FUNCTIONAL ANNOTATION OF FARM ANIMAL GENOMES

EDITED BY: Hans Cheng, Amanda Jane Chamberlain, Elisabetta Giuffra, Christa Kuehn, Christopher K. Tuggle and Huaijun Zhou

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

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 978-2-88971-829-0
DOI 10.3389/978-2-88971-829-0

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

FUNCTIONAL ANNOTATION OF FARM ANIMAL GENOMES

Topic Editors:
Hans Cheng, Agricultural Research Service, United States
Amanda Jane Chamberlain, Agriculture Victoria, Australia
Elisabetta Giuffra, University Paris-Saclay, INRAE, France
Christa Kuehn, Research Institute for Farm Animal Biology (FBN), Germany
Christopher K. Tuggle, Iowa State University, United States
Huaijun Zhou, University of California, Davis, United States

Citation: Cheng, H., Chamberlain, A. J., Giuffra, E., Kuehn, C., Tuggle, C. K., Zhou, H., eds. (2021). Functional Annotation of Farm Animal Genomes. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-829-0

Table of Contents

06 ***Editorial: Functional Annotation of Animal Genomes***
Amanda Jane Chamberlain, Hans H. Cheng, Elisabetta Giuffra, Christa Kuehn, Christopher K. Tuggle and Huaijun Zhou

09 ***Global Analysis of Transcription Start Sites in the New Ovine Reference Genome (*Oar rambouillet v1.0*)***
Mazdak Salavati, Alex Caulton, Richard Clark, Iveta Gazova, Timothy P. L. Smith, Kim C. Worley, Noelle E. Cockett, Alan L. Archibald, Shannon M. Clarke, Brenda M. Murdoch, Emily L. Clark on behalf of The Ovine FAANG Project Consortium

30 ***Genome-Wide Histone Modifications and CTCF Enrichment Predict Gene Expression in Sheep Macrophages***
Alisha T. Massa, Michelle R. Mousel, Maria K. Herndon, David R. Herndon, Brenda M. Murdoch and Stephen N. White

50 ***Identification of a Goat Intersexuality-Associated Novel Variant Through Genome-Wide Resequencing and Hi-C***
Guang-Xin E, Dong-Ke Zhou, Zhu-Qing Zheng, Bai-Gao Yang, Xiang-Long Li, Lan-Hui Li, Rong-Yan Zhou, Wen-Hui Nai, Xun-Ping Jiang, Jia-Hua Zhang, Qiong-Hua Hong, Yue-Hui Ma, Ming-Xing Chu, Hui-Jiang Gao, Yong-Ju Zhao, Xing-Hai Duan, Yong-Meng He, Ri-Su Na, Yan-Guo Han, Yan Zeng, Yu Jiang and Yong-Fu Huang

61 ***Generation of a Biobank From Two Adult Thoroughbred Stallions for the Functional Annotation of Animal Genomes Initiative***
Callum G. Donnelly, Rebecca R. Bellone, Erin N. Hales, Annee Nguyen, Scott A. Katzman, Ghislaine A. Dujovne, Kelly E. Knickelbein, Felipe Avila, Ted S. Kalbfleisch, Elena Giulotto, Nicole B. Kingsley, Jocelyn Tanaka, Elizabeth Esdaile, Sichong Peng, Anna Dahlgren, Anna Fuller, Michael J. Mienaltowski, Terje Raudsepp, Verena K. Affolter, Jessica L. Petersen and Carrie J. Finno

67 ***"Adopt-a-Tissue" Initiative Advances Efforts to Identify Tissue-Specific Histone Marks in the Mare***
N. B. Kingsley, Natasha A. Hamilton, Gabriella Lindgren, Ludovic Orlando, Ernie Bailey, Samantha Brooks, Molly McCue, T. S. Kalbfleisch, James N. MacLeod, Jessica L. Petersen, Carrie J. Finno and Rebecca R. Bellone

76 ***A de novo Full-Length mRNA Transcriptome Generated From Hybrid-Corrected PacBio Long-Reads Improves the Transcript Annotation and Identifies Thousands of Novel Splice Variants in Atlantic Salmon***
Sigmund Ramberg, Bjørn Høyheim, Tone-Kari Knutsdatter Østbye and Rune Andreassen

98 ***Accessing Livestock Resources in Ensembl***
Fergal J. Martin, Astrid Gall, Michal Szpak and Paul Flicek

118 ***Network Analyses Predict Small RNAs That Might Modulate Gene Expression in the Testis and Epididymis of* Bos indicus *Bulls***
Andressa O. de Lima, Juliana Afonso, Janette Edson, Esteban Marcellin, Robin Palfreyman, Laercio R. Porto-Neto, Antonio Reverter and Marina R. S. Fortes

132 ***Comparative Analysis of the Circular Transcriptome in Muscle, Liver, and Testis in Three Livestock Species***
Annie Robic, Chloé Cerutti, Christa Kühn and Thomas Faraut

151 ***Large-Scale Multiplexing Permits Full-Length Transcriptome Annotation of 32 Bovine Tissues From a Single Nanopore Flow Cell***
Michelle M. Halstead, Alma Islas-Trejo, Daniel E. Goszczynski, Juan F. Medrano, Huaijun Zhou and Pablo J. Ross

164 ***Characterizing Genetic Regulatory Elements in Ovine Tissues***
Kimberly M. Davenport, Alisha T. Massa, Suraj Bhattarai, Stephanie D. McKay, Michelle R. Mousel, Maria K. Herndon, Stephen N. White, Noelle E. Cockett, Timothy P. L. Smith, Brenda M. Murdoch on behalf of The Ovine FAANG Project Consortium

184 ***Successful ATAC-Seq From Snap-Frozen Equine Tissues***
Sichong Peng, Rebecca Bellone, Jessica L. Petersen, Theodore S. Kalbfleisch and Carrie J. Finno

194 ***The FAANG Data Portal: Global, Open-Access, "FAIR", and Richly Validated Genotype to Phenotype Data for High-Quality Functional Annotation of Animal Genomes***
Peter W. Harrison, Alexey Sokolov, Akshatha Nayak, Jun Fan, Daniel Zerbino, Guy Cochrane and Paul Flicek

202 ***Tissue Resources for the Functional Annotation of Animal Genomes***
Michèle Tixier-Boichard, Stéphane Fabre, Sophie Dhorne-Pollet, Adeline Goubil, Hervé Acloque, Silvia Vincent-Naulleau, Pablo Ross, Ying Wang, Ganrea Chanthavixay, Hans Cheng, Catherine Ernst, Vicki Leesburg, Elisabetta Giuffra and Huaijun Zhou and the Collaborative Working Group

