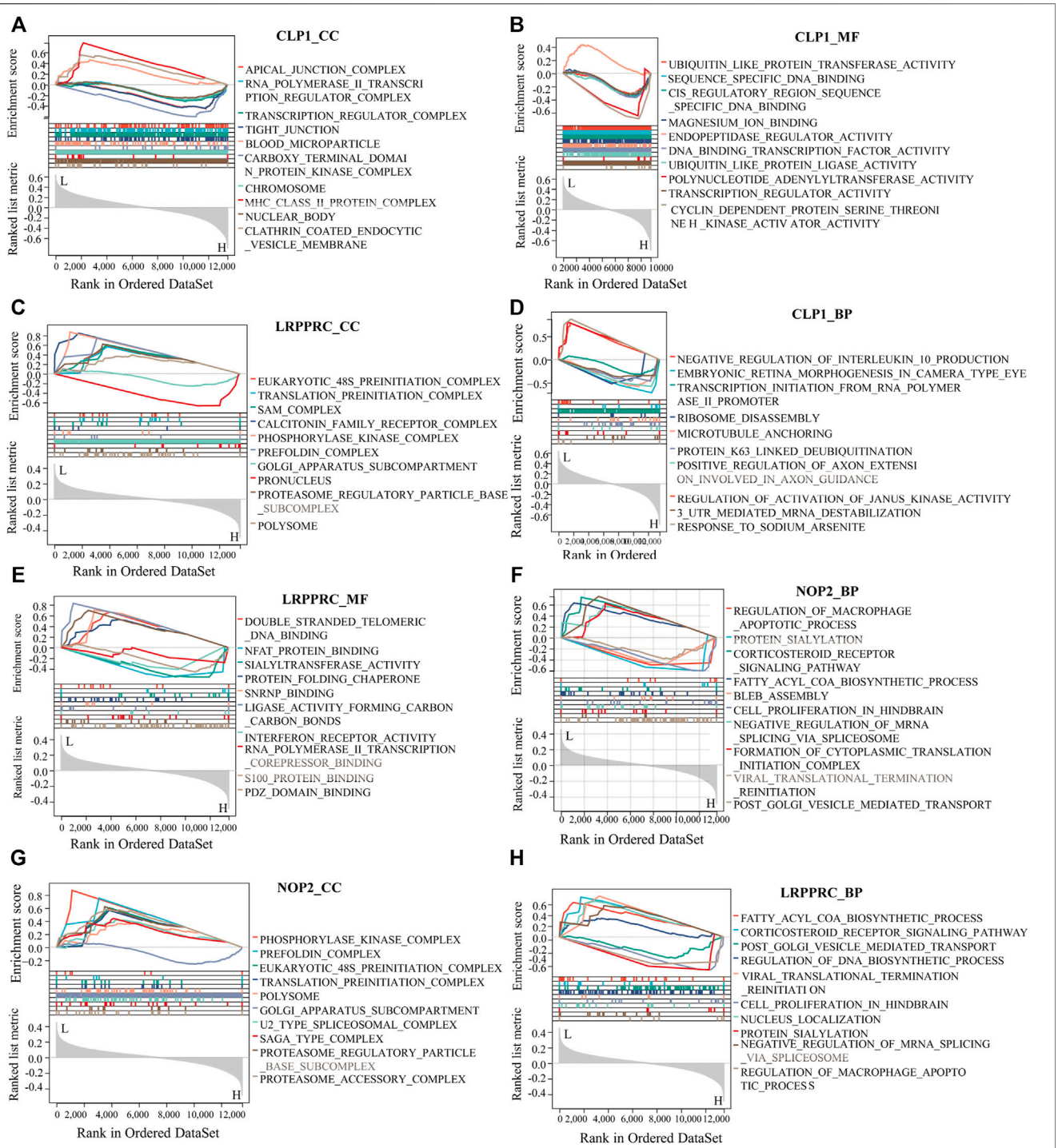


FIGURE 9 | GSEA-GO analysis of key genes. **(A–H)**: The results of functional enrichment of GSEA-GO (including BP, CC, and MF) showing CLP1, YTHDC1, LRPPRC, and NOP2 respectively displayed.

As an important post-transcriptional regulator, RNA modification participates in the biological processes of a variety of eukaryotes, profoundly affects the development of organisms, and plays a key regulatory role in a variety of diseases (Su and Randau 2011; Meyer and Jaffrey 2014). Nevertheless, the mechanism of immune cell infiltration and RNA modification in RA has not been fully elucidated. In addition, the relationship between the two in the occurrence and development of RA has not yet been clarified. The purpose of our research is to explore the significance of RNA modification

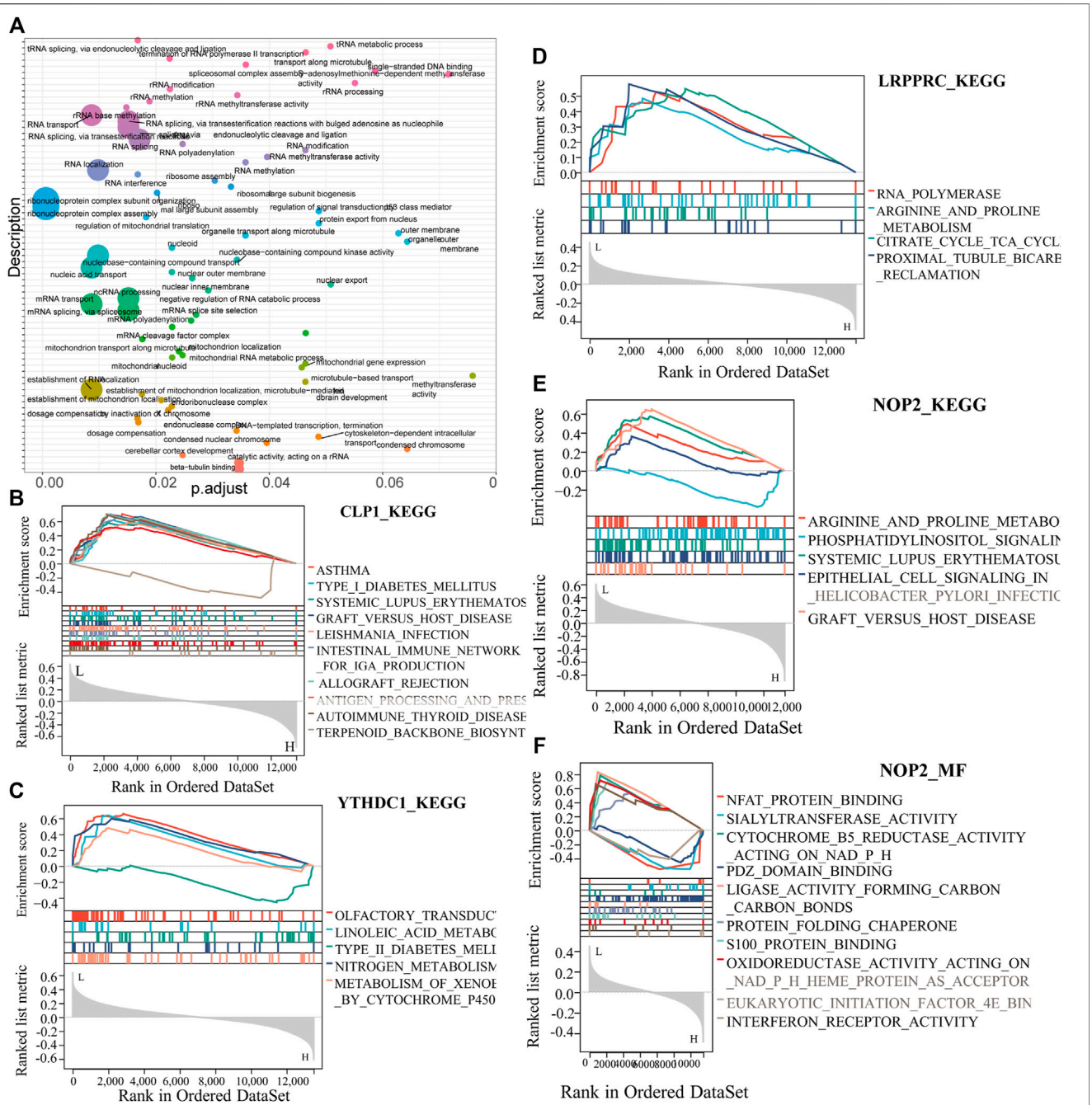


FIGURE 10 | Key gene GESA-KEGG analysis. **(A)**: Summary of key gene GSEA-KEGG functions, the horizontal axis represents the P-adjust value, **(B–E)**: CLP1, YTHDC1, LRPPRC, NOP2 GSEA-KEGG function enrichment results. **(F)**: NOP2-MF function analysis.

regulators in RA. The main goal of this research is to explore the relationship between RA and RNA modification and immune cell infiltration and to find clinically applicable serum markers.

In our research, 8 RNA modification regulators with significant differences in expression between RA synovial tissue and normal tissue were initially found from 5 RNA modification regulator gene sets, and 10 abnormal immune infiltrations in RA synovial tissue of immune cells were identified as well. On this basis, we identified 4 hub RNA modification regulator genes (YTHDC1, LRPPRC, NOP2, and CLP1) through cluster analysis. Combined with the analysis of the level of immune cell infiltration, the hub genes and infiltrating immune cell-related RA prediction model were established respectively, in which CLP1 gene and Tregs cells had the highest influence weights in the two models individually. CLP1 exhibits evident differential expression as

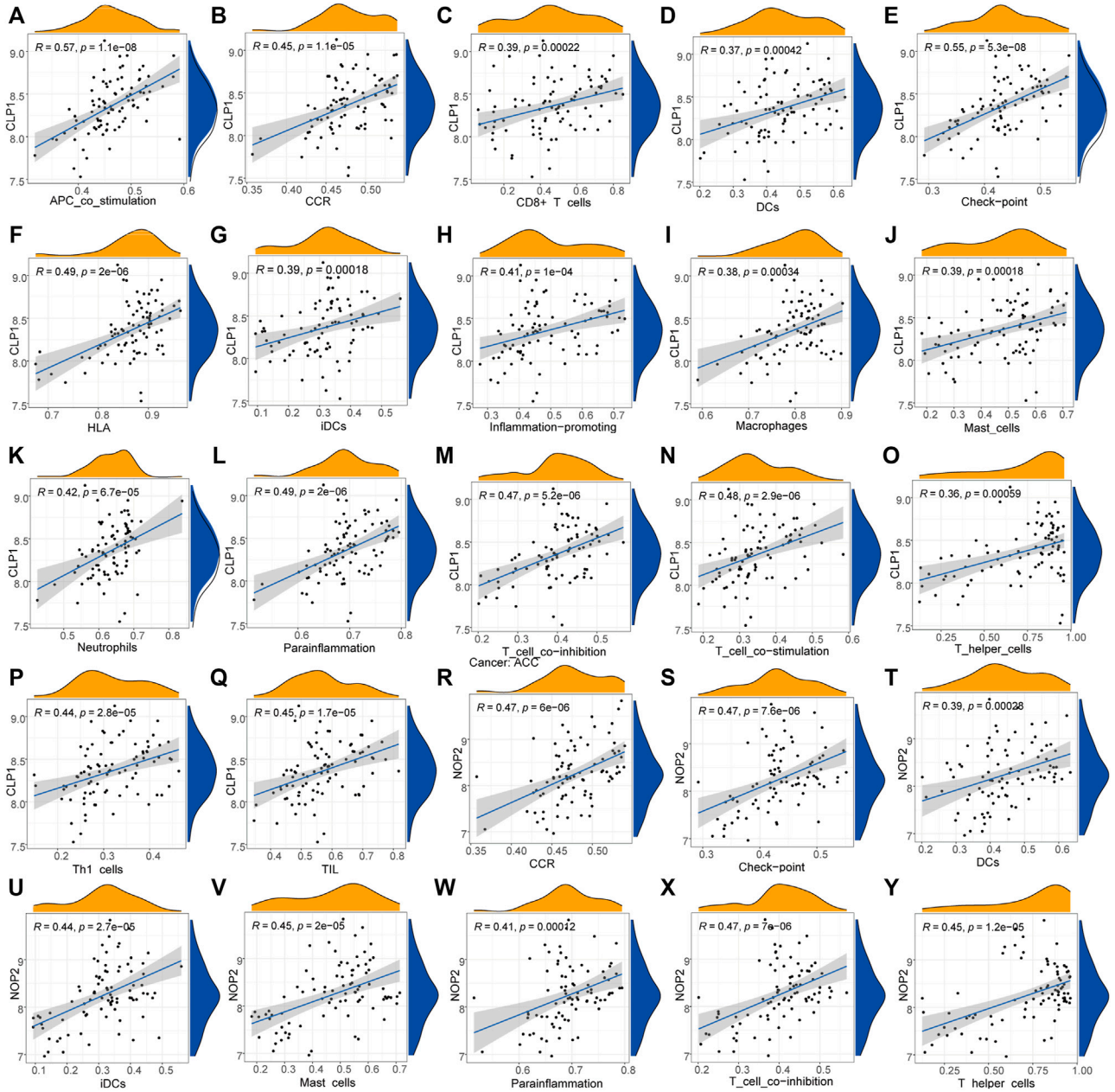


FIGURE 11 | Correlation between key genes and immunity.

shown by the molecular subtype analysis of differentially expressed genes here, hence we define CLP1 as a diagnostic marker for RA. In addition, we developed a nomogram based on 5 immune cells (T cells CD8, T cells CD4 memory resting, T cells regulatory (Tregs), macrophages M0, neutrophils) that can be used to predict the prognosis of RA patients. Furthermore, we also explored the interaction between the four key RNA modifiers and miRNA and TF. GSEA-GO and KEGG-GSEA analysis of hub genes were conducted, and further exploration of the relationship between the diagnostic marker CLP1 and immune cell infiltration was progressed.

CLP1 is one of the constituent proteins of the cleavage factor protein complex. It participates in the cleavage of the 3′untranslated region of newly synthesized mRNA molecules in the process of gene transcription. It is one of the post-transcriptional modifications necessary to produce mature mRNA (Hardy and Norbury 2016) (Ustyantsev et al., 2017). In addition, CLP1 is also involved in the precursor tRNA splicing process and plays an important role in the tRNA splicing endonuclease (TSEN) complex (Schaffer et al., 2014). In addition, Hiroyuki Fujinami et al. pointed out that CLP1 is the main RNA kinase in mice and is mainly used to

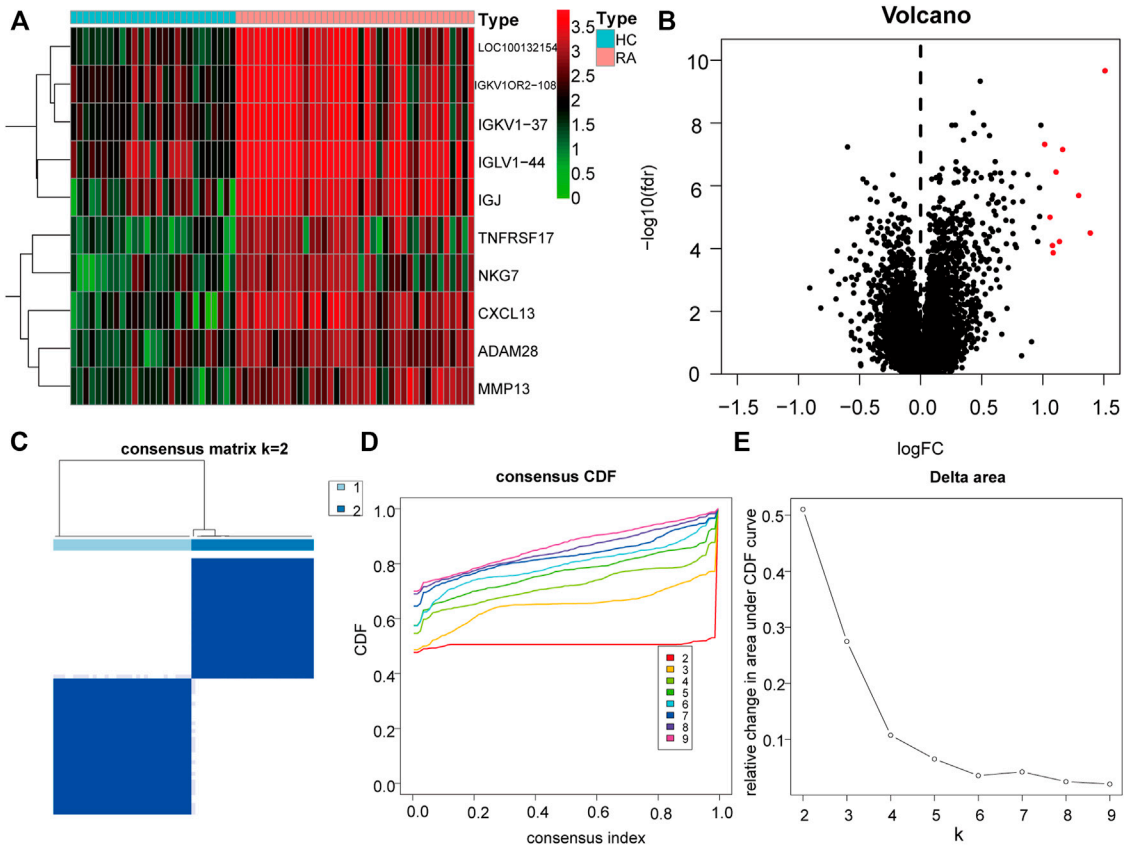


FIGURE 12 | Differential Genes and Molecular Types of Rheumatoid Arthritis. **(A)**: The heat map depicts the differential genes between RA and normal synovial tissues, the top ten highly expressed genes in RA were selected, **(B)**: The volcano map depicts the differences in gene expression between RA and normal synovial tissues, and the top ten highly expressed genes in RA tissues in the category were chosen. **(C–E)**: Clustering and grouping of synovial samples based on differential genes for rheumatoid arthritis. **(D)**: PCA analysis under different groups, where red is cluster A and blue is cluster B. **(E)**: Differences in the expression of key genes in different groups.

phosphorylate the 5′end of RNA in the siRNA pathway (Fujinami et al., 2020). In disease research, Caitlin E Monaghan et al. found that homozygous mutations in CLP1 can affect the processing of mRNA 3′and ultimately lead to neurodegenerative disease, namely pontocerebellar hypoplasia type 10 (Monaghan et al., 2021); while Kitti Szoták-Ajtay et al. found that CLP1 knockout mice have abnormal lung expansion function, and Clp1 $^{K/K}$ embryos in late pregnancy showed impaired prenatal lymphatic function and impaired lung expansion (Szoták-Ajtay et al., 2020). However, it is worth noting that the existing relationship between CLP1 and immune cells is unclear. The study by Clotilde Guyon et al. pointed out that CLP1-mediated 3′UTR shortening may be involved in the expression of thymic medullary epithelial cells and the process of antigen presentation, thereby affecting the occurrence and development of autoimmune diseases (Guyon et al., 2020). The research of Kitti Szoták-Ajta and Clotilde Guyon confirmed the feasibility of our choice of CLP1 as a diagnostic indicator of RA to a certain extent.

RA is a chronic autoimmune illness characterized by persistent inflammation that has historically hampered therapeutic management (O'Neil and Kaplan 2019). Although the immunization research for RA continues to deepen, the RA-related immunization manner is still unclear. In our research, Tregs cells are considered to play an important role in the occurrence and development of RA. Existing studies believe that Tregs cells can alleviate the progression of RA by inhibiting a variety of inflammatory cytokines produced by the synovium of bones and joints (Shi and Chi 2019; Sjaastad et al., 2021); and in the hypoxic environment of synovium, T cells will be induced to differentiate by synovial fibroblasts. This leads to a decrease in the number of Tregs cells and an increase in the number of Th17 cells, which aggravates the progression of RA (Ding et al., 2020). After analyzing the correlation between CLP1 and immune cells, we found that CLP1 was highly positively correlated with T cells follicular helper (Tfh) infiltration, but was highly negatively correlated with mast cells expression. Recent research has discovered that uncontrolled Tfh cell expansion can be seen in a variety of systemic autoimmune disorders (Yao et al., 2021). the number of circulating Tfh-like (cTfh-like) cells, their subtypes, and synovial infiltrating T helper cell is linked to RA patients' disease activity (Lu et al., 2021). The involvement of

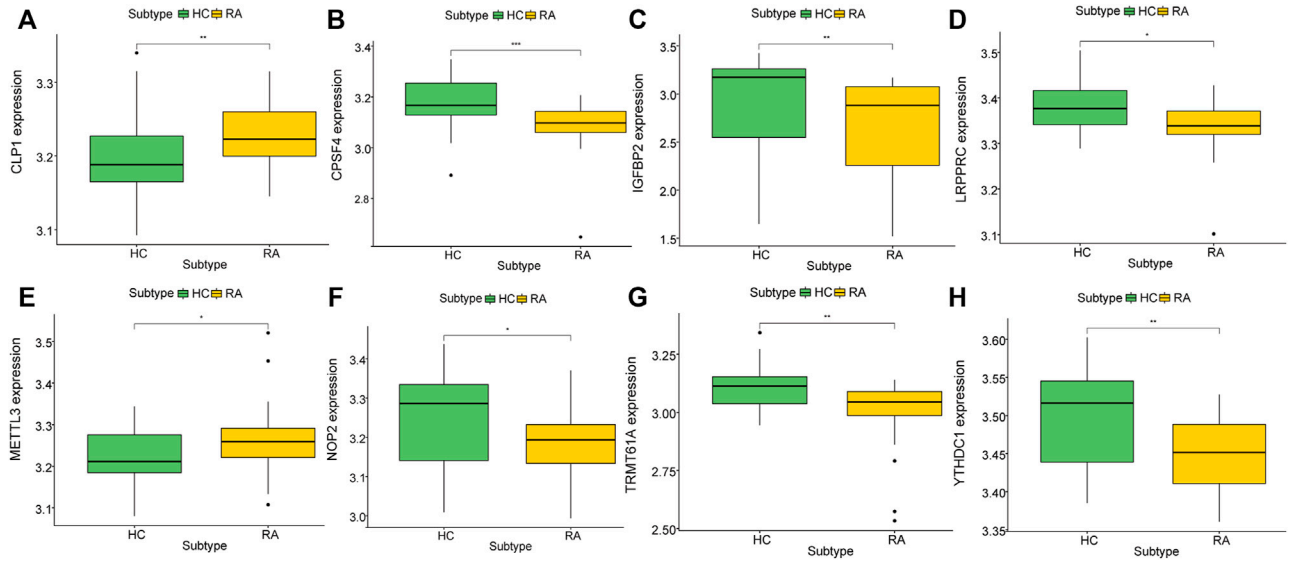


FIGURE 13 | Correlation between key genes and molecular typing of rheumatoid arthritis. **(A–H)**: The expression of different rheumatoid arthritis differential genes and hub genes differed under molecular classification groupings, only CLP1 showed significant differences, thus we define it as a diagnostic marker.

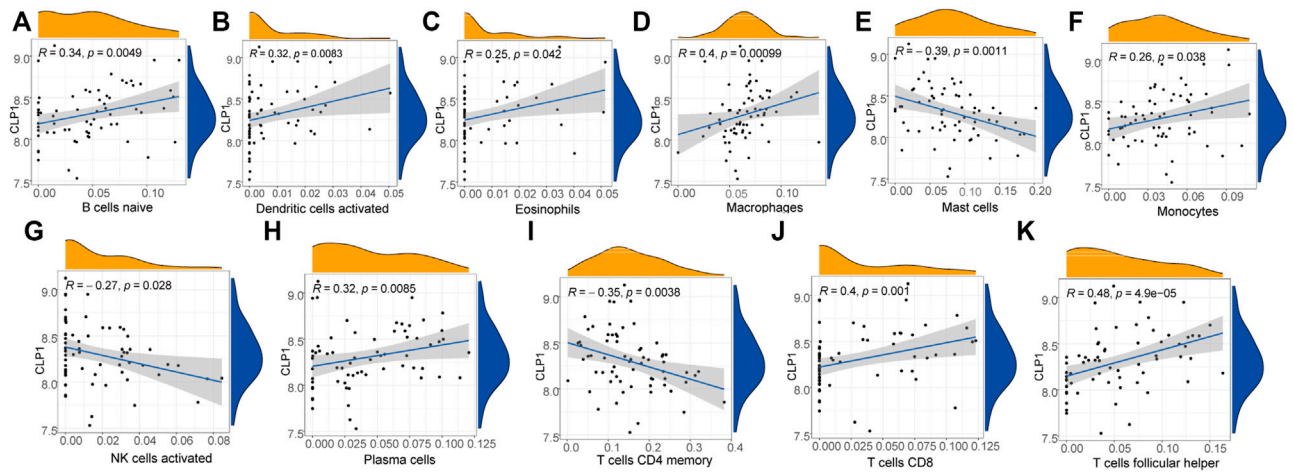


FIGURE 14 | Diagnostic marker gene and immune correlation analysis. **(A–K)**: CLP1 gene correlation analysis with immune cells, the slope is the correlation size, and the Pvalue represents the significance level.

mast cells in the development of RA is currently unknown. In different reports, mast cells have two contradictory effects: pro-inflammatory and anti-inflammatory. For example, in the animal model of mast cell defect, the model animal shows resistance to arthritis (Pitman et al., 2011); in the same way, a recent study by Julio Ramı´rez et al. found that the density of mast cells was increased in 23 RA patients undergoing joint synovial tissue biopsy (Kim et al., 2021). On the contrary, the mRNA levels of synovial mast cell-specific genes in naive early RA patients with DMARD are negatively correlated with the severity of the disease (Rivellese et al., 2015). In addition, the levels of serum trypsin and synovial mRNA are negatively correlated with systemic inflammation through the detection of CRP levels (Rossini et al., 2016). The results of our study tend to show the negative correlation effect of mast cells infiltration in RA, which may be related to the heterogeneity of RA itself. The upregulation of CLP1 may affect mast cells in RA synovium through some unknown pathways which needs further study.

In the subsequent construction of the miRNA/TF-mRNA network, we found some miRNAs that regulate CLP1 are associated with RA. Interestingly, the role of these miRNAs in the pathogenesis of RA has not yet been studied. In addition, we have also noticed that there is a wide-ranging relationship between the modifier NOP2 and TF, and the role of this regulatory factor in RA is not yet

clear. In addition, the results of GSEA-GO showed that CLP1 is also involved in the apoptosis process of macrophages. Studies have pointed out that the death of activated macrophages is related to the pathogenesis of RA (Yang et al., 2020). According to the report, the anti-apoptotic ability of pro-inflammatory macrophages in the synovial fluid of RA patients is higher than that of anti-inflammatory macrophages (Balogh et al., 2018), but the role of CLP1 and macrophage apoptosis in RA has not been elucidated. The KEGG-GSEA analysis results suggest that CLP1 is also involved in arginine and proline metabolism, systemic lupus erythematosus, graft-versus-host disease, and other disease processes, which also confirms the role of CLP1 in autoimmune system diseases.

In addition to this evidence, a recent review by Jiajie Tu et al. also pointed out that the interaction of macrophages with T cell populations has significant implications for the progression of RA. In general, macrophages can recruit or differentiate T cells toward pro-inflammatory subtypes (e.g., toward Th17); in turn, different T cell types can shift the balance of monocyte/macrophage differentiation toward disruption of Osteocytes, aggravating joint damage, and prompt macrophages to secrete various cytokines. Our findings also reveal the significant infiltration changes of macrophages and T cell subsets in the process of RA, which adds favorable evidence for the search for RA immunotherapy targets (Tu et al., 2021).

Although this research investigated the manner of RNA modification and immune cell infiltration in RA, there are still several limitations in our study. Firstly, even if the reliability of bioinformatics technology is promoted rapidly, further studies and experiments should be conducted to verify the RNA modification manner in RA. Secondly, 3 datasets were introduced and 66 joint synovial tissue samples were included in this study, but other examples such as serum samples in RA are unavailable. On the other hand, the degree of RA in samples should be classified. With the improvement based on the limitations above, our study could be verified solidly. Nevertheless, our research indicated a potential diagnostic and prognostic biomarker, CLP1, in RA, which could be the new therapeutic target in the clinic. Otherwise, we reveal the relationships among RNA modification, immune cell infiltration, and RA, which could update the acknowledge of the advance and onset in RA.