213 ***Putative Causal Variants Are Enriched in Annotated Functional Regions From Six Bovine Tissues***
Claire P. Prowse-Wilkins, Jianghui Wang, Ruidong Xiang, Josie B. Garner, Michael E. Goddard and Amanda J. Chamberlain

230 ***Reference Transcriptomes of Porcine Peripheral Immune Cells Created Through Bulk and Single-Cell RNA Sequencing***
Juber Herrera-Uribe, Jayne E. Wiarda, Sathesh K. Sivasankaran, Lance Daharsh, Haibo Liu, Kristen A. Byrne, Timothy P. L. Smith, Joan K. Lunney, Crystal L. Loving and Christopher K. Tuggle

257 ***RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and* Cis*-Regulation Analysis by Allele-Specific Expression in Livestock Species***
Frédéric Jehl, Fabien Degalez, Maria Bernard, Frédéric Lecerf, Laetitia Lagoutte, Colette Désert, Manon Coulée, Olivier Bouchez, Sophie Leroux, Behnam Abasht, Michèle Tixier-Boichard, Bertrand Bed'hom, Thierry Burlot, David Gourichon, Philippe Bardou, Hervé Acloque, Sylvain Foissac, Sarah Djebali, Elisabetta Giuffra, Tatiana Zerjal, Frédérique Pitel, Christophe Klopp and Sandrine Lagarrigue

274 ***Transcriptomes of an Array of Chicken Ovary, Intestinal, and Immune Cells and Tissues***

Eliah G. Overbey, Theros T. Ng, Pietro Catini, Lisa M. Griggs, Paul Stewart, Suzana Tkalcic, R. David Hawkins and Yvonne Drechsler

292 ***Watch Out for a Second SNP: Focus on Multi-Nucleotide Variants in Coding Regions and Rescued Stop-Gained***

Fabien Degalez, Frédéric Jehl, Kévin Muret, Maria Bernard, Frédéric Lecerf, Laetitia Lagoutte, Colette Désert, Frédérique Pitel, Christophe Klopp and Sandrine Lagarrigue

300 ***PacBio Iso-Seq Improves the Rainbow Trout Genome Annotation and Identifies Alternative Splicing Associated With Economically Important Phenotypes***

Ali Ali, Gary H. Thorgaard and Mohamed Salem

320 ***PacBio Single-Molecule Long-Read Sequencing Provides New Light on the Complexity of Full-Length Transcripts in Cattle***

Tianpeng Chang, Bingxing An, Mang Liang, Xinghai Duan, Lili Du, Wentao Cai, Bo Zhu, Xue Gao, Yan Chen, Lingyang Xu, Lupei Zhang, Huijiang Gao and Junya Li

frontiers in Genetics

EDITORIAL
published: 15 October 2021
doi: 10.3389/fgene.2021.768626

Editorial: Functional Annotation of Animal Genomes

Amanda Jane Chamberlain[1], *Hans H. Cheng*[2]*, *Elisabetta Giuffra*[3], *Christa Kuehn*[4], *Christopher K. Tuggle*[5] *and Huaijun Zhou*[6]

[1]Agriculture Victoria, Agribio, Centre for AgriBiosciences, Bundoora, VIC, Australia, [2]Avian Disease and Oncology Laboratory, Washington D.C., MI, United States, [3]Paris-Saclay University, INRAE, AgroParisTech, GABI, Jouy-en-Josas, Jouy-en-Josas, France, [4]Research Institute for Farm Animal Biology (FBN), Dummerstorf, Germany, [5]Iowa State University, Ames, IA, United States, [6]University of California Davis, Davis, CA, United States

Keywords: genome annotation, farm animals, genomics, sequencing, bioinformatics, resources

Editorial on the Research Topic

Functional Annotation of Farm Animal Genomes

All fields of biology have been greatly influenced by the generation of complete and well-annotated genome assemblies. This impact is most apparent with the findings and resulting applications from the Human Genome Project (HGP), which has transformed biomedical science. The original justification for having a genome assembly was to get a complete "parts list" with the primary goal being the identification and location of all genes. However, it soon became readily apparent that genomes were much more than just sequences that code for proteins; protein-coding regions account for ~1.5% of the human genome and similar results were obtained in analyzing the genomes of domesticated and other farmed animal species. Thus, current efforts have been focused on finding relevant functional elements, such as non-coding elements that regulate when, where, and how much specific genes and/or particular isoforms are expressed.

To address the need for annotation of farm animal genomes, the Functional Annotation of Animal Genomes (FAANG) Consortium was launched in 2015. Like other research consortia, FAANG (www.faang.org) is committed to sharing data rapidly and before publication for the benefit of the whole community (www.faang.org/data-share-principle), with data and metadata (standardized details on samples, laboratory and bioinformatic protocols applied with a comprehensiveness more than current practice) being collected in the FAANG Data Portal (https://data.faang.org/home).

A Research Topic call for papers was made to provide the opportunity to report on the ongoing efforts to annotate farm animal genomes and inform genomic biology. We believed that such a Research Topic would be timely as a historical marker of such efforts, as the pilot FAANG projects were being completed and a number of larger-scale projects are underway in Australia, the United States and Europe. Many groups responded to this call. The Research Topic also offered the opportunity to establish reference-settings for FAANG with respect to methods and protocols. We are pleased that 21 papers, representing eight species as well as two species-agnostic resource reports, are presented in this collection. Below, we summarize reports with complementary themes, focused on one or more of the following topics:

NEW BIOLOGICAL AND BIOINFORMATIC RESOURCES FOR THE COMMUNITY

As a community, it is important to have shared resources to minimize duplication of effort, standardize wet-lab protocols, and consistent and readily-available bioinformatic pipelines. Such

OPEN ACCESS

Edited and reviewed by:
Francisco Peñagaricano, University of Wisconsin-Madison, United States

***Correspondence:**
Hans H. Cheng
hans.cheng@usda.gov

Specialty section:
This article was submitted to Functional Annotation of Animal, a section of the journal Frontiers in Genetics

Received: *31 August 2021*
Accepted: *30 September 2021*
Published: *15 October 2021*

Citation:
Chamberlain AJ, Cheng HH, Giuffra E, Kuehn C, Tuggle CK and Zhou H (2021) Editorial: Functional Annotation of Animal Genomes. Front. Genet. 12:768626. doi: 10.3389/fgene.2021.768626

efforts are a hallmark of the FAANG community from its inception, and several groups reported the completion of different community resources. An European-US effort to describe tissue samples, as well as sample collection protocols and associated metadata across two early FAANG pilot projects, was provided by Tixier-Boichard et al. Several groups reported equine community resources. Donnelley et al. highlighted development of a stallion tissue biobank, a community collaboration to "sponsor" individual tissues to expand epigenetics data for mares was described (Kingsley et al.), and documentation of protocols for measuring chromatin accessible sites using ATAC-seq of equine tissues was contributed (Peng et al.). Two groups outlined FAANG community data resources, including a description of available livestock data and annotation tools at Ensembl (Martin et al.) and an account of the current status and resources available at the FAANG data portal (Harrison et al.).