CONCLUSION

In summary, we selected 4 key RNA regulatory factors and 5 infiltrating immune cells to establish a prediction model and nomogram of RA in this study and selected CLP1 as an important diagnostic marker in RA. On this basis, we also discussed the role of Tregs, Tfh, and mast cells in the occurrence and development of RA, hoping to provide new ideas for better clinical diagnosis and treatment of RA.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

ZZ, SH, and ST planned the research concept and designed it, made provision of study material, collected data and analyzed them, wrote and approved the manuscript. XL, JR, and XY searched for data, wrote programming code. JL, MW, and ME collected pictures and graphs as well as edited them SZ and DZ collected data and analyzed them, wrote and approved and helped correct the manuscript.

FUNDING

This research is funded by Shenzhen Key Laboratory of Musculoskeletal Tissue Reconstruction and Function Restoration and Shenzhen People's Hospital (Project number: ZDSYS20200811143752005), Guangzhou Science and Technology Project (Grant No. 201904010060, Effect and mechanism of S100A4 on collagen-induced arthritis (CIA) model in mice.), National Natural Science Foundation of China (Project number:81401766), the central universities(Project number:21619348), National natural science Foundation of China(Project number:81901650) and Science and Technology Projects in Guangzhou(Project number:2021020200460).

REFERENCES

An, J., Lai, J., Sajjanhar, A., Batra, J., Wang, C., and Nelson, C. C. (2015). J-circos: an Interactive Circos Plotter. *Bioinformatics* 31 (9), 1463–1465. doi:10.1093/bioinformatics/btu842

Baldwin, A. S., Jr. (2001). Series Introduction: the Transcription Factor NF-kappaB and Human Disease. *J. Clin. Invest.* 107 (1), 3–6. doi:10.1172/JCI11891

Balogh, E., Veale, D. J., McGarry, T., Orr, C., Szekanecz, Z., Ng, C. T., et al. (2018). Oxidative Stress Impairs Energy Metabolism in Primary Cells and Synovial Tissue of Patients with Rheumatoid Arthritis. *Arthritis Res. Ther.* 20 (1), 95. doi:10.1186/s13075-018-1592-1

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2007). NCBI GEO: Mining Tens of Millions of Expression Profiles-Ddatabase and Tools Update. *Nucleic Acids Res.* 35 (Database issue), D760–D765. doi:10.1093/nar/gkl887

Broeren, M. G., de Vries, M., Bennink, M. B., Arntz, O. J., Blom, A. B., Koenders, M. I., et al. (2016). Disease-Regulated Gene Therapy with Anti-inflammatory Interleukin-10 under the Control of the CXCL10 Promoter for the Treatment of Rheumatoid Arthritis. *Hum. Gene Ther.* 27 (3), 244–254. doi:10.1089/hum.2015.127

Brzustewicz, E., Henc, I., Daca, A., Szarecka, M., Sochocka-Bykowska, M., Witkowski, J., et al. (2017). Autoantibodies, C-Reactive Protein, Erythrocyte Sedimentation Rate and Serum Cytokine Profiling in Monitoring of Early Treatment. *Cent. Eur. J. Immunol.* 42 (3), 259–268. doi:10.5114/ceji.2017.70968

Chang, L. H., Huang, H. S., Wu, P. T., Jou, I. M., Pan, M. H., Chang, W. C., et al. (2012). Role of Macrophage CCAAT/enhancer Binding Protein Delta in the Pathogenesis of Rheumatoid Arthritis in Collagen-Induced Arthritic Mice. *PloS one* 7 (9), e45378. doi:10.1371/journal.pone.0045378

Chen, H., Yao, J., Bao, R., Dong, Y., Zhang, T., Du, Y., et al. (2021). Cross-talk of Four Types of RNA Modification Writers Defines Tumor Microenvironment and Pharmacogenomic Landscape in Colorectal Cancer. *Mol. Cancer* 20 (1), 29. doi:10.1186/s12943-021-01322-w

Cong, P., Wu, T., Huang, X., Liang, H., Gao, X., Tian, L., et al. (2021). Identification of the Role and Clinical Prognostic Value of Target Genes of m6A RNA Methylation Regulators in Glioma. *Front. Cell Dev. Biol.* 9, 709022. doi:10.3389/fcell.2021.709022

Coutant, F., and Miossec, P. (2020). Evolving Concepts of the Pathogenesis of Rheumatoid Arthritis with Focus on the Early and Late Stages. *Curr. Opin. Rheumatol.* 32 (1), 57–63. doi:10.1097/BOR.0000000000000664

Davis, S., and Meltzer, P. S. (2007). GEOquery: a Bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23 (14), 1846–1847. doi:10.1093/bioinformatics/btm254

Ding, Y., Wang, L., Wu, H., Zhao, Q., and Wu, S. (2020). Exosomes Derived from Synovial Fibroblasts under Hypoxia Aggravate Rheumatoid Arthritis by Regulating Treg/Th17 Balance. *Exp. Biol. Med. (Maywood)* 245 (14), 1177–1186. doi:10.1177/1535370220934736

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of Transcription in Human Cells. *Nature* 489 (7414), 101–108. doi:10.1038/nature11233

Firestein, G. S., and McInnes, I. B. (2017). Immunopathogenesis of Rheumatoid Arthritis. *Immunity* 46 (2), 183–196. doi:10.1016/j.immuni.2017.02.006

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33 (1), 1–22. doi:10.18637/jss.v033.i01

Fujinami, H., Shiraishi, H., Hada, K., Inoue, M., Morisaki, I., Higa, R., et al. (2020). CLP1 Acts as the Main RNA Kinase in Mice. *Biochem. Biophys. Res. Commun.* 525, 129–134. doi:10.1016/j.bbrc.2020.02.066

Gao, Y., Vasic, R., Song, Y., Teng, R., Liu, C., Gbyli, R., et al. (2020). m6A Modification Prevents Formation of Endogenous Double-Stranded RNAs and Deleterious Innate Immune Responses during Hematopoietic DevelopmentA Modification Prevents Formation of Endogenous Double-Stranded RNAs and Deleterious Innate Immune Responses during Hematopoietic Development. *Immunity* 52 (6), 1007–e8. doi:10.1016/j.immuni.2020.05.003

Ge, Y., Chen, Z., Fu, Y., Xiao, X., Xu, H., Shan, L., et al. (2021). Identification and Validation of Hub Genes of Synovial Tissue for Patients with Osteoarthritis and Rheumatoid Arthritis. *Hereditas* 158 (1), 37. doi:10.1186/s41065-021-00201-0

Gui, J., and Li, H. (2005). Penalized Cox Regression Analysis in the High-Dimensional and Low-Sample Size Settings, with Applications to Microarray Gene Expression Data. *Bioinformatics* 21 (13), 3001–3008. doi:10.1093/bioinformatics/bti422

Guyon, C., Jmari, N., Padonou, F., Li, Y. C., Ucar, O., Fujikado, N., et al. (2020). Aire-dependent Genes Undergo Clp1-Mediated 3'UTR Shortening Associated with Higher Transcript Stability in the Thymus. *Elife* 9, 52985. doi:10.7554/eLife.52985

Hardy, J. G., and Norbury, C. J. (2016). Cleavage Factor Im (CFIm) as a Regulator of Alternative Polyadenylation. *Biochem. Soc. Trans.* 44 (4), 1051–1057. doi:10.1042/BST20160078

Isozaki, T., Amin, M. A., Arbab, A. S., Koch, A. E., Ha, C. M., Edhayan, G., et al. (2014). Inhibitor of DNA Binding 1 as a Secreted Angiogenic Transcription Factor in Rheumatoid Arthritis. *Arthritis Res. Ther.* 16 (2), R68. doi:10.1186/ar4507

Jonkhout, N., Tran, J., Smith, M. A., Schonrock, N., Mattick, J. S., and Novoa, E. M. (2017). The RNA Modification Landscape in Human Disease. *Rna* 23 (12), 1754–1769. doi:10.1261/rna.063503.117

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* 45 (D1), D353–d361. doi:10.1093/nar/gkw1092

Kim, K. W., Kim, B. M., Won, J. Y., Min, H. K., Lee, K. A., Lee, S. H., et al. (2021). Regulation of Osteoclastogenesis by Mast Cell in Rheumatoid Arthritis. *Arthritis Res. Ther.* 23 (1), 124. doi:10.1186/s13075-021-02491-1

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection. *Cell Syst.* 1 (6), 417–425. doi:10.1016/j.cels.2015.12.004

Littlejohn, E. A., and Monrad, S. U. (2018). Early Diagnosis and Treatment of Rheumatoid Arthritis. *Prim. Care* 45 (2), 237–255. doi:10.1016/j.pop.2018.02.010

Liu, S., Guo, Y., Lu, L., Lu, J., Ke, M., Xu, T., et al. (2020). Fibrinogen-Like Protein 1 Is a Novel Biomarker for Predicting Disease Activity and Prognosis of Rheumatoid Arthritis. *Front. Immunol.* 11, 579228. doi:10.3389/fimmu.2020.579228

Lu, J., Wu, J., Xia, X., Peng, H., and Wang, S. (2021). Follicular Helper T Cells: Potential Therapeutic Targets in Rheumatoid Arthritis. *Cell Mol. Life Sci.* 78 (12), 5095–5106. doi:10.1007/s00018-021-03839-1

Luo, Z., Chen, S., and Chen, X. (2021). CircMAPK9 Promotes the Progression of Fibroblast-like Synoviocytes in Rheumatoid Arthritis via the miR-140-3p/PPM1A axis. *J. Orthop. Surg. Res.* 16 (1), 395. doi:10.1186/s13018-021-02550-y

Meyer, K. D., and Jaffrey, S. R. (2014). The Dynamic Epitranscriptome: N6-Methyladenosine and Gene Expression Control. *Nat. Rev. Mol. Cell Biol.* 15 (5), 313–326. doi:10.1038/nrm3785

Monaghan, C. E., Adamson, S. I., Kapur, M., Chuang, J. H., and Ackerman, S. L. (2021). The Clp1 R140H Mutation Alters tRNA Metabolism and mRNA 3' Processing in Mouse Models of Pontocerebellar Hypoplasia. *Proc. Natl. Acad. Sci. U. S. A.* 11 (39), 118. doi:10.1073/pnas.2110730118

Myngbay, A., Bexeitov, Y., Adilbayeva, A., Assylbekov, Z., Yevstratenko, B. P., Aitzhanova, R. M., et al. (2019). CTHRC1: A New Candidate Biomarker for Improved Rheumatoid Arthritis Diagnosis. *Front. Immunol.* 10, 1353. doi:10.3389/fimmu.2019.01353

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry. *Nat. Biotechnol.* 37 (7), 773–782. doi:10.1038/s41587-019-0114-2

O'Neil, L. J., and Kaplan, M. J. (2019). Neutrophils in Rheumatoid Arthritis: Breaking Immune Tolerance and Fueling Disease. *Trends Mol. Med.* 25 (3), 215–227. doi:10.1016/j.molmed.2018.12.008

Pathan, M., Keerthikumar, S., Ang, C. S., Gangoda, L., Quek, C. Y., Williamson, N. A., et al. (2015). FunRich: An Open Access Standalone Functional Enrichment and Interaction Network Analysis Tool. *Proteomics* 15 (15), 2597–2601. doi:10.1002/pmic.201400515

Pitman, N., Asquith, D. L., Murphy, G., Liew, F. Y., and McInnes, I. B. (2011). Collagen-induced Arthritis Is Not Impaired in Mast Cell-Deficient Mice. *Ann. Rheum. Dis.* 70 (6), 1170–1171. doi:10.1136/ard.2010.134528

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43 (7), e47. doi:10.1093/nar/gkv007

Rivellese, F., Suurmond, J., Habets, K., Dorjée, A. L., Ramamoorthi, N., Townsend, M. J., et al. (2015). Ability of Interleukin-33- and Immune Complex-Triggered Activation of Human Mast Cells to Down-Regulate Monocyte-Mediated Immune Responses. *Arthritis Rheumatol.* 67 (9), 2343–2353. doi:10.1002/art.39192

Rossini, M., Viapiana, O., Zanoni, G., Tripi, G., Idolazzi, L., Fracassi, E., et al. (2016). Serum Levels of Tryptase Suggest That Mast Cells Might Have an Antiinflammatory Role in Rheumatoid Arthritis: Comment on the Article by Rivellese et al. *Arthritis Rheumatol.* 68 (3), 769. doi:10.1002/art.39506

Schaffer, A. E., Eggens, V. R., Caglayan, A. O., Reuter, M. S., Scott, E., Coufal, N. G., et al. (2014). CLP1 Founder Mutation Links tRNA Splicing and Maturation to Cerebellar Development and Neurodegeneration. *Cell* 157 (3), 651–663. doi:10.1016/j.cell.2014.03.049

Shi, H., and Chi, H. (2019). Metabolic Control of Treg Cell Stability, Plasticity, and Tissue-specific Heterogeneity. *Front. Immunol.* 10, 2716. doi:10.3389/fimmu.2019.02716

Sjaastad, L. E., Owen, D. L., Tracy, S. I., and Farrar, M. A. (2021). Phenotypic and Functional Diversity in Regulatory T Cells. *Front. Cell Dev. Biol.* 9, 715901. doi:10.3389/fcell.2021.715901

Smolen, J. S., Aletaha, D., and McInnes, I. B. (2016). Rheumatoid Arthritis. *Lancet* 388 (10055), 2023–2038. doi:10.1016/S0140-6736(16)30173-8

Soifer, H. S., Rossi, J. J., and Saetrom, P. (2007). MicroRNAs in Disease and Potential Therapeutic Applications. *Mol. Ther.* 15 (12), 2070–2079. doi:10.1038/sj.mt.6300311

Su, A. A., and Randau, L. (2011). A-to-I and C-To-U Editing within Transfer RNAs. *Biochem. (Mosc)* 76 (8), 932–937. doi:10.1134/S0006297911080098

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-wide Experimental Datasets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi:10.1093/nar/gky1131

Szoták-Ajtay, K., Szőke, D., Kovács, G., Andréka, J., Brenner, G. B., Giricz, Z., et al. (2020). Reduced Prenatal Pulmonary Lymphatic Function Is Observed in Clp1 K/K Embryos with Impaired Motor Functions Including Fetal Breathing Movements in Preparation of the Developing Lung for Inflation at Birth. *Front. Bioeng. Biotechnol.* 8, 136. doi:10.3389/fbioe.2020.00136

Thomas, P. D. (2017). "The Gene Ontology and the Meaning of Biological Function," in *Methods in Molecular Biology* (Clifton, NJ) 1446, 15–24. doi:10.1007/978-1-4939-3743-1_2

Tu, J., Huang, W., Zhang, W., Mei, J., and Zhu, C. (2021). A Tale of Two Immune Cells in Rheumatoid Arthritis: The Crosstalk between Macrophages and T Cells in the Synovium. *Front. Immunol.* 12, 655477. doi:10.3389/fimmu.2021.655477

Ustyantsev, I. G., Golubchikova, J. S., Borodulina, O. R., and Kramerov, D. A. (2017). Canonical and Noncanonical RNA Polyadenylation. *Mol. Biol. Mosk.* 51 (2), 262–273. doi:10.7868/S0026898417010189

Wang, Y., Hou, L., Yuan, X., Xu, N., Zhao, S., Yang, L., et al. (20201975). LncRNA NEAT1 Targets Fibroblast-like Synoviocytes in Rheumatoid Arthritis via the miR-410-3p/YY1 Axis. *Front. Immunol.* 11, 1975. doi:10.3389/fimmu.2020.01975

Weyand, C. M., and Goronzy, J. J. (2021). The Immunology of Rheumatoid Arthritis. *Nat. Immunol.* 22 (1), 10–18. doi:10.1038/s41590-020-00816-x

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: a Class Discovery Tool with Confidence Assessments and Item Tracking. *Bioinformatics* 26 (12), 1572–1573. doi:10.1093/bioinformatics/btq170

Woetzel, D., Huber, R., Kupfer, P., Pohlers, D., Pfaff, M., Driesch, D., et al. (2014). Identification of Rheumatoid Arthritis and Osteoarthritis Patients by Transcriptome-Based Rule Set Generation. *Arthritis Res. Ther.* 16 (2), R84. doi:10.1186/ar4526

Wu, C. Y., Yang, H. Y., Luo, S. F., and Lai, J. H. (2021a). From Rheumatoid Factor to Anti-citrullinated Protein Antibodies and Anti-carbamylated Protein Antibodies for Diagnosis and Prognosis Prediction in Patients with Rheumatoid Arthritis. *Int. J. Mol. Sci.* 22 (2), 686. doi:10.3390/ijms22020686

Wu, S., Li, X. F., Wu, Y. Y., Yin, S. Q., Huang, C., and Li, J. (2021b). N6 -Methyladenosine and Rheumatoid Arthritis: A Comprehensive Review. *Front. Immunol.* 12, 731842. doi:10.3389/fimmu.2021.731842

Yang, X., Chang, Y., and Wei, W. (2020). Emerging Role of Targeting Macrophages in Rheumatoid Arthritis: Focus on Polarization, Metabolism and Apoptosis. *Cell Prolif.* 53 (7), e12854. doi:10.1111/cpr.12854

Yao, Y., Chen, C. L., Yu, D., and Liu, Z. (2021). Roles of Follicular Helper and Regulatory T Cells in Allergic Diseases and Allergen Immunotherapy. *Allergy* 76 (2), 456–470. doi:10.1111/all.14639

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omics* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zhao, He, Tang, Lai, Ren, Yu, Lin, Wang, El Akkawi, Zeng and Zha. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Cell and Developmental Biology

REVIEW
published: 08 June 2022
doi: 10.3389/fcell.2022.879465

Regulation of 3D Organization and Its Role in Cancer Biology

Anghui Peng[1,2], *Wang Peng*[3], *Ruiqi Wang*[4], *Hao Zhao*[5], *Xinyang Yu*[1,2]* and *Yihao Sun*[1,2]*

[1]Zhuhai Interventional Medical Center, Zhuhai Precision Medical Center, Zhuhai People's Hospital, Zhuhai Hospital Affiliated with Jinan University, Jinan University, Zhuhai, China, [2]Guangdong Provincial Key Laboratory of Tumor Interventional Diagnosis and Treatment, Zhuhai Institute of Translational Medicine, Zhuhai People's Hospital Affiliated with Jinan University, Jinan University, Zhuhai, China, [3]Department of Oncology, Liuzhou People's Hospital, Liuzhou, China, [4]Department of Pharmacy, Zhuhai People's Hospital, Zhuhai Hospital Affiliated with Jinan University, Jinan University, Zhuhai, China, [5]The First College of Clinical Medical Science, China Three Gorges University, Yichang, China

Three-dimensional (3D) genomics is the frontier field in the post-genomics era, its foremost content is the relationship between chromatin spatial conformation and regulation of gene transcription. Cancer biology is a complex system resulting from genetic alterations in key tumor oncogenes and suppressor genes for cell proliferation, DNA replication, cell differentiation, and homeostatic functions. Although scientific research in recent decades has revealed how the genome sequence is mutated in many cancers, high-order chromosomal structures involved in the development and fate of cancer cells represent a crucial but rarely explored aspect of cancer genomics. Hence, dissection of the 3D genome conformation of cancer helps understand the unique epigenetic patterns and gene regulation processes that distinguish cancer biology from normal physiological states. In recent years, research in tumor 3D genomics has grown quickly. With the rapid progress of 3D genomics technology, we can now better determine the relationship between cancer pathogenesis and the chromatin structure of cancer cells. It is becoming increasingly explicit that changes in 3D chromatin structure play a vital role in controlling oncogene transcription. This review focuses on the relationships between tumor gene expression regulation, tumor 3D chromatin structure, and cancer phenotypic plasticity. Furthermore, based on the functional consequences of spatial disorganization in the cancer genome, we look forward to the clinical application prospects of 3D genomic biomarkers.

Keywords: chromatin, spatial structure, cancer, super-enhancer, oncogene

OPEN ACCESS

Edited by:
William C Cho,
QEH, Hong Kong, SAR China

Reviewed by:
Thomas Liehr,
Friedrich Schiller University Jena, Germany
Zaki A Sherif,
Howard University, United States

***Correspondence:**
Xinyang Yu
xyu26@buffalo.edu
Yihao Sun
syh@alumni.tongji.edu.cn

Specialty section:
This article was submitted to Epigenomics and Epigenetics, a section of the journal Frontiers in Cell and Developmental Biology

Received: *19 February 2022*
Accepted: *27 April 2022*
Published: *08 June 2022*

Citation:
Peng A, Peng W, Wang R, Zhao H, Yu X and Sun Y (2022) Regulation of 3D Organization and Its Role in Cancer Biology. Front. Cell Dev. Biol. 10:879465. doi: 10.3389/fcell.2022.879465

INTRODUCTION

In 1885, Carl Rabl first introduced the concept of the regional organization of interphase chromosomes inside the animal nucleus. He predicted the orientation of interphase chromosomes and the occupation of different regions throughout interphase, providing insights into the chromosomal arrangement in the nucleus (Cremer and Cremer, 2006b). In 1909, Theodor Boveri put forward the term chromosomal territories (CTs) and noted that each chromosome visible in the cell nucleus occupies a distinct part of the nuclear space (Cremer and Cremer, 2006a). At the end of the 20th century, Cremer and colleagues confirmed the presence of CTs using electro fluorescence imaging (Cremer et al., 1982), revealing the dynamic architecture of chromatin and disseminating potential implications in the functional compartmentalization of the nucleus.

The 2-m-long DNA in eukaryotic cells is highly compacted into the nucleus in the form of chromatin, with nucleosomes as basic subunits that organize DNA and histones into a compact chromatin state (Handoko et al., 2011). Epigenetic modifications of histones affect the affinity of DNA-bound proteins, leading to changes in chromatin configuration (Zhu and Li, 2016). At higher levels, euchromatin and heterochromatin regions are often spatially separated in the same CT (Hildebrand and Dekker, 2020). Pioneering studies have confirmed that chromosome location, chromatin structure, and transcriptional regulation are closely intertwined (Rennie et al., 2018).

The human genome comprises more than 2,000 topologically associated domains (TADs), covering more than 90% of the genome (Gorkin et al., 2014). TAD boundaries act as effective insulators that distinguish transcriptional regulatory activities from potential targets, thereby increasing the frequency of chromosomal contacts (Handoko et al., 2011; Bonev et al., 2017). One of the key factors in the formation of the TAD boundary is the interaction between the zinc finger transcription factor CTCF and the multi-subunit protein complex cohesion (Szabo et al., 2019). TAD acts as a combination of self-interacting domains bound by multiple nested CTCFs (Rao et al., 2014). However, the mechanism by which CTCF isolates chromatin interactions between TADs has not been comprehensively elucidated. To date, two hypothetical models, the handcuff model, and the ring extrusion model have been proposed to explain it. On the one hand, in the handcuff model, CTCF spans TAD boundaries, and the two ends of TAD are connected by CTCF protein which recruits cohesion (Vietri Rudan and Hadjur, 2015; Dixon et al., 2016). On the other hand, the ring extrusion model proposes that the mammalian genome is divided into TADs in the megabase range on average, with a pair of tethered DNA-binding units sliding along the DNA in opposite directions to form DNA loops, with the DNA between the units extruding out (Dekker and Mirny, 2016). It can predict the binding specificity of the CTCF protein (Xi and Beer, 2021).

In addition to the enrichment of CTCF domains, the TAD boundary also contains a large number of DNA elements such as housekeeping genes, tRNAs, and short interspersed element (SINEs) retrotransposons (Lupianez et al., 2016). TAD organization divides chromatin compartments into type A (open domain, gene-rich) and type B (closed domain, gene-poor), which alternate along the chromosome and are approximately 5 Mb long (Dekker et al., 2013; Hildebrand and Dekker, 2020). A high-resolution multiple interactions map of the 4.5 Mb domain in the mouse X chromosome inactivation center showed that intra-TADs interactions were stronger than inter-TADs interactions (Nora et al., 2012). In general, TAD is highly conserved in different cell types, whereas compartments A and B, and gene expression patterns in open chromatin loci, are highly cell type- and tissue-specific (Thurman et al., 2012; Fortin and Hansen, 2015). Complex DNA topologies, including polymer loops, are frequently coupled with specific interaction kinetics of proteins and DNA molecules on target sequences (Zhang et al., 2006). Transcriptional regulation plays a critical role in lineage differentiation and cell fate determination in eukaryotes. This complex transcriptional system comprises a series of regulatory elements, such as enhancers and super-enhancers (SEs) that finely tune target gene expression (Wray et al., 2003; Prieto and Maeshima, 2019). Enhancers are short cis-regulatory elements, whereas SEs spanning dozens of kilobases are clusters of putative enhancers playing decisive roles in defining cellular identity (Pott and Lieb, 2015; Peng and Zhang, 2018). In human cells, most enhancers interact remotely with the promoters of target genes, whereas only a few enhancers regulate proximal promoters (Mora et al., 2016). Enhancers play an important role in the active establishment of chromatin loops. Because enhancers can be physically associated with the promoter of the target gene by 3D circularization or tracking, chromatin interactions are not always linearly proximal (Lettice et al., 2003; Montavon et al., 2011; Proudhon et al., 2015). The effects of long-range promoter-enhancer interactions appear to be mediated, in part, by loop formation. In other words, the loop structure enables the long-range regulation of target genes.

Chromatin structure alterations are a major cause of transcriptional dysregulation in various diseases, including cancer. The stable 3D chromatin state ensures precise gene expression by organizing regulatory elements and gene loci at close spatial distances, thereby ensuring the normal structure and function of the genome (Ma et al., 2019). The specific subsets of oncogenes expressed by each cell are directly related to gene regulation and transcriptional activity (Vicente-Duenas et al., 2013). The 3D genome structure of tumor cells is clearly distinguishable, and its TAD structure is smaller than that of normal cells (Taberlay et al., 2016). Tumorigenesis is often accompanied by a large number of mutations, and the mutated genes are high efficiently transcribed in broadly accessible chromatin regions. Transcribed regions are reassigned to greater spatial proximity, enabling genes to share regulatory elements and transcriptional factors (TFs) (Mourad et al., 2014). Alterations in the chromatin spatial structure of tumor cells promote the formation of different combinations of enhancers and oncogenes in the dynamic transcription process. Additionally, the causal relationship between heterochromatin dysfunction and increased genomic instability is a well-established mechanism underlying cancer progression. Given the significance of genome topology, an increasing number of unsolved issues are related to how it affects human cancer biology.

Main Technologies of 3D Genomics

Over the past few decades, an increasing number of tools have been developed to study the physical organization and transcriptional regulation of genomes. Advanced techniques have made it possible to capture alterations in chromatin conformation during different developmental stages inside the nucleus. Gradually, more and more technologies aimed at 3D chromatin spatial detection have emerged. The major 3D genomics techniques are listed in **Table 1**.

Dekker et al. developed chromatin conformation capture (3C), which, along with its derived technologies, such as 4C, 5C, HiC, and ChIA-PET, has allowed genome-scale detection of long-range interactions between specific sites of chromatin in

TABLE 1 | Main technologies of 3D genomics.