REFERENCE TRANSCRIPTOMES FOR CATALOGING FUNCTION AND PREDICTING REGULATORY RELATIONSHIPS

Transcriptomic resources are very much akin to the generation of a reference genome assembly, by providing important baseline functional knowledge for highly relevant tissues of each species. Given the continuous improvement in sequencing technologies, it was not surprising that many papers utilized the latest platforms (e.g., long-read sequencing, single cell RNA-seq) to define RNA transcripts and splice variants, as well as chromatin accessibility and epigenetic modifications at RNA-expressing genes. These efforts reflect the breadth of the community in targeting many farmed species spanning fish, birds and mammals for different tissues, developmental stages, and cell types. For example, RNA-seq-based transcriptomes for 10 tissues or isolated cell populations from chickens was summarized by Overbey et al. Iso-seq-based (Chang et al.) and Nanopore-based (Halstead et al.) transcriptomes of a large number of cattle tissues was generated and used to identify full-length transcripts and alternative splicing isoforms, which were often tissue-specifically expressed. Long-read transcriptome technology was also used by Ali et al. on fourteen tissues to improve the annotation of the Rainbow trout genome, as well as identify splice isoforms associated with traits of economic importance in aquaculture. Similar long-read transcriptome analysis of several tissues in Salmon genome annotation was shown to substantially improve the transcript catalog for this species (Ramberg et al.). A new RNA isoform, circular RNA, was cataloged across public and new RNA-seq datasets for three tissues from sheep, cattle, and pig (Robic et al.), and both tissue- and developmental stage-associated differences in abundance of circular RNAs was detected. Further, RNA-seq analysis of eight specific flow-sorted populations of peripheral blood mononuclear cells (PBMC) was compared to single-cell RNA-seq analysis of PBMC by Herrera-Uribe et al. Both datatypes extended annotation for the pig genome, identified co-expressed genes for all major PBMC types in porcine blood, and showed many specific cell types could be matched to human PBMC cell-specific transcriptomes. Finally, co-expression analysis between RNAs and miRNAs across different stages of spermatogenesis was used to predict miRNA regulatory targets in this important process (de Lima et al.).

LARGE SCALE FUNCTIONAL ANNOTATIONS: INSIGHTS FROM THE OVINE AND CAPRINE FAANG PROJECTS

Chromatin accessibility patterns and epigenomic modifications were reported as outcomes of the ovine FAANG project. The work from Davenport et al. and Massa et al. is setting high standards for analyzing histone modification, transcription factor binding and/or whole genome-wide methylation analyses. The authors demonstrated that the level of activity at the functional genomic elements found correlated with nearby transcriptomic expression. Further exploration of transcription start sites (Salavati et al.) confirmed the spatial association of active genomic elements and initiation of transcription. Furthermore, E and colleagues used whole genome sequencing and Hi-C to provide mechanistic insights as to the biological basis for polled intersex syndrome (PIS) leads to reproductive disorders in goats.

THE USE OF FUNCTIONAL GENOMIC DATA TO PREDICT CAUSAL VARIANTS

An ultimate goal of both basic and applied genomics is to connect genotype to phenotype, and multiple groups reported progress in linking genetic variation with the molecular phenotype of RNA expression, which has seen substantial advancement in Genotype-Tissue Expression (GTEx) studies in humans and model species. By analyzing chicken tissues for which both RNA-seq and genomic DNA sequence were available in two populations, Jehl et al. developed thresholds for variant calling and showed the value of existing RNA-seq datasets for reliable SNP detection in allele-specific expression (ASE) and future GTEx studies. In a second report from this group and again investigating chicken RNA-seq data, Degalez et al. reported on the value of haplotype-aware variant annotation and the interest to consider multi-nucleotide variants in the coding regions. Prowse-Wilkins et al. produced and integrated histone modification and CTCF data across six tissues from lactating dairy cows to identify partitions of the genome predicted to comprise functional regions in these tissues. Importantly, they then showed the level of activity of these functional regions were correlated with nearby gene expression and such regions were enriched for putative causal variants. Interestingly, the level of enrichment improved where regions were correlated with the level of expression and was greatest for QTL for milk production traits. This work provided strong evidence for the core

hypothesis of the FAANG project; that form follows function and cataloging genome functional elements can be used to find important (e.g., predictive) variation likely causing phenotypic differences.

As exemplified by this collection, the efforts produced by groups throughout the world indicate the future of FAANG is very bright. Having said this, the value of the insights provided by the currently more comprehensive efforts in human and biomedical models is clear, and significantly more progress will be needed to fully exploit the public investment in animal agricultural genomics. Especially challenging will be the validation of predicted functional elements and the verification of casual variants associated with complex traits, as each polymorphism may have only a small effect. However, the next major advancements in translation of farm animal genome functional variation into prediction of biological phenotype will come from such precise knowledge of individual genomes.

We close by congratulating each of the contributing authors for their outstanding work, and extend our appreciation to all of the reviewers for their time and effort to improve each submission.

AUTHOR CONTRIBUTIONS

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

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 Chamberlain, Cheng, Giuffra, Kuehn, Tuggle 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 October 2020
doi: 10.3389/fgene.2020.580580

Global Analysis of Transcription Start Sites in the New Ovine Reference Genome (*Oar rambouillet v1.0*)

Mazdak Salavati[1,2†], Alex Caulton[3,4†], Richard Clark[5], Iveta Gazova[1,6], Timothy P. L. Smith[7], Kim C. Worley[8], Noelle E. Cockett[9], Alan L. Archibald[1], Shannon M. Clarke[3], Brenda M. Murdoch[10] and Emily L. Clark[1,2] on behalf of The Ovine FAANG Project Consortium*

[1] The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Edinburgh, United Kingdom, [2] Centre for Tropical Livestock Genetics and Health (CTLGH), Roslin Institute, University of Edinburgh, Midlothian, United Kingdom, [3] AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand, [4] Genetics Otago, Department of Biochemistry, University of Otago, Dunedin, New Zealand, [5] Genetics Core, Edinburgh Clinical Research Facility, The University of Edinburgh, Edinburgh, United Kingdom, [6] MRC Human Genetics Unit, The University of Edinburgh, Edinburgh, United Kingdom, [7] USDA, Agricultural Research Service, U.S. Meat Animal Research Center, Clay Center, NE, United States, [8] Baylor College of Medicine, Houston, TX, United States, [9] Department of Animal, Dairy and Veterinary Sciences, Utah State University, Logan, UT, United States, [10] Department of Animal, Veterinary and Food Sciences, University of Idaho, Moscow, ID, United States