Technologies	Characteristics	Advantages	Limitation	Reference
3C	The interaction mode is one versus one	Precisely detects the interaction between two target regions	Low throughput; low resolution	Dekker et al. (2002)
4C	Reverse PCR; the interaction mode is one versus all	Detects the interactions between one target region with genome	Interaction data are prone to bias	Simonis et al. (2006)
5C	Multiple Primer design; the interaction mode is many versus many	Detects interactions among multiple regions	Low coverage and difficult-to-assess PCR redundancy	Dostie et al. (2006)
Hi-C	Interaction mode is all versus all	High-throughput detection of genome-wide interactions	High cost of sequencing; difficult to analyze because of the large amount of data	Lieberman-Aiden et al. (2009)
Capture-C	Target domain capture	Provide an unbiased, high-resolution map of cis interactions for hundreds of genes in a single experiment.	Sampling is limited to a defined domain of chromatin	Hughes et al. (2014)
3D FISH	DNA imaging scheme in single cells	Highly multiplexed detection of a genomic region of interest	Harsh treatments are required to prepare the chromatin for the FISH probes	Solovei et al. (2002)
DNase-HiC	Endonuclease DNase I replaces the restriction endonuclease	Higher effective resolution than traditional Hi-C libraries	DNase exhibits sequence bias at cleavage sites with low GC content	Ramani et al. (2016)
Micro-C	Micrococcal nuclease replaces the restriction endonuclease restriction enzymes	Able to access shorter-range interactions at higher resolution	Cannot capture long-range interactions	de Souza, (2015)
ChIP-seq	Genome-wide profiling of DNA-binding proteins, histone modifications, or nucleosomes	High resolution, low noise, great coverage, and decreased cost of sequencing	Difficulty in analyzing data owing to bias	Park, (2009)
ATAC-seq	DNA accessibility with hyperactive Tn5 transposase	Fast and sensitive detection for genome-wide chromatin accessibility	Difficult to achieve ideally cut fragments	Buenrostro et al. (2015)
ChIA-PET	Protein-centric chromatin conformation method	High-throughput detection of protein-mediated genome-wide interactions	Difficult to obtain specific antibodies for protein detection	Li et al. (2017)
HiChIP	Protein-centric chromatin conformation method	More efficient and lower input requirement than ChIA-PET; multi-scale genome architecture with greater signal to the background than *in situ* Hi-C	Biased signal owing to the enrichment of target binding sites	Mumbach et al. (2016)

candidate regions (Dekker et al., 2002; Sati and Cavalli, 2017), revealing the hierarchical structure of chromosomes and providing information on the organization and interaction of chromatin in different cell types. Chromosome conformation capture-on-chip (4C) can generate a genome-wide interaction map of multiple sites with a bait sequence (Simonis et al., 2006). Chromosome conformation capture carbon copy (5C) technology allows for the chromatin interactions of a large number of genes by drawing an interaction map between multiple loci (Dostie et al., 2006). High-throughput chromosome conformation capture (Hi-C) technology, which can capture all chromatin interactions in the whole genome, is currently a robust tool over mass capture technologies to identify chromatin loops and describe TAD compartment conditions (Lieberman-Aiden et al., 2009).

Although HiC reveals TADs as conserved features of chromatin organization, it is limited to the observation of thousands of cells and the reliance on restriction enzymes for fragmentation. Some techniques can make up for these limitations. FISH on 3D-preserved nuclei (3D-FISH) in combination with 3D-microscopy and image reconstruction provides detailed information on the chromatin architecture by visualizing individual chromosomes at the interphase stage, thus providing direct evidence for CTs in the nucleus at the single-cell level (Solovei et al., 2002; Cremer and Cremer, 2010). DNase Hi-C and Micro-C use DNase I and micrococcal nuclease (MNase), respectively, instead of digesting cross-linked genomes, generating mononucleosomes, and inferring genome structure maps at single-nucleosome resolution (de Souza, 2015; Ramani et al., 2016). The relentless development of 3D genomic techniques led to cutting-edge technologies. Capture-C yields hundreds of fold fragment enrichment, significantly improving the detection efficiency of local interactions in target chromatin regions (Hughes et al., 2014). Chromatin immunoprecipitation (ChIP) technology is an effective tool for investigating TFs and histone modifications (Park, 2009). The assay for transposase-accessible chromatin (ATAC) technique reveals the chromatin state of most noncoding functional elements in the whole genome (Buenrostro et al., 2015). In addition, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (Li et al., 2017) and HiChIP (Mumbach et al., 2016) can comprehensively capture specific protein-mediated interactions in the whole genome. Combined with high-throughput sequencing, these technologies provide a way to understand how eukaryotic genomes fold and organize inside the nucleus.

From the perspective of the multi-omics level of DNA mutation, epigenetic alterations, histone modification, 3D conformation, and transcriptional regulation, Hi-C is generally combined with one or more additional techniques (whole-genome sequencing (WGS), ChIP-seq, ATAC-seq, and RNA-seq) to investigate the transcriptional regulation and pathogenetic mechanisms of cancers (**Figure 1**). With the rapid development

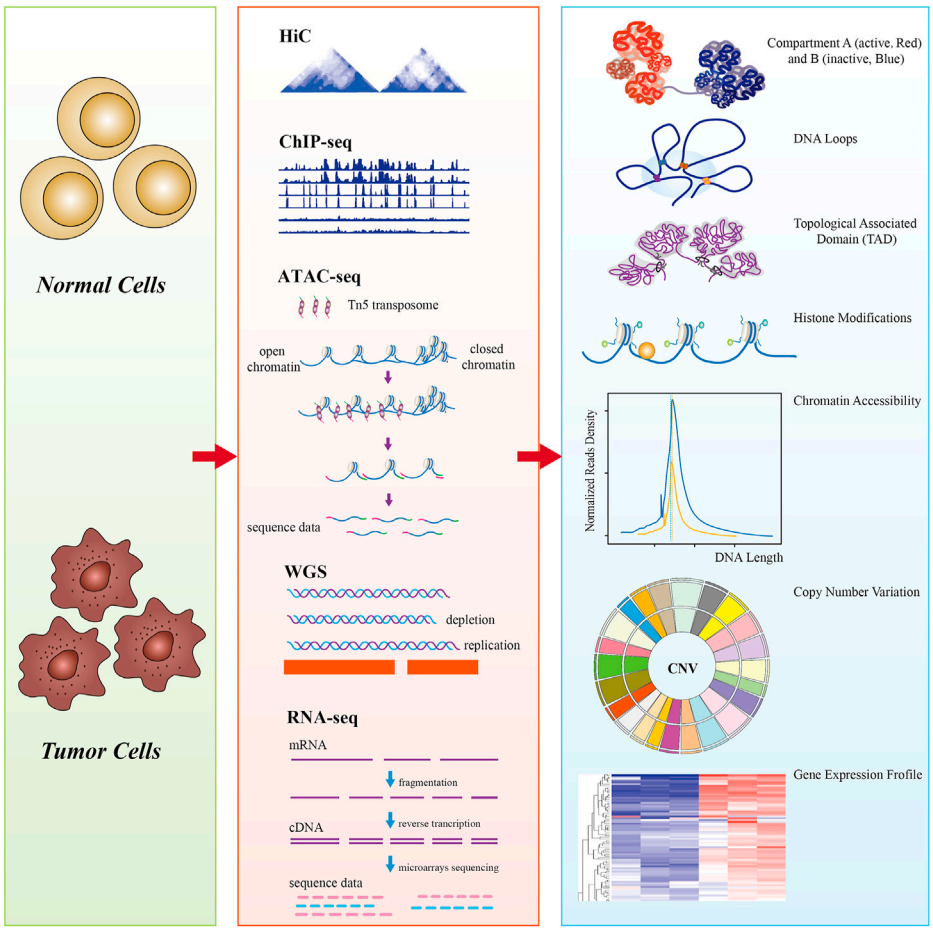


FIGURE 1 | A schematic diagram of multi-omics analysis between normal cells (control) and tumor cells. Hi-C data showed that tumor chromosome territories could be partitioned into A (active, red) and B (inactive, blue) compartments, chromatin is folded into topologically associating domains (TADs) (100–1,000 kb), and enhancer–promoter loops (10–500 kb); ChIP-seq revealed tumor genome-wide epigenetic changes, such as histone modifications; ATAC-seq detects tumor genomic chromatin accessibility using Tn5 transposase-specific recognition cleavage of open chromatin; whole-genome sequencing (WGS) detects tumor chromatin structural variations, including copy number variations (CNVs); genome-wide detection of tumor-specific genes by RNA-seq. Multi-omics reveals the hierarchical structures of 3D genome organization, transcription regulation, and structure variation mechanisms of the whole tumor genome at the genetic, epigenetic, and RNA levels.

of single-cell technology, multi-omics have opened up new avenues for revealing the tumor cell pathogenesis and underlying regulatory mechanisms.

Transcriptional Dysregulation Participates in Oncogenesis

The smooth operation of gene expression patterns plays a fundamental role in the finely-tuned regulation of gene expression. Transcriptional dysregulation triggers carcinogenesis, including abnormal cellular processes, such as hyperproliferation, immortality, metastasis, and immune escape (Liu et al., 2018; Gupta et al., 2020). Cis-regulatory elements control tissue-specific gene expression underlying tumor cell development, determining cell identity and cell fate (Huang et al., 2021). SEs can activate oncogene expression, irrespective of the distance or orientation to their transcription start sites (Tang et al., 2020). The SE-associated transcription program is key to revealing the mechanism of tumorigenesis (Zhang T. et al., 2020). In a wide array of cancer types, intensive transcription of oncogenes in cancer cells is often promoted by SEs (Sengupta and George, 2017). *MYC* is a classical SE-associated oncogene involved in global gene transcription amplification. The *MYC* members of human oncogenes include *c-MYC*, *MYCL*, and *MYCN*. *MYCN* protein is an oncogenic driver that functions in transcriptional programs similar to those of *MYC* (Zanotti et al., 2021). Cutting-edge research has pointed out that the association of SEs with multiple oncogenes is acquired during hepatocarcinogenesis, and the increase in SEs at *MYC* and *MYCN* was observed in hepatocellular carcinoma (HCC) cells (Tsang et al., 2019). In osteosarcoma and multiple myeloma (MM), most SE-amplified genes are bound by *MYC* (Loven et al., 2013; Chen et al., 2018). Xiang et al. found that a super-enhancer of approximately 150 kb located 515 kb upstream of *MYC* forms

a chromatin loop with *MYC* in human colorectal cancer (Xiang et al., 2014). *c-MYC* is abnormally highly expressed in the process of B cell carcinogenesis owing to the chromatin space remodeling (Jiang S. et al., 2017). The inhibition of transcriptional cyclin-dependent kinases (CDKs) leads to global repression of *MYCN*-dependent transcriptional amplification and sustained growth of neuroblastoma cells. In line with this, the upregulation of the active transcriptional program in neuroblastoma cells is promoted by the development of SEs (Chipumuro et al., 2014). Yuan et al. integrated RNA-seq and ChIP-seq to explore SE-mediated transcriptional dysregulation in nasopharyngeal carcinoma (NPC) cells by screening 19 SE-associated candidate genes (Yuan et al., 2017). They validated that five genes (*BCAR1*, *F3*, *LDLR*, *TBC1D2*, and *TP53TG1*) sustain the cell survival and promote proliferation of NPC. DNA-binding motif analysis has shown that *ETS2* is a potential SE-promoting TF during NPC tumorigenesis (Yuan et al., 2017). *RUNX1* and *DNAJB1*, identified as SE-associated oncogenes in esophageal squamous cell carcinoma (OSCC), significantly promote OSCC cell proliferation (Jiang Y. Y. et al., 2017). In prostate cancer cell lines and tissues, two enhancers located 63 kb upstream and 48 kb downstream of the *PTBP3* region were identified to specifically loop to the *PTBP3* promoter (Kubiak et al., 2019). Overall, cell type-specific gene transcriptional dysregulation is the hallmark of malignancies and is primarily underpinned by alterations in SEs. The dependence on SE-driven transcription in cancer biology greatly benefits tumorigenesis. Aberrant cell growth and proliferation prompted by dysregulated transcriptional progression renders cancer highly invasive and unconducive to clinical therapy.

Alterations of 3D Genome Architecture in Cancers

In many tumor types, decompressed heterochromatin leads to decreased chromosomal stability, DNA damage, fragmented DNA folding, and activated transcription, eventually triggering the malignant transformation in the early stage of carcinogenesis (Xu et al., 2020). Transcriptional differences increase gene expression in the transition domain of type B to type A compartment, promoting interactions in type A compartments on chr16-22 in breast cancer (Barutcu et al., 2015). Significant differences in the stereotypical folding of each chromosome which boosts gene expression in B-type to A-type compartment conversion regions were observed in genome-wide chromatin conformation between normal epithelial cells and breast cancer cells (Barutcu et al., 2015). In T cell acute lymphoblastic leukemia (T-ALL), the loss of boundary sites of TADs, which may support the gene regulation theory by promoting enhancer promoter interactions and isolating different regulatory units, can activate oncogenes insulated neighborhoods (Hnisz et al., 2016). Li et al. found that the alteration of CTCF binding, which disrupts the robustness of the TAD boundary, interferes with the oncogenic transcription program of the *TAL1* gene, dramatically altering leukemogenic processes. The polarity and organization of the TAD boundary depend on the CTCF orientation (Li et al., 2020). Kloetgen et al. integrated Hi-C, RNA-seq, and CTCF ChIP-seq technologies, revealing that TAD boundary disruption is associated with increased enhancer promoter interactions and chromatin accessibility (Kloetgen et al., 2020). Zhou et al. uncovered 24 dynamic patterns characterizing 3D genome recompartmentalization accompanied by lower CTCF binding at the TAD boundary in estradiol (E2)-induced breast cancer cells (Zhou et al., 2019). The conformation of the 3D chromatin genome is a deeper layer of inter-tumor heterogeneity. In glioblastoma, specific boundary loss causes the enhancer to interact abnormally with the oncogene *PDGFRA* (Lettice et al., 2003). The immune-related gene *CD276,* which co-expressed with stem cell genes, displays increased accessibility in glioblastoma stem cells to achieve a shared 3D genome state that triggers self-renewal. It is thought that genome instability destroys the normal transcription program (Johnston et al., 2019). Collectively, high-resolution 3D tumor genome maps provide global insights for evaluating cancer transcription programs, genome stability, and compartment conversions. The integration of information on loops, territories, and compartment construction contributes to a comprehensive understanding of tumor genome organization and etiology.