OPEN ACCESS

Edited by:
Huaijun Zhou, University of California, Davis, United States

Reviewed by:
Oleg Gusev, RIKEN, Japan
Zhihua Jiang, Washington State University, United States

***Correspondence:**
Emily L. Clark
emily.clark@roslin.ed.ac.uk

† These authors have contributed equally to this work

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

Received: *06 July 2020*
Accepted: *09 September 2020*
Published: *23 October 2020*

Citation:
Salavati M, Caulton A, Clark R, Gazova I, Smith TPL, Worley KC, Cockett NE, Archibald AL, Clarke SM, Murdoch BM and Clark EL (2020) Global Analysis of Transcription Start Sites in the New Ovine Reference Genome (Oar rambouillet v1.0). Front. Genet. 11:580580. doi: 10.3389/fgene.2020.580580

The overall aim of the Ovine FAANG project is to provide a comprehensive annotation of the new highly contiguous sheep reference genome sequence (*Oar rambouillet v1.0*). Mapping of transcription start sites (TSS) is a key first step in understanding transcript regulation and diversity. Using 56 tissue samples collected from the reference ewe Benz2616, we have performed a global analysis of TSS and TSS-Enhancer clusters using Cap Analysis Gene Expression (CAGE) sequencing. CAGE measures RNA expression by 5′ cap-trapping and has been specifically designed to allow the characterization of TSS within promoters to single-nucleotide resolution. We have adapted an analysis pipeline that uses TagDust2 for clean-up and trimming, Bowtie2 for mapping, CAGEfightR for clustering, and the Integrative Genomics Viewer (IGV) for visualization. Mapping of CAGE tags indicated that the expression levels of CAGE tag clusters varied across tissues. Expression profiles across tissues were validated using corresponding polyA+ mRNA-Seq data from the same samples. After removal of CAGE tags with <10 read counts, 39.3% of TSS overlapped with 5′ ends of 31,113 transcripts that had been previously annotated by NCBI (out of a total of 56,308 from the NCBI annotation). For 25,195 of the transcripts, previously annotated by NCBI, no TSS meeting stringent criteria were identified. A further 14.7% of TSS mapped to within 50 bp of annotated promoter regions. Intersecting these predicted TSS regions with annotated promoter regions (±50 bp) revealed 46% of the predicted TSS were "novel" and previously un-annotated. Using whole-genome bisulfite sequencing data from the same tissues, we were able to determine that a proportion of these "novel" TSS were

hypo-methylated (32.2%) indicating that they are likely to be reproducible rather than "noise". This global analysis of TSS in sheep will significantly enhance the annotation of gene models in the new ovine reference assembly. Our analyses provide one of the highest resolution annotations of transcript regulation and diversity in a livestock species to date.

Keywords: ovine, TSS, CAGE, WGBS, promoter, enhancer, transcriptome, FAANG

INTRODUCTION

The Functional Annotation of Animal Genomes (FAANG) consortium is a concerted international effort to use molecular assays, developed during the Human ENCODE project (Birney et al., 2007), to annotate the majority of functional elements in the genomes of domesticated animals (Andersson et al., 2015; Giuffra and Tuggle, 2019). Toward this aim, the overarching goal of the Ovine FAANG project (Murdoch, 2019) is to provide a comprehensive annotation of the new highly contiguous reference genome for sheep, *Oar rambouillet v1.0*.[1] The Ovine FAANG project is developing a deep and robust dataset of expressed elements and regulatory features in the sheep genome as a resource for the livestock genomics community. Here, we describe a global analysis of transcription start sites (TSS) using Cap Analysis Gene Expression (CAGE) sequencing.

Cap Analysis Gene Expression measures RNA expression by 5′ cap-trapping to identify the 5′ ends of both polyadenylated and non-polyadenylated RNAs including lncRNAs and miRNAs, and has been specifically designed to allow the characterization of TSS within promoters to single-nucleotide resolution (Takahashi et al., 2012). By using 5′-cap capture, we avoid transcripts that have been 5′ degraded. Conventional RNA-Seq and cDNA datasets can be "contaminated" with such degradation products and data from transcripts where first strand cDNA synthesis was incomplete. These "contaminants" can give rise to erroneous transcript/gene models with false 5′ ends. The level of resolution provided by CAGE allows investigation of the regulatory inputs driving transcript expression and construction of transcriptional networks to study, for example, the genetic basis for disease susceptibility (Baillie et al., 2017) or for systematic analysis of transcription start sites through development (Lizio et al., 2017). Using CAGE sequencing technology, the FANTOM5 consortium generated a comprehensive annotation of TSS for the human genome, which included the major primary cell and tissue types (Forrest et al., 2014).

The goal of this study was to annotate TSS and TSS-Enhancer clusters in the ovine genome (*Oar rambouillet v1.0*). Our approach was to perform CAGE analysis on 55 tissues and one type of primary immune cell (alveolar macrophages). Tissues representing all the major organ systems were collected from Benz2616, the Rambouillet ewe used to generate the *Oar rambouillet v1.0* reference assembly. CAGE tags for each tissue sample clustered with a high level of specificity according to their expression profiles as measured by mRNA-Seq. Mapping of CAGE tags indicated that a large proportion of detected TSS did not overlap with the current annotated 5′ end of transcripts. The reproducibility of these "novel" TSS was tested using whole-genome DNA methylation profiles from a subset of the same tissues.

DNA methylation plays a key role in the regulation of gene expression and the maintenance of genome stability (Ibeagha-Awemu and Zhao, 2015), and is the most highly studied epigenetic mark. In mammalian species, DNA methylation occurs primarily at cytosine-phosphate-guanine dinucleotides (CpG) and to a lesser extent at CHH and CHG sites (where C, cytosine; H, adenine, guanine, or thymine; and G, guanine) (An et al., 2018). Generally, DNA methylation in the promoter region of genes represses transcription, inhibiting elongation by transcriptional machinery. Methylation over TSS represses transcription initiation whereas, conversely, methylation within gene bodies stimulates elongation and influences alternative splicing of transcripts (Jones, 2012; Lev Maor et al., 2015; An et al., 2018). Using DNA methylation profiles, we were able to determine the proportion of "novel" TSS in our dataset that were likely true signals of transcription initiation based on a hypomethylated state rather than being an artifact of CAGE sequencing.

We provide the annotation of TSS in the ovine genome as tracks in a genome browser via the Track Hub Registry and visualize these in the R package GViz, ensuring the data are accessible and useable to the livestock genomics community. The global analysis of TSS we present here will significantly enhance the annotation of gene models in the new ovine reference assembly demonstrating the utility of the datasets generated by the Ovine FAANG project and providing a foundation for future work.

MATERIALS AND METHODS

Animals

Tissues were collected from an adult female Rambouillet sheep at the Utah Veterinary Diagnostic Laboratory on April 29, 2016. At the time of sample collection, Benz2616 was approximately 6 years of age and after a thorough veterinary examination confirmed to be healthy. Benz 2616 was donated to the project by the USDA. Sample collection methods were planned and tested over 15 months in 2015−2016, and a description of these is available via the FAANG Data Coordination Centre.[2]

[1] https://www.ncbi.nlm.nih.gov/assembly/GCF_002742125.1/

[2] https://data.faang.org/api/fire_api/samples/USU_SOP_Ovine_Benz2616_Tissue_Collection_20160426.pdf

Sample Collection

Necropsy of Benz2616 was performed by a veterinarian to ensure proper identification of tissues, and a team of scientists on hand provided efficient and rapid transfer of tissue sections to containers which were snap frozen in liquid nitrogen before transfer to −80°C for long-term storage. Alveolar macrophages were collected by bronchoalveolar lavage as described in Cordier et al. (1990). Details of all 100 samples collected from Benz2616 are included in the BioSamples database under submission GSB-7268, group accession number SAMEG329607[3] and associated information is recorded according to FAANG metadata specifications (Harrison et al., 2018). The FAANG assays, as described later, were generated from a subset of tissues for CAGE (56 tissues), polyA+ mRNA-Seq (58 tissues), and whole-genome bisulfite sequencing (WGBS) (8 tissues) (**Figure 1**).

CAGE Library Preparation and Analysis

RNA Isolation for CAGE Library Preparation

Frozen tissues (60–100 mg per sample) were homogenized by grinding with a mortar and pestle on dry ice and RNA was isolated using TRIzol Reagent (Invitrogen) according to the manufacturer's instructions. After RNA isolation, 10 μg of RNA per sample was treated with DNase I (NEB) then column purified using a RNeasy MinElute kit (Qiagen), according to the manufacturer's instructions. Full details of the RNA extraction protocol are available via the FAANG Data Coordination Centre.[4] Each RNA sample was run on an Agilent BioAnalyzer to ensure RNA integrity was sufficiently high ($RIN^e > 6$). Details of RNA purity metrics for each sample are included in **Supplementary Table 1**. RNA samples were then stored at −80°C for downstream analysis.

CAGE Library Preparation and Sequencing

Cap Analysis Gene Expression libraries were prepared for each sample as described in Takahashi et al. (2012) from a starting quantity of 5 μg of DNase treated total RNA. Random primers were used to ensure conversion of all 5′ cap-trapping RNAs according to Takahashi et al. (2012). The full protocol is available via the FAANG Data Coordination Centre.[5] Libraries were prepared in batches of eight and pooled. Sequencing was performed on the Illumina HiSeq 2500 platform by multiplexing eight samples on one lane to generate approximately 20 million 50 bp single-end reads per sample. Eight of the available fifteen 5′ linker barcodes from Takahashi et al. (2012) were used for multiplexing: ACG, GAT, CTT, ATG, GTA, GCC, TAG, and TGG. In total, eight separate library pools were generated and spread across two HiSeq 2500 flow cells. Details of barcodes assigned to each sample and pool IDs are included in **Supplementary Table 1**.

Processing and Mapping of CAGE Libraries

All sequence data were processed using in-house scripting (bash and R) on the University of Edinburgh high-performance computing facility (Edinburgh, 2020). The analysis protocol for CAGE is available via the FAANG Data Coordination Centre[6] and summarized in **Figure 2**. To de-multiplex the data, we used the FastX toolkit version 0.014 (Hannon Lab, 2017) for short read pre-processing. We then used TagDust2 v.2.33 (Lassmann, 2015) to extract mappable reads from the raw data and for read clean-up to remove the *EcoP1* site and barcode, according to the recommendations of the FANTOM5 consortium (e.g., Bertin et al., 2017). This process resulted in cleaned reads approximately 27 nt in length (hereafter referred to as CAGE tags) which were mapped to the Rambouillet Benz2616 genome available from NCBI (*Oar rambouillet v1.0* GCA_002742125.1) using Bowtie2 v.2.3.5.1 in −very-sensitive mode equivalent to options *-D 20 -R 3 -N 0 -L 20 -i S,1,0.50* (Langmead and Salzberg, 2012). Multi-mapped reads were identified using Bowtie2 v.2.3.5.1 in –very-sensitive mode and excluded from the rest of the analysis. The mapped BAM files were then processed for base-pair resolution strand-specific read counts using bedtools v.2.29.0 (Quinlan and Hall, 2010). Metrics for the attrition of raw reads at each stage of the analysis pipeline are included in **Supplementary File 1**, **Section 1.1** For the bedGraph files to be used in the CAGEfightR package, they were converted to bigWig format using UCSCs tool BedGraphToBigWig (Kent et al., 2010).

Normalization and Mapping of CAGE Tags

For normalization and clustering of CAGE tags (as CAGE Tags-Per-Million Mapped: CTPM), we used the software package CAGEfightR v.1.5.1 (Thodberg and Sandelin, 2019). The normalization was performed by dividing CAGE tag counts in each predicted cluster by the total mapped CAGE tags in the sample, multiplied by 1×10^6. To perform these analyses, we created a custom BSgenome object (a container of the genomic sequence) for sheep from *Oar rambouillet v1.0* using the BSgenome Bioconductor package v.1.53.1 (Pages, 2020). Distribution metrics of CAGE tags across the genome were annotated and analyzed using the TxDB transcript ID assignment and Genomic Features package v.1.36.4 (Lawrence et al., 2013). The TxDB object was created using the NCBI gff3 gene annotation file from NCBI *Oar rambouillet v1.0* GCA_002742125.1 (*GCF_002742125.1_Oar_rambouillet_v1.0_genomic.gff release 103*).