Tumor Structure Variation and TAD Boundary

In most cancers, structural variants promote oncogenesis through a variety of mechanisms, including the genome with complete or partial chromosomal gain and loss. A comprehensive understanding of the entire cancer system is required to dissect the interplay between higher-order chromatin structures and somatic mutations (Harbers et al., 2021). Multiple structural and numerical chromosomal aberrations lead to profound changes in the structure and function of the genome, including translocations, insertions, point mutations, copy number variations (CNVs), and chromosomal aneuploidy (Teixeira and Heim, 2005). These variations are hallmarks of most cancer genomes. Cancer epigenetics and genetics may have complementary roles in this regard. A typical example is the Philadelphia chromosome (Ph) first discovered by Nowell and Hungerford and described as a typically short chromosome 22 recurring in tumor cells of patients with chronic myelogenous leukemia (CML) (Nowell et al., 1960).

Increasing evidence has demonstrated that chromosomal translocation coupled with the disruption of 3D genome organization plays a role in carcinogenesis. A study on carcinogenic translocation events suppressed by tyrosyl-DNA phosphodiesterase 2 (*TDP2*) found that the loss of non-homologous end joining (NHEJ) repair during transcription disrupts genome stability (Ramsden and Nussenzweig, 2021). The frequency of translocation selection is related to the spatial contact probability of interaction sites. In MM, CNV breakpoints overlap with the TAD boundaries. By integrating Hi-C, WGS, and RNA-seq data of MM cell lines, Wu and colleagues identified 56 inter-chromosomal translocations with multiple inter-chromosomal interactions. The intensity of the overall spatial

interaction between chromosomes of MM cell lines is significantly higher than that of normal B cells, indicating that the 3D conformation of the cancer cell genome is affected by inter-chromosomal translocations during MM development (Wu et al., 2017). Another cause of tumor genome instability is double-strand breaks (DSBs) during gene transcription, possibly resulting in chromosomal translocation. Translocations are likely to occur at hotspots of DSBs in regions with extreme spatial proximity (Zhang et al., 2012). Furthermore, specific 3D FISH chromatin landscapes unveil gene activity-related changes containing spatial relationships of DNA-proteins and translocation in human cancers (Kocanova et al., 2018; Kulasinghe et al., 2020).

Tumor structure variations are involved in cancers, as they can affect TAD integrity, reorganize specific enhancer promoter interactions, and alter gene expression (Anania and Lupianez, 2020). Insulator proteins such as CTCF bind to the TAD boundary, preventing the interactions of genes and regulatory elements between different TADs (Kim et al., 2007). However, a recent study documented that TAD boundary destruction can alter the TAD structure and establish new TADs (Ulianov et al., 2016). New domains can also be established without destroying the TAD boundaries. For example, genomic rearrangement with breakpoints in TADs leads to their breakage and fusion, ultimately activating oncogenes and ultimately triggering tumorigenesis (Groschel et al., 2014; Northcott et al., 2014). Dixon et al. found extensive deletion of enhancers at the distal end of the region where the structural mutations occurred. Enhancers are located near genes that are frequently mutated in cancers (Dixon et al., 2018). Prostate cancer cells retain the ability to segment their genome into megabase-sized TAD regions and establish new smaller cancer-specific TADs, whose boundaries mostly appear in the CNV area (Taberlay et al., 2016). Although the genome of tumors typically has more TADs, their average TAD size is smaller than that of normal cells (Wu et al., 2017). Oncogene dysregulation can be caused by the loss or reduced activity of TAD boundaries. Gain-of-function mutations in IDH are characteristic of the main pathological and treatment prognostic categories of gliomas. Flavahan and colleagues found that CTCF binding sites are significantly reduced in IDH mutant gliomas, allowing a potent enhancer to aberrantly contact and activate *PDGFRA* expression (Flavahan et al., 2016). CTCF site depletion at the TAD boundary and variation in chromatin structure are found in the aberrant expression of pathogenicity-related genes in some cancers. A general genome-wide dysregulation of gene expression associated with TAD boundaries has been found in B cell precursor acute lymphoblastic leukemia (BCP ALL) in hyperdiploid children. Hyperdiploid ALL shows abnormal chromosome morphology, whereas low expression of CTCF and cohesin is observed in hyperdiploid ALL (Yang et al., 2019).

Overall, structural variation in the chromosomal aberration program of the cancer epigenome leads to chromatin remodeling and dysregulated gene expression, whose malignant mechanism is related to the destruction of TAD boundaries.

Outlook of 3D Genomics in Tumor Diagnosis and Treatment

Because cancer is characterized by morphological changes in the cell nucleus, exploring the chromatin structure in cancer is expected to help identify candidate biomarkers (**Figure 2**). High-throughput analysis of genome-wide histone modifications shows that in almost all cancer types, a group of genes have unique epigenetic characteristics that are closely related to different stages and different kinds of tumors. Currently, epigenetic markers are used as effective biomarkers in early clinical screening and the prediction of patient diagnosis and treatment response. Identifying specific histone signatures associated with each type of cancer enables not only a more accurate diagnosis and prognosis, but also lays the foundation for the design and evaluation of epigenetic agents (Audia and Campbell, 2016). The use of inhibitors of DNA methyltransferases and HDACs is clinically effective for several cancers. For instance, several KDMs in the family of histone lysine demethylases have been implicated in the development of various cancers, and are thus considered potential drug targets. KDM inhibitors have potential value for elucidating tumor cell function and tumor therapy (Hoffmann et al., 2012; McAllister et al., 2016). Moreover, HDAC8 knockdown initiates a similar differentiation program as selective small-molecule inhibitors in neuroblastoma cells (Oehme et al., 2009). Clinical implications of biological programs allow the design of HDAC8-selective small-molecule inhibitors for cancer cell suppression.

The mechanisms by which oncogenes control myriad cellular processes to induce tumorigenesis expose the fragility and difficulties in treatment. Transcriptional inhibitors are potential therapeutic agents for treating certain tumors. In *MYC*-dependent cancers, interference of chromatin-dependent signal transduction with RNA polymerase II (RNA Pol II) and inhibition of RNA Pol II transcription initiation and elongation are therapeutic principles in malignancies. CDK7, a member of a family of CDKs involved in regulating RNAPII initiation, pause, and elongation, preferentially binds to SE and activates SE-related gene expression (Larochelle et al., 2012). The selective targeting of mechanisms that promote the overall transcriptional amplification in tumor cells renders CDK7 inhibition an effective target for the treatment of cancers driven by specific oncogenes (Chipumuro et al., 2014). The blockade of CDK7 function is expected to suppress the expression of genes primed for transcription. For instance, CDK7 inhibitors commonly repress *MYC*, an oncogene overexpressed in 70% of human cancers. THZ1 is a small-molecule covalent inhibitor of CDK7 that blocks *MYC/MYCN* transcription in *MYC/MYCN*-amplified cells by irreversibly inhibiting CDK7. The unique SE landscape of *MYCN*-amplified cells also determines their sensitivity to THZ1, which protects normal cells from toxicity (Chipumuro et al., 2014). THZ1 also exerts a prominent anticancer effect on HCC and NPC (Yuan et al., 2017; Tsang et al., 2019). The

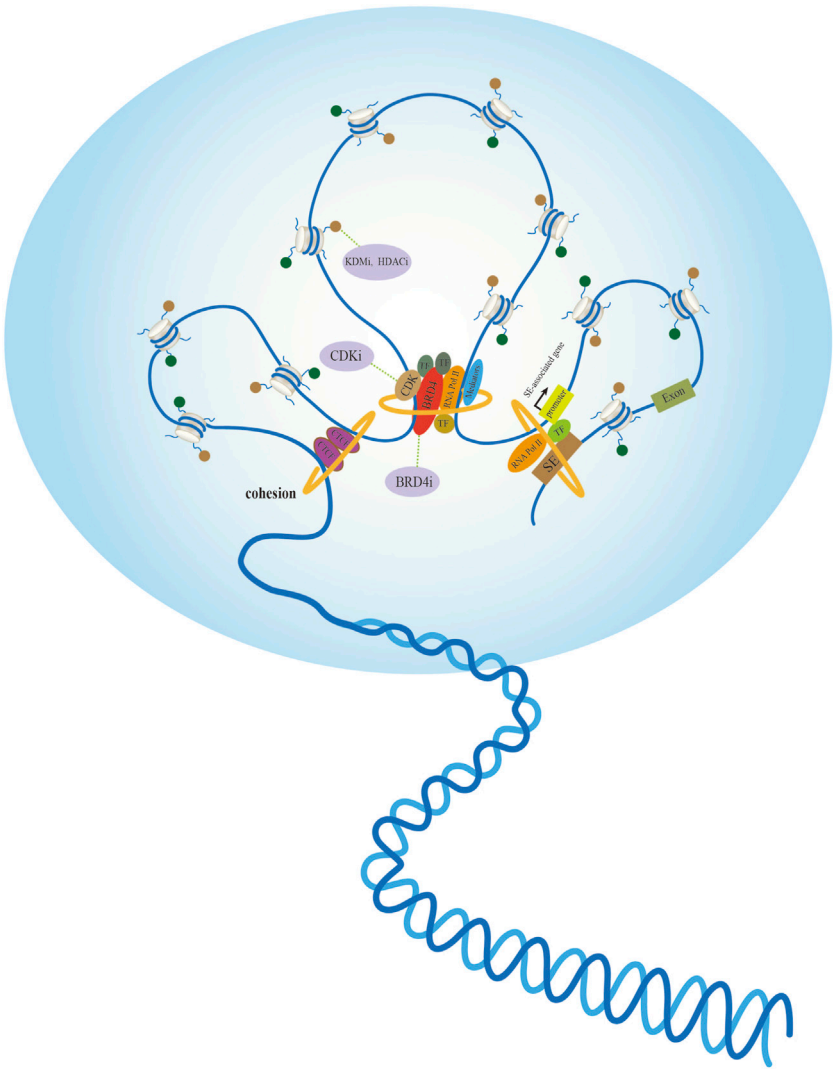


FIGURE 2 | Active chromatin hubs of tumor nuclear morphology and potential anticancer targets. Left: The internal structure of chromatin loop formed by spatial contacts in CTCF binding sites. Middle: Multiple proteins containing transcription factors (TFs) recruit mediators and RNA polymerase II (RNA Pol II) participates in nuclear transcription *via* different mechanisms. Small-molecule inhibitors exert anticancer effects by targeting tumor-promoting proteins or histone modifications. Right: Spatial dimension of SE-associated gene regulation in a gene-specific manner, transcription factor (TFs) binding to super-enhancers (SE) facilitates interaction with promoters with large genomic distances.

bromodomain and extra-terminal domain (BET) family protein BRD4 has been recognized as a general regulator that couples the acetylation state of chromatin with Pol II elongation (Jang et al., 2005). Transcriptional dysregulation of BRD4 promotes the transcriptional activation of specific downstream targets that promote malignancies (Yang et al., 2005; Muhar et al., 2018). BRD4 is closely associated with tumorigenesis and has shown therapeutic potential in preclinical models (Shi and Vakoc, 2014; Wu et al., 2019). JQ1 is a small-molecule BRD4 inhibitor that targets the acetyllysine-recognition domain (bromodomain) of a putative coactivator involved in transcription initiation and elongation to repress *MYC* transcriptional function by the competitive displacement of chromatin-bound coactivators. Bromodomain inhibitors may be an ideal model system for agent mechanism and translational research on *MYC* pathway inhibitors (Delmore et al., 2011; Donati et al., 2018). Surprisingly, dihydroergotamine (DHE), an *NR4A*-induced drug, showed similar efficacy as JQ1 in inhibiting SE-dependent *MYC* transcription and AML growth in mouse xenografts (Call et al., 2020). It implies that DHE is a promising alternative therapeutic strategy for BET inhibitors in AML. These small-molecule inhibitors provide novel therapeutic strategies for specific malignant diseases. However, the limitations of poor prognosis and the emergence of drug resistance render their therapeutic effects unsatisfactory. Notably, combinatorial therapy with BRD4i and histone deacetylase inhibitors (HDACi) showed strong synergy in reducing tumor burden and inhibiting tumor progression (Mazur et al., 2015). Combining JQ1 and THZ1 in

treating head and neck squamous cell carcinoma (HNSCC) effectively inhibited tumor growth and reduced toxicity and drug resistance, resulting in a better prognosis for patients (Zhang W. et al., 2020). A synergistic effect was also observed coupling BRD4 inhibitors (BRD4i) and CDK inhibitors (CDKi) in the treatment of medulloblastoma (Bandopadhayay et al., 2019).

As multiple cancer subtypes are rapidly emerging, epigenetic modulators of specific modifications and small-molecule inhibitors of tumor-promoting factors have become entailing hallmarks. The combined inhibition of these regulatory proteins is an alternative therapeutic strategy for cancer clinics. Whether epigenetic alterations and transcriptional regulations are the cause or the result of altered cellular states, they have potential value as biomarkers for disease diagnosis or as targets for therapeutic intervention. In the long run, systematic interrogation of cancer entities and pathologies of aberrant chromatin folding will uncover new vulnerabilities and novel therapeutic targets in personalized therapy.

CONCLUSION

The integrity of the 3D hierarchical structure of chromosome entities throughout the life cycle of human cells is important for the proper deployment of cell-type-specific gene expression programs. Abnormalities in chromosomal integrity and structure, such as aberrant chromatin folding, compartment conversions, disruption of TAD boundaries, and rewiring of promoter-enhancer interactions generally lead to malignant transformation *via* dysregulated gene expression. The interplay between transcription and genome conformation is the driving force behind cell fate determination, and 3D genome structure plays a critical role in characterizing cancer, thus having profound clinical implications. With the deepening of research on the higher-order chromatin structure of tumor cells, we might gain a more comprehensive understanding of the pathophysiology of carcinogenesis, ultimately promoting the development of clinical cancer treatment.

AUTHOR CONTRIBUTIONS

AP wrote the manuscript. WP, RW, and HZ performed the design of the review. YS and XY were responsible for document retrieval. RW prepared the tables. AP and WP prepared the figures. YS and XY read and revised the manuscript. All authors have read and approved the final manuscript.

FUNDING

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (2020A1515110057), the Xiangshan Talented Scientific Research Foundation of Zhuhai People's Hospital (2020XSYC-07, 2022XSYC-02), and the National Natural Science Foundation of China (32100561, 32100447).