Clustering of CAGE Tags

To annotate TSS in the *Oar rambouillet v1.0* genome assembly, we first generated expression read counts for each tag (bp resolution). Tags with <10 read counts were removed first then any tags that were not present in at least 37/56 tissues (i.e., two-thirds of the tissues) were also removed. This conservative representation threshold was introduced to ensure CAGE tags included in downstream analysis were reproducible. In the absence of additional biological replicates, we based this on the assumption that a CAGE tag was

[3]https://www.ebi.ac.uk/biosamples/samples/SAMEG329607

[4]https://data.faang.org/api/fire_api/assays/USDA_SOP_RNA_Extraction_From_Tissue_20180626.pdf

[5]https://data.faang.org/api/fire_api/assays/ROSLIN_SOP_CAGE-library-preparation_20190903.pdf

[6]https://data.faang.org/api/fire_api/analysis/ROSLIN_SOP_CAGE_analysis_pipeline_20191029.pdf

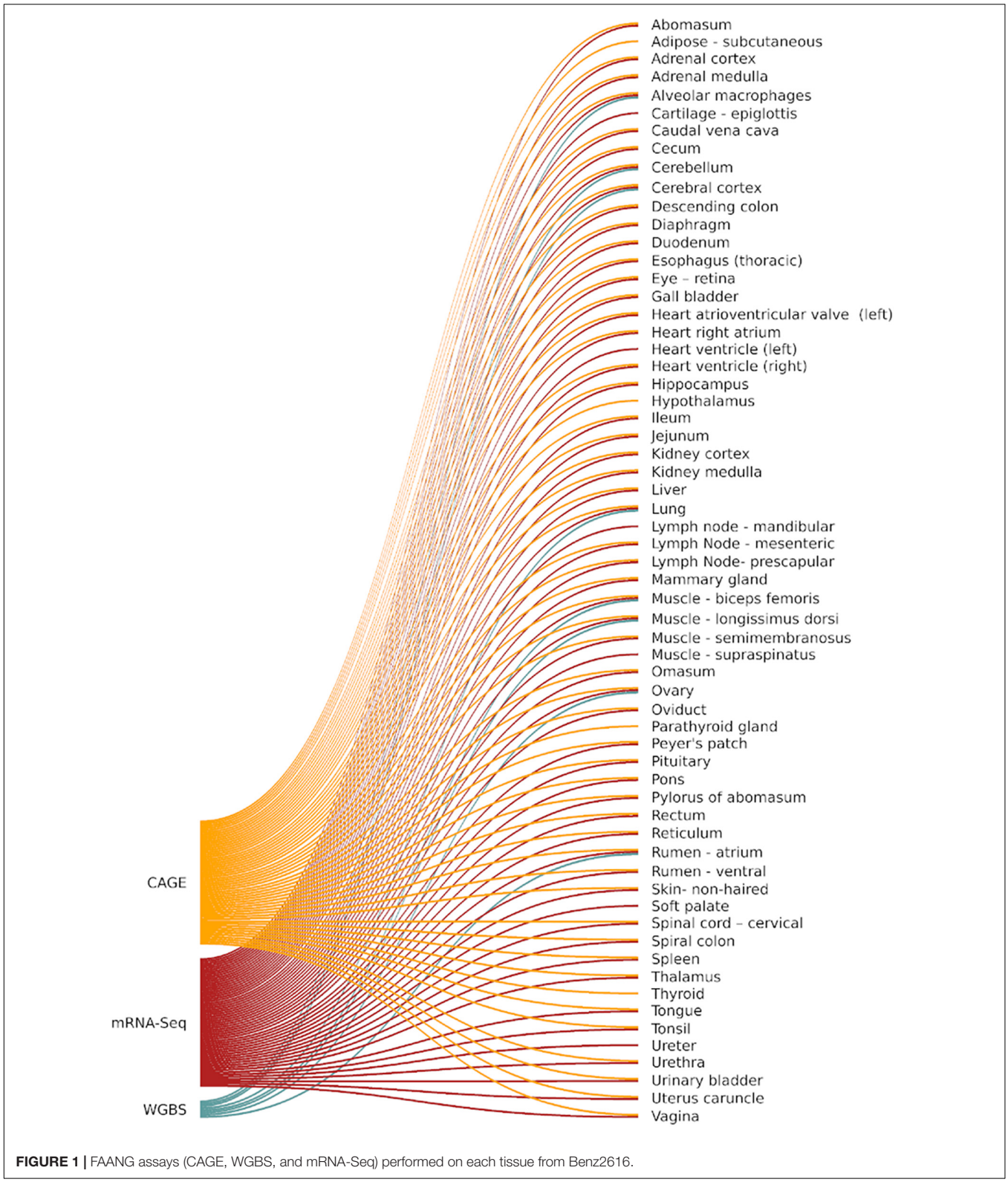


FIGURE 1 | FAANG assays (CAGE, WGBS, and mRNA-Seq) performed on each tissue from Benz2616.

more likely to be reproducible if it was shared across multiple tissues. However, it should be noted that this method would reduce sensitivity to putative highly tissue-specific TSS and this is discussed later. Gene annotation from NCBI's GTF file (*GCF_002742125.1_Oar_rambouillet_v1.0_genomic.gtf release 103*) was used to validate the coordinates of predicted

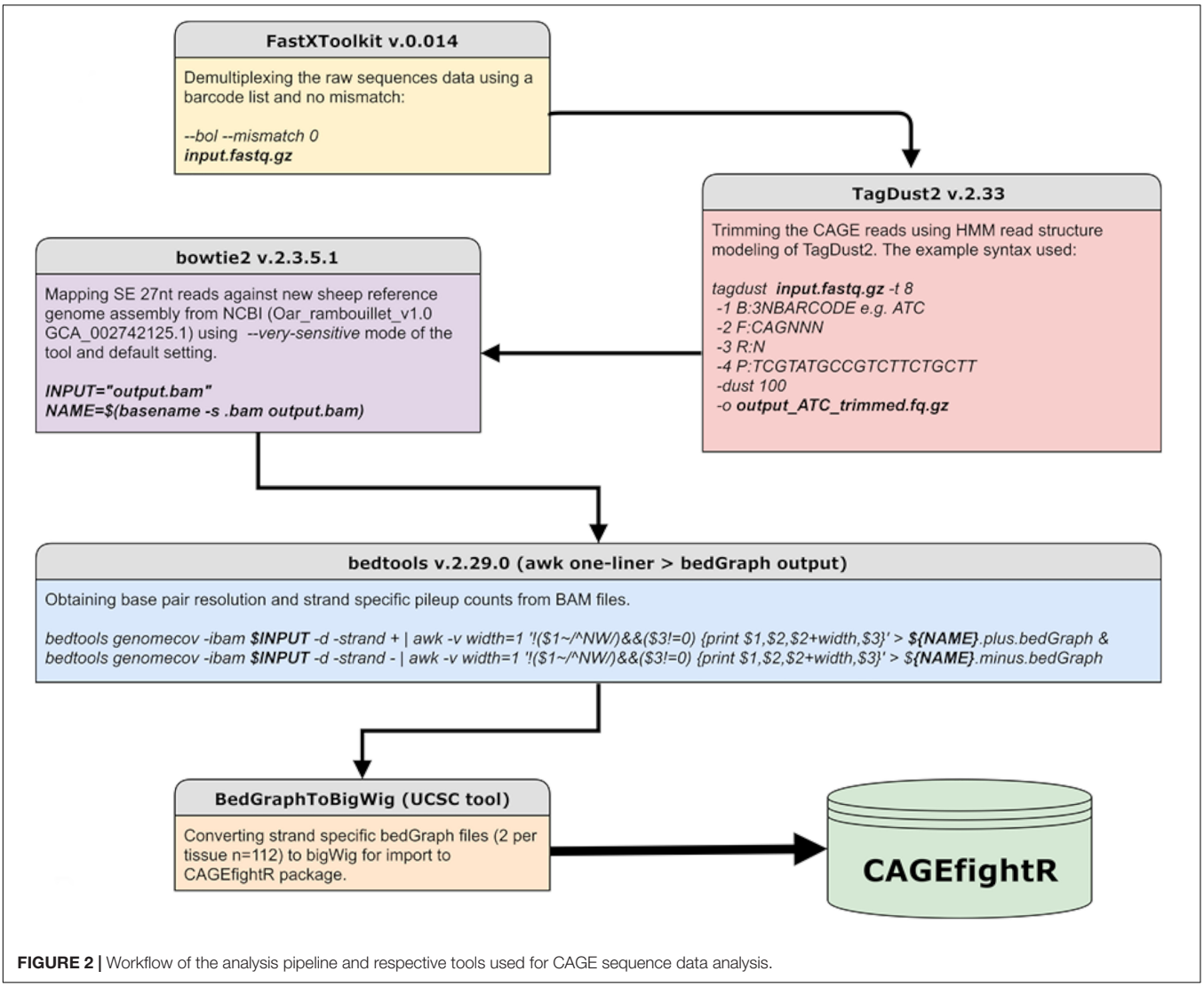


FIGURE 2 | Workflow of the analysis pipeline and respective tools used for CAGE sequence data analysis.

CAGE clusters (i.e., residing within or outside the promoter of annotated genes). Five thresholds for representation, of CAGE tags (excluding intergenic and intronic tags) across tissues, were compared (one tissue, five tissues, one-third of the tissues, half of the tissues, two-thirds of the tissues, and all of the tissues). The proportion of CAGE tag clusters within (tagged by unique gene IDs) or outside the promoter region (untagged) was used to compare each threshold. Highly stringent filtering (56/56 representation) found CAGE tag clusters associated with 2,949 genes (out of 30,862 genes annotated by NCBI) representing putative TSS for genes expressed in all 56 tissues. A reduction of the threshold to two-thirds (37/56 tissues) resulted in 13,912 genes (31,113 transcripts) associated with CAGE tag clusters. Reducing the threshold further to one-third of tissues resulted in a high proportion of CAGE tag clusters that were not associated with genes ("untagged") (41.6%) and 18,005 associated with genes (39,458 transcripts). According to this criterion, we selected the two-thirds threshold. Although highly stringent, this provided only the highest confidence TSS tag clusters, associated with widely expressed genes and widely used promoters, for the analysis of the dataset we present here. Further details of this comparison are included in **Supplementary File 1**, **Section 1.2**.

Transcription start sites expression profiles (as CTPM) were then regenerated for each tissue using the CAGEfightR v.1.5.1 quickTSS, quickEnhancers, and findLinks functions (Thodberg and Sandelin, 2019). The CAGE tags clustered (1) uni-directionally (according to the sense or anti-sense flag of the mapped CAGE tag) into predicted TSS and (2) bi-directionally, using the TSS-Enhancer detection algorithm from CAGEfightR (Thodberg and Sandelin, 2019), into correlated TSS and enhancer (TSS-Enhancer) clusters. Bi-directional (TSS-Enhancer) clusters are defined as clusters of CAGE tags that are located on the opposing strand within 400−1,000 bp proximity of the center of a promoter (Thodberg and Sandelin, 2019). The bi-directional clusters outside of this range were excluded from this analysis according to the previously described method in Thodberg et al. (2019). The concept of uni-directional and bi-directional clustering is illustrated in **Figure 3**.

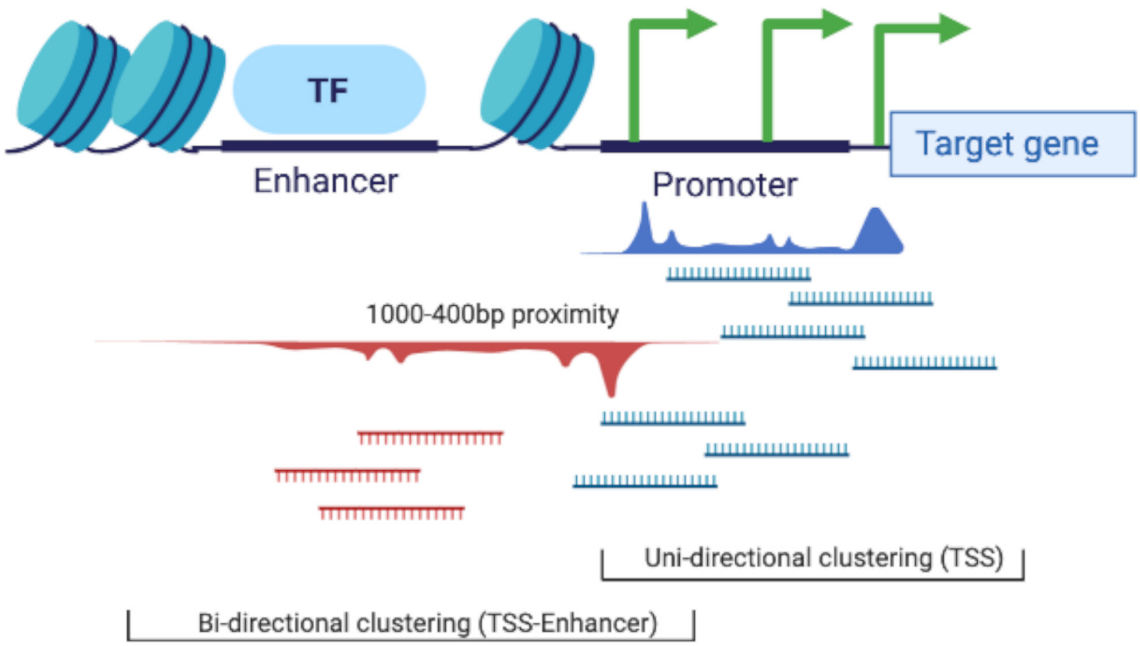


FIGURE 3 | Schematic representation of the two clustering algorithms used in the CAGEfightR package for TSS (uni-directional) and TSS-Enhancer (bi-directional) clustering.