ACKNOWLEDGMENTS

We are grateful to Editage (www.editage.com) for language editing.

REFERENCES

Anania, C., and Lupiáñez, D. G. (2020). Order and Disorder: Abnormal 3D Chromatin Organization in Human Disease. *Brief. Funct. Genomics* 19 (2), 128–138. doi:10.1093/bfgp/elz028

Audia, J. E., and Campbell, R. M. (2016). Histone Modifications and Cancer. *Cold Spring Harb. Perspect. Biol.* 8 (4), a019521. doi:10.1101/cshperspect.a019521

Bandopadhayay, P., Piccioni, F., O'Rourke, R., Ho, P., Gonzalez, E. M., and Buchan, G. (2019). Neuronal Differentiation and Cell-Cycle Programs Mediate Response to BET-Bromodomain Inhibition in MYC-Driven Medulloblastoma. *Nat. Commun.* 10 (1), 2400. doi:10.1038/s41467-019-10307-9

Barutcu, A. R., Lajoie, B. R., McCord, R. P., Tye, C. E., Hong, D., Messier, T. L., et al. (2015). Chromatin Interaction Analysis Reveals Changes in Small Chromosome and Telomere Clustering between Epithelial and Breast Cancer Cells. *Genome Biol.* 16, 214. doi:10.1186/s13059-015-0768-0

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell.* 171 (3), 557–572. e524. doi:10.1016/j.cell.2017.09.043

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-wide. *Curr. Protoc. Mol. Biol.* 109, 21 29 21–2129. doi:10.1002/0471142727.mb2129s109 29

Call, S. G., Duren, R. P., Panigrahi, A. K., Nguyen, L., Freire, P. R., Grimm, S. L., et al. (2020). Targeting Oncogenic Super Enhancers in MYC-dependent AML Using a Small Molecule Activator of NR4A Nuclear Receptors. *Sci. Rep.* 10 (1), 2851. doi:10.1038/s41598-020-59469-3

Chen, D., Zhao, Z., Huang, Z., Chen, D. C., Zhu, X. X., Wang, Y. Z., et al. (2018). Super Enhancer Inhibitors Suppress MYC Driven Transcriptional Amplification and Tumor Progression in Osteosarcoma. *Bone Res.* 6, 11. doi:10.1038/s41413-018-0009-8

Chipumuro, E., Marco, E., Christensen, C. L., Kwiatkowski, N., Zhang, T., Hatheway, C. M., et al. (2014). CDK7 Inhibition Suppresses Super-enhancer-linked Oncogenic Transcription in MYCN-Driven Cancer. *Cell.* 159 (5), 1126–1139. doi:10.1016/j.cell.2014.10.024

Cremer, T., Cremer, C., Baumann, H., Luedtke, E. K., Sperling, K., Teuber, V., et al. (1982). Rabl's Model of the Interphase Chromosome Arrangement Tested in Chinese Hamster Cells by Premature Chromosome Condensation and Laser-UV-Microbeam Experiments. *Hum. Genet.* 60 (1), 46–56. doi:10.1007/BF00281263

Cremer, T., and Cremer, C. (2006b). Rise, Fall and Resurrection of Chromosome Territories: a Historical Perspective. Part II. Fall and Resurrection of Chromosome Territories during the 1950s to 1980s. Part III. Chromosome Territories and the Functional Nuclear Architecture: Experiments and Models from the 1990s to the Present. *Eur. J. Histochem* 50 (4), 223–272.

Cremer, T., and Cremer, C. (2006a). Rise, Fall and Resurrection of Chromosome Territories: a Historical Perspective. Part I. The Rise of Chromosome Territories. *Eur. J. Histochem* 50 (3), 161–176.

Cremer, T., and Cremer, M. (2010). Chromosome Territories. *Cold Spring Harb. Perspect. Biol.* 2 (3), a003889. doi:10.1101/cshperspect.a003889

de Souza, N. (2015). Genomics. Micro-C Maps of Genome Structure. *Nat. Methods* 12 (9), 812. doi:10.1038/nmeth.3575

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the Three-Dimensional Organization of Genomes: Interpreting Chromatin Interaction Data. *Nat. Rev. Genet.* 14 (6), 390–403. doi:10.1038/nrg3454

Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. *Cell.* 164 (6), 1110–1121. doi:10.1016/j.cell.2016.02.007

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. *Science* 295 (5558), 1306–1311. doi:10.1126/science.1067799

Delmore, J. E., Issa, G. C., Lemieux, M. E., Rahl, P. B., Shi, J., Jacobs, H. M., et al. (2011). BET Bromodomain Inhibition as a Therapeutic Strategy to Target C-Myc. *Cell.* 146 (6), 904–917. doi:10.1016/j.cell.2011.08.017

Dixon, J. R., Gorkin, D. U., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell.* 62 (5), 668–680. doi:10.1016/j.molcel.2016.05.018

Dixon, J. R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V. T., et al. (2018). Integrative Detection and Analysis of Structural Variation in Cancer Genomes. *Nat. Genet.* 50 (10), 1388–1398. doi:10.1038/s41588-018-0195-8

Donati, B., Lorenzini, E., and Ciarrocchi, A. (2018). BRD4 and Cancer: Going beyond Transcriptional Regulation. *Mol. Cancer* 17 (1), 164. doi:10.1186/s12943-018-0915-9

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a Massively Parallel Solution for Mapping Interactions between Genomic Elements. *Genome Res.* 16 (10), 1299–1309. doi:10.1101/gr.5571506

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., et al. (2016). Insulator Dysfunction and Oncogene Activation in IDH Mutant Gliomas. *Nature* 529 (7584), 110–114. doi:10.1038/nature16490

Fortin, J. P., and Hansen, K. D. (2015). Reconstructing A/B Compartments as Revealed by Hi-C Using Long-Range Correlations in Epigenetic Data. *Genome Biol.* 16, 180. doi:10.1186/s13059-015-0741-y

Gorkin, D. U., Leung, D., and Ren, B. (2014). The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell. Stem Cell.* 14 (6), 762–775. doi:10.1016/j.stem.2014.05.017

Groschel, S., Sanders, M. A., Hoogenboezem, R., de Wit, E., Bouwman, B. A. M., Erpelinck, C., et al. (2014). A Single Oncogenic Enhancer Rearrangement Causes Concomitant EVI1 and GATA2 Deregulation in Leukemia. *Cell.* 157 (2), 369–381. doi:10.1016/j.cell.2014.02.019

Gupta, R., Janostiak, R., and Wajapeyee, N. (2020). Transcriptional Regulators and Alterations that Drive Melanoma Initiation and Progression. *Oncogene* 39 (48), 7093–7105. doi:10.1038/s41388-020-01490-x

Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., et al. (2011). CTCF-mediated Functional Chromatin Interactome in Pluripotent Cells. *Nat. Genet.* 43 (7), 630–638. doi:10.1038/ng.857

Harbers, L., Agostini, F., Nicos, M., Poddighe, D., Bienko, M., and Crosetto, N. (2021). Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data from the Cancer Genome Atlas. *Front. Oncol.* 11, 700568. doi:10.3389/fonc.2021.700568

Hildebrand, E. M., and Dekker, J. (2020). Mechanisms and Functions of Chromosome Compartmentalization. *Trends Biochem. Sci.* 45 (5), 385–396. doi:10.1016/j.tibs.2020.01.002

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A. L., Bak, R. O., Li, C. H., et al. (2016). Activation of Proto-Oncogenes by Disruption of Chromosome Neighborhoods. *Science* 351 (6280), 1454–1458. doi:10.1126/science.aad9024

Hoffmann, I., Roatsch, M., Schmitt, M. L., Carlino, L., Pippel, M., Sippl, W., et al. (2012). The Role of Histone Demethylases in Cancer Therapy. *Mol. Oncol.* 6 (6), 683–703. doi:10.1016/j.molonc.2012.07.004

Huang, H., Hu, J., Maryam, A., Huang, Q., Zhang, Y., Ramakrishnan, S., et al. (2021). Defining Super-enhancer Landscape in Triple-Negative Breast Cancer by Multiomic Profiling. *Nat. Commun.* 12 (1), 2242. doi:10.1038/s41467-021-22445-0

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., et al. (2014). Analysis of Hundreds of Cis-Regulatory Landscapes at High Resolution in a Single, High-Throughput Experiment. *Nat. Genet.* 46 (2), 205–212. doi:10.1038/ng.2871

Jang, M. K., Mochizuki, K., Zhou, M., Jeong, H. S., Brady, J. N., and Ozato, K. (2005). The Bromodomain Protein Brd4 Is a Positive Regulatory Component of P-TEFb and Stimulates RNA Polymerase II-dependent Transcription. *Mol. Cell.* 19 (4), 523–534. doi:10.1016/j.molcel.2005.06.027

Jiang, S., Zhou, H., Liang, J., Gerdt, C., Wang, C., Ke, L., et al. (2017a). The Epstein-Barr Virus Regulome in Lymphoblastoid Cells. *Cell. Host Microbe* 22 (4), 561–573. e564. doi:10.1016/j.chom.2017.09.001

Jiang, Y. Y., Lin, D. C., Mayakonda, A., Hazawa, M., Ding, L. W., Chien, W. W., et al. (2017b). Targeting Super-enhancer-associated Oncogenes in Oesophageal Squamous Cell Carcinoma. *Gut* 66 (8), 1358–1368. doi:10.1136/gutjnl-2016-311818

Johnston, M. J., Nikolic, A., Ninkovic, N., Guilhamon, P., Cavalli, F. M. G., Seaman, S., et al. (2019). High-resolution Structural Genomics Reveals New Therapeutic Vulnerabilities in Glioblastoma. *Genome Res.* 29 (8), 1211–1222. doi:10.1101/gr.246520.118

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., et al. (2007). Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell.* 128 (6), 1231–1245. doi:10.1016/j.cell.2006.12.048

Kloetgen, A., Thandapani, P., Ntziachristos, P., Ghebrechristos, Y., Nomikou, S., Lazaris, C., et al. (2020). Three-dimensional Chromatin Landscapes in T Cell Acute Lymphoblastic Leukemia. *Nat. Genet.* 52 (4), 388–400. doi:10.1038/s41588-020-0602-9

Kocanova, S., Goiffon, I., and Bystricky, K. (2018). 3D FISH to Analyse Gene Domain-specific Chromatin Re-modeling in Human Cancer Cell Lines. *Methods* 142, 3–15. doi:10.1016/j.ymeth.2018.02.013

Kubiak, M., Jurek, A., Kaminska, K., Kowalewski, J., Huang, S., and Lewandowska, M. A. (2019). Chromosome Conformation Capture Reveals Two Elements that Interact with the PTBP3 (ROD1) Transcription Start Site. *Int. J. Mol. Sci.* 20 (2). doi:10.3390/ijms20020242

Kulasinghe, A., Lim, Y., Kapeleris, J., Warkiani, M., O'Byrne, K., and Punyadeera, C. (2020). The Use of Three-Dimensional DNA Fluorescent *In Situ* Hybridization (3D DNA FISH) for the Detection of Anaplastic Lymphoma Kinase (ALK) in Non-small Cell Lung Cancer (NSCLC) Circulating Tumor Cells. *Cells* 9 (6). doi:10.3390/cells9061465

Larochelle, S., Amat, R., Glover-Cutter, K., Sanso, M., Zhang, C., Allen, J. J., et al. (2012). Cyclin-dependent Kinase Control of the Initiation-To-Elongation Switch of RNA Polymerase II. *Nat. Struct. Mol. Biol.* 19 (11), 1108–1115. doi:10.1038/nsmb.2399

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., et al. (2003). A Long-Range Shh Enhancer Regulates Expression in the Developing Limb and Fin and Is Associated with Preaxial Polydactyly. *Hum. Mol. Genet.* 12 (14), 1725–1735. doi:10.1093/hmg/ddg180

Li, X., Luo, O. J., Wang, P., Zheng, M., Wang, D., Piecuch, E., et al. (2017). Long-read ChIA-PET for Base-Pair-Resolution Mapping of Haplotype-specific Chromatin Interactions. *Nat. Protoc.* 12 (5), 899–915. doi:10.1038/nprot.2017.012

Li, Y., Liao, Z., Luo, H., Benyoucef, A., Kang, Y., Lai, Q., et al. (2020). Alteration of CTCF-Associated Chromatin Neighborhood Inhibits TAL1-Driven Oncogenic Transcription Program and Leukemogenesis. *Nucleic Acids Res.* 48 (6), 3119–3133. doi:10.1093/nar/gkaa098

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326 (5950), 289–293. doi:10.1126/science.1181369

Liu, Y., Ao, X., Ding, W., Ponnusamy, M., Wu, W., Hao, X., et al. (2018). Critical Role of FOXO3a in Carcinogenesis. *Mol. Cancer* 17 (1), 104. doi:10.1186/s12943-018-0856-3

Loven, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., et al. (2013). Selective Inhibition of Tumor Oncogenes by Disruption of Super-enhancers. *Cell.* 153 (2), 320–334. doi:10.1016/j.cell.2013.03.036

Lupianez, D. G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends Genet.* 32 (4), 225–237. doi:10.1016/j.tig.2016.01.003

Ma, Y., McKay, D. J., and Buttitta, L. (2019). Changes in Chromatin Accessibility Ensure Robust Cell Cycle Exit in Terminally Differentiated Cells. *PLoS Biol.* 17 (9), e3000378. doi:10.1371/journal.pbio.3000378

Mazur, P. K., Herner, A., Mello, S. S., Wirth, M., Hausmann, S., Sanchez-Rivera, F. J., et al. (2015). Combined Inhibition of BET Family Proteins and Histone Deacetylases as a Potential Epigenetics-Based Therapy for Pancreatic Ductal Adenocarcinoma. *Nat. Med.* 21 (10), 1163–1171. doi:10.1038/nm.3952

McAllister, T. E., England, K. S., Hopkinson, R. J., Brennan, P. E., Kawamura, A., and Schofield, C. J. (2016). Recent Progress in Histone Demethylase Inhibitors. *J. Med. Chem.* 59 (4), 1308–1329. doi:10.1021/acs.jmedchem.5b01758

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., et al. (2011). A Regulatory Archipelago Controls Hox Genes Transcription in Digits. *Cell.* 147 (5), 1132–1145. doi:10.1016/j.cell.2011.10.023

Mora, A., Sandve, G. K., Gabrielsen, O. S., and Eskeland, R. (2016). In the Loop: Promoter-Enhancer Interactions and Bioinformatics. *Brief. Bioinform* 17 (6), 980–995. doi:10.1093/bib/bbv097

Mourad, R., Hsu, P. Y., Juan, L., Shen, C., Koneru, P., Lin, H., et al. (2014). Estrogen Induces Global Reorganization of Chromatin Structure in Human Breast Cancer Cells. *PLoS One* 9 (12), e113354. doi:10.1371/journal.pone.0113354

Muhar, M., Ebert, A., Neumann, T., Umkehrer, C., Jude, J., Wieshofer, C., et al. (2018). SLAM-seq Defines Direct Gene-Regulatory Functions of the BRD4-MYC axis. *Science* 360 (6390), 800–805. doi:10.1126/science.aao2793

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., et al. (2016). HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture. *Nat. Methods* 13 (11), 919–922. doi:10.1038/nmeth.3999

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., et al. (2012). Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre. *Nature* 485 (7398), 381–385. doi:10.1038/nature11049

Northcott, P. A., Lee, C., Zichner, T., Stutz, A. M., Erkek, S., Kawauchi, D., et al. (2014). Enhancer Hijacking Activates GFI1 Family Oncogenes in Medulloblastoma. *Nature* 511 (7510), 428–434. doi:10.1038/nature13379

Nowell, P. C., Hungerford, D. A., and Nowell, P. C. (1960). A Minute Chromosome in Human Chronic Granulocytic Leukemia. *Science* 132, 1497.