Identification of Shared TSS or TSS-Enhancer Clusters Across Tissues

Transcription start sites or TSS-Enhancer clusters that were shared across tissues were identified by investigating the CTPM expression profile of each of the tissues using correlation-based and mutual information (MI) distance matrices (Priness et al., 2007; Reshef et al., 2018). This method of MI-based clustering tolerates missingness and outlier-induced grouping errors in gene expression profiles (Priness et al., 2007). Using this method, we assumed that the CTPM expression profile, for each cluster, could vary across tissues. However, for a predicted TSS or TSS-Enhancer cluster to be considered high-confidence and associated with widely expressed genes and widely used promoters, it must be present in at least two-thirds of the tissues (37/56) in the dataset.

Identification of Tissue-Specific TSS or TSS-Enhancer Clusters

The two-thirds representation threshold applied previously would remove all tissue-specific CAGE tag clusters. To overcome this, a rerun of the clustering algorithm was performed with the two-thirds representation threshold removed. Tissue-specific uni-directional TSS clusters that were only present in 1/56 tissues were identified by filtering for CAGE tags with >10 expressed counts to create a data frame. The data frame was then filtered tissue by tissue to only retain uni-directional TSS clusters present in each tissue separately. This process was then repeated for the TSS-Enhancer clusters.

Annotation of "Novel" TSS in the Ovine Genome

We expected given the diversity of tissues sampled that we would detect a significant number of "novel", previously unannotated TSS. The CAGE tag uni-directional clusters (TSS) were annotated using the mergeByOverlay function of the GenomicFeatures package in R and the custom TxDB object as follows:

mergeByOverlaps(subject = TSS, query = promoters(txdb, upstream = 25, downstream = 25, use.names = T,c("tx_name," "GENEID")), maxgap = 25, type = "any"). The TxDB object calculates the range of the promoter based on the 5′UTR and first CDS codon coordinates. In each tissue, any putative TSS region within 50 bp range of the promoter coordinate of a gene model was considered "annotated". In addition, we expanded this range to 400 bp to determine whether this would identify significantly more unannotated TSS further from the promoter. A reverse sub setting of the 50 bp window region was performed as follows: *subsetByOverlaps(x = TSS, ranges = annotated, invert = TRUE).* These regions were considered "novel" TSS previously unannotated in the assembly. This process was repeated for every tissue separately ($n = 56$).

Comparative Analysis of WGBS and CAGE Data

Preparation of Genomic DNA From Tissue

Extraction of DNA for bisulfite sequencing was performed using a phenol:chloroform:isoamyl alcohol method. Briefly, approximately 1 g frozen tissue was pulverized and resuspended in 2.26 ml of digestion buffer (10 mM Tris−HCl, 400 mM

NaCl, 2 mM EDTA, pH 8.0) with 200 μl of SDS 10% and 60 μl RnaseA (10 mg/ml) (Sigma-Aldrich, St. Louis, MO, United States). RNA degradation proceeded for 1 h at 37°C with gentle shaking. Next, 25 μl of proteinase K (20 mg/ml) (Sigma-Aldrich) was added to the suspension and incubated overnight (approximately 16 h) at 37°C with gentle shaking. The viscous lysate was transferred to a 2 ml Phase Lock tube (VWR, Radnor, PA, United States) and extracted twice with Tris–HCl-saturated phenol:chloroform:isoamyl alcohol (25:24:1) pH 8.0, followed by extraction with 2.5 ml chloroform. The DNA was precipitated by addition of 5.5 ml of 100% ethanol and 250 μl of 3 M sodium acetate to the aqueous phase in a 15 ml conical tube, mixed by gentle inversion until the DNA became visible. The DNA was removed with a bent Pasteur pipette hook, washed in 5 ml 70% cold ethanol, air dried then resuspended in 250 μl−1 ml of 1 × TE, and stored at −20°C until use. DNA concentration was quantified fluorometrically on the Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, United States) using the Qubit dsDNA HS Assay Kit. The purity of the extractions was determined via 260/280 and 260/230 ratios measured on the NanoDrop 8000 (Thermo Fisher Scientific) and DNA integrity was assessed by 1% agarose gel electrophoresis. The protocol is available via the FAANG Data Coordination Centre.[7]

Whole-Genome Bisulfite Conversion and Sequencing

Library preparation and sequencing of seven tissues and one cell type (**Figure 1**), selected to include a representative from all major organ systems, were performed by The Garvan Institute of Medical Research, Darlinghurst, Sydney, NSW, Australia. Un-methylated lambda DNA was added at 0.5% of the total sample DNA concentration before bisulfite conversion as a conversion efficiency control. DNA conversion was carried out using the EZ DNA Methylation-Gold Kit (Zymo Research, CA, United States) following the manufacturer's instructions. The Accel-NGS Methyl-seq DNA kit (Swift Biosciences, MI, United States) for single indexing was used to prepare the libraries, following the manufacturer's instructions. Libraries were pooled together and sequenced across six lanes of a flow cell on an Illumina HiSeq X platform using paired-end chemistry for 150 bp reads (min 10× coverage). The protocol is available via the FAANG Data Coordination Centre.[8]

WGBS Data Processing

Paired-end Illumina WGBS data were processed and analyzed using in-house scripting (bash and R) and a range of purpose-built bioinformatics tools on the AgResearch and University of Edinburgh high-performance computing facilities. The analysis protocol for WGBS is available via the FAANG Data Coordination Centre[9] and summarized in the next section.

Briefly, FASTQ files for each sample, run across multiple lanes, were merged together. TrimGalore v.0.5.0[10] was used to trim raw reads to remove adapter oligos, poor-quality bases (phred score less than 20), and the low-complexity sequence tag introduced during Accel-NGS Methyl-seq DNA kit library preparation as follows: *trim_galore -q 20 –fastqc –paired –clip_R2 18 –three_prime_clip_R1 18 –retain_unpaired –o Trim_out INPUT_R1.fq.gz INPUT_R2.fq.gz*.

A bisulfite-sequencing amenable reference genome was built using the *Oar rambouillet v1.0*, GenBank accession number: GCA_002742125.1 genome with the BSSeeker2 script *bs_seeker2-build.py* using bowtie v2.3.4.3 (Langmead and Salzberg, 2012) and default parameters. The Enterobacteria phage lambda genome available from NCBI (accession number NC_001416) was added to the *Oar rambouillet v1.0* genome as an extra chromosome to enable alignment of the unmethylated lambda DNA conversion control reads. Paired-end, trimmed reads were aligned to the reference genome using the BSSeeker2 script *bs_seeker2-align.py* and bowtie v2.3.4.3 (Langmead and Salzberg, 2012