Oehme, I., Deubzer, H. E., Wegener, D., Pickert, D., Linke, J. P., Hero, B., et al. (2009). Histone Deacetylase 8 in Neuroblastoma Tumorigenesis. *Clin. Cancer Res.* 15 (1), 91–99. doi:10.1158/1078-0432.CCR-08-0684

Park, P. J. (2009). ChIP-seq: Advantages and Challenges of a Maturing Technology. *Nat. Rev. Genet.* 10 (10), 669–680. doi:10.1038/nrg2641

Peng, Y., and Zhang, Y. (2018). Enhancer and Super-enhancer: Positive Regulators in Gene Transcription. *Anim. Model. Exp. Med.* 1 (3), 169–179. doi:10.1002/ame2.12032

Pott, S., and Lieb, J. D. (2015). What Are Super-enhancers? *Nat. Genet.* 47 (1), 8–12. doi:10.1038/ng.3167

Prieto, E. I., and Maeshima, K. (2019). Dynamic Chromatin Organization in the Cell. *Essays Biochem.* 63 (1), 133–145. doi:10.1042/EBC20180054

Proudhon, C., Hao, B., Raviram, R., Chaumeil, J., and Skok, J. A. (2015). Long-Range Regulation of V(D)J Recombination. *Adv. Immunol.* 128, 123–182. doi:10.1016/bs.ai.2015.07.003

Ramani, V., Cusanovich, D. A., Hause, R. J., Ma, W., Qiu, R., Deng, X., et al. (2016). Mapping 3D Genome Architecture through *In Situ* DNase Hi-C. *Nat. Protoc.* 11 (11), 2104–2121. doi:10.1038/nprot.2016.126

Ramsden, D. A., and Nussenzweig, A. (2021). Mechanisms Driving Chromosomal Translocations: Lost in Time and Space. *Oncogene* 40 (25), 4263–4270. doi:10.1038/s41388-021-01856-9

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell.* 159 (7), 1665–1680. doi:10.1016/j.cell.2014.11.021

Rennie, S., Dalby, M., van Duin, L., and Andersson, R. (2018). Transcriptional Decomposition Reveals Active Chromatin Architectures and Cell Specific Regulatory Interactions. *Nat. Commun.* 9 (1), 487. doi:10.1038/s41467-017-02798-1

Sati, S., and Cavalli, G. (2017). Chromosome Conformation Capture Technologies and Their Impact in Understanding Genome Function. *Chromosoma* 126 (1), 33–44. doi:10.1007/s00412-016-0593-6

Sengupta, S., and George, R. E. (2017). Super-Enhancer-Driven Transcriptional Dependencies in Cancer. *Trends Cancer* 3 (4), 269–281. doi:10.1016/j.trecan.2017.03.006

Shi, J., and Vakoc, C. R. (2014). The Mechanisms behind the Therapeutic Activity of BET Bromodomain Inhibition. *Mol. Cell.* 54 (5), 728–736. doi:10.1016/j.molcel.2014.05.016

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., et al. (2006). Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture-On-Chip (4C). *Nat. Genet.* 38 (11), 1348–1354. doi:10.1038/ng1896

Solovei, I., Cavallo, A., Schermelleh, L., Jaunin, F., Scasselati, C., Cmarko, D., et al. (2002). Spatial Preservation of Nuclear Chromatin Architecture during Three-Dimensional Fluorescence *In Situ* Hybridization (3D-FISH). *Exp. Cell. Res.* 276 (1), 10–23. doi:10.1006/excr.2002.5513

Szabo, Q., Bantignies, F., and Cavalli, G. (2019). Principles of Genome Folding into Topologically Associating Domains. *Sci. Adv.* 5 (4), eaaw1668. doi:10.1126/sciadv.aaw1668

Taberlay, P. C., Achinger-Kawecka, J., Lun, A. T., Buske, F. A., Sabir, K., Gould, C. M., et al. (2016). Three-dimensional Disorganization of the Cancer Genome Occurs Coincident with Long-Range Genetic and Epigenetic Alterations. *Genome Res.* 26 (6), 719–731. doi:10.1101/gr.201517.115

Tang, F., Yang, Z., Tan, Y., and Li, Y. (2020). Super-enhancer Function and its Application in Cancer Targeted Therapy. *NPJ Precis. Oncol.* 4, 2. doi:10.1038/s41698-020-0108-z

Teixeira, M. R., and Heim, S. (2005). Multiple Numerical Chromosome Aberrations in Cancer: what Are Their Causes and what Are Their Consequences? *Semin. Cancer Biol.* 15 (1), 3–12. doi:10.1016/j.semcancer.2004.09.006

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., et al. (2012). The Accessible Chromatin Landscape of the Human Genome. *Nature* 489 (7414), 75–82. doi:10.1038/nature11232

Tsang, F. H., Law, C. T., Tang, T. C., Cheng, C. L., Chin, D. W., Tam, W. V., et al. (2019). Aberrant Super-enhancer Landscape in Human Hepatocellular Carcinoma. *Hepatology* 69 (6), 2502–2517. doi:10.1002/hep.30544

Ulianov, S. V., Khrameeva, E. E., Gavrilov, A. A., Flyamer, I. M., Kos, P., Mikhaleva, E. A., et al. (2016). Active Chromatin and Transcription Play a Key Role in Chromosome Partitioning into Topologically Associating Domains. *Genome Res.* 26 (1), 70–84. doi:10.1101/gr.196006.115

Vicente-Duenas, C., Romero-Camarero, I., Cobaleda, C., and Sanchez-Garcia, I. (2013). Function of Oncogenes in Cancer Development: a Changing Paradigm. *EMBO J.* 32 (11), 1502–1513. doi:10.1038/emboj.2013.97

Vietri Rudan, M., and Hadjur, S. (2015). Genetic Tailors: CTCF and Cohesin Shape the Genome during Evolution. *Trends Genet.* 31 (11), 651–660. doi:10.1016/j.tig.2015.09.004

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., et al. (2003). The Evolution of Transcriptional Regulation in Eukaryotes. *Mol. Biol. Evol.* 20 (9), 1377–1419. doi:10.1093/molbev/msg140

Wu, P., Li, T., Li, R., Jia, L., Zhu, P., Liu, Y., et al. (2017). 3D Genome of Multiple Myeloma Reveals Spatial Genome Disorganization Associated with Copy Number Variations. *Nat. Commun.* 8 (1), 1937. doi:10.1038/s41467-017-01793-w

Wu, Y., Wang, Y., Diao, P., Zhang, W., Li, J., Ge, H., et al. (2019). Therapeutic Targeting of BRD4 in Head Neck Squamous Cell Carcinoma. *Theranostics* 9 (6), 1777–1793. doi:10.7150/thno.31581

Xi, W., and Beer, M. A. (2021). Loop Competition and Extrusion Model Predicts CTCF Interaction Specificity. *Nat. Commun.* 12 (1), 1046. doi:10.1038/s41467-021-21368-0

Xiang, J. F., Yin, Q. F., Chen, T., Zhang, Y., Zhang, X. O., Wu, Z., et al. (2014). Human Colorectal Cancer-specific CCAT1-L lncRNA Regulates Long-Range Chromatin Interactions at the MYC Locus. *Cell. Res.* 24 (5), 513–531. doi:10.1038/cr.2014.35

Xu, J., Ma, H., Ma, H., Jiang, W., Mela, C. A., Duan, M., et al. (2020). Super-resolution Imaging Reveals the Evolution of Higher-Order Chromatin Folding in Early Carcinogenesis. *Nat. Commun.* 11 (1), 1899. doi:10.1038/s41467-020-15718-7

Yang, M., Vesterlund, M., Siavelis, I., Moura-Castro, L. H., Castor, A., Fioretos, T., et al. (2019). Proteogenomics and Hi-C Reveal Transcriptional Dysregulation in High Hyperdiploid Childhood Acute Lymphoblastic Leukemia. *Nat. Commun.* 10 (1), 1519. doi:10.1038/s41467-019-09469-3

Yang, Z., Yik, J. H., Chen, R., He, N., Jang, M. K., Ozato, K., et al. (2005). Recruitment of P-TEFb for Stimulation of Transcriptional Elongation by the Bromodomain Protein Brd4. *Mol. Cell.* 19 (4), 535–545. doi:10.1016/j.molcel.2005.06.029

Yuan, J., Jiang, Y. Y., Mayakonda, A., Huang, M., Ding, L. W., Lin, H., et al. (2017). Super-Enhancers Promote Transcriptional Dysregulation in Nasopharyngeal Carcinoma. *Cancer Res.* 77 (23), 6614–6626. doi:10.1158/0008-5472.CAN-17-1143

Zanotti, S., Vanhauwaert, S., Van Neste, C., Olexiouk, V., Van Laere, J., Verschuuren, M., et al. (2021). MYCN-induced Nucleolar Stress Drives an

Early Senescence-like Transcriptional Program in hTERT-Immortalized RPE Cells. *Sci. Rep.* 11 (1), 14454. doi:10.1038/s41598-021-93863-9

Zhang, T., Song, X., Zhang, Z., Mao, Q., Xia, W., Xu, L., et al. (2020a). Aberrant Super-enhancer Landscape Reveals Core Transcriptional Regulatory Circuitry in Lung Adenocarcinoma. *Oncogenesis* 9 (10), 92. doi:10.1038/s41389-020-00277-9

Zhang, W., Ge, H., Jiang, Y., Huang, R., Wu, Y., Wang, D., et al. (2020b). Combinational Therapeutic Targeting of BRD4 and CDK7 Synergistically Induces Anticancer Effects in Head and Neck Squamous Cell Carcinoma. *Cancer Lett.* 469, 510–523. doi:10.1016/j.canlet.2019.11.027

Zhang, Y., McCord, R. P., Ho, Y. J., Lajoie, B. R., Hildebrand, D. G., Simon, A. C., et al. (2012). Spatial Organization of the Mouse Genome and its Role in Recurrent Chromosomal Translocations. *Cell.* 148 (5), 908–921. doi:10.1016/j.cell.2012.02.002

Zhang, Y., Smith, C. L., Saha, A., Grill, S. W., Mihardja, S., Smith, S. B., et al. (2006). DNA Translocation and Loop Formation Mechanism of Chromatin Remodeling by SWI/SNF and RSC. *Mol. Cell.* 24 (4), 559–568. doi:10.1016/j.molcel.2006.10.025

Zhou, Y., Gerrard, D. L., Wang, J., Li, T., Yang, Y., Fritz, A. J., et al. (2019). Temporal Dynamic Reorganization of 3D Chromatin Architecture in Hormone-Induced Breast Cancer and Endocrine Resistance. *Nat. Commun.* 10 (1), 1522. doi:10.1038/s41467-019-09320-9

Zhu, P., and Li, G. (2016). Structural Insights of Nucleosome and the 30-nm Chromatin Fiber. *Curr. Opin. Struct. Biol.* 36, 106–115. doi:10.1016/j.sbi.2016.01.013

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Peng, Peng, Wang, Zhao, Yu and Sun. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

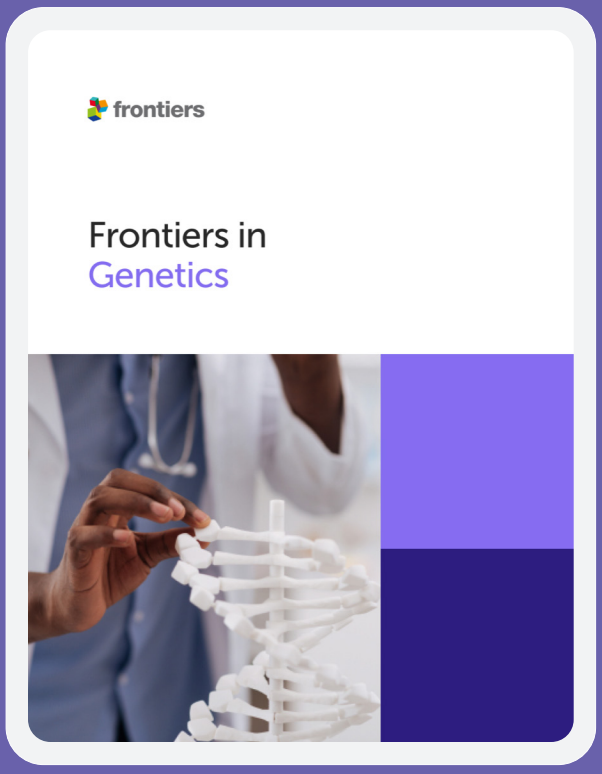
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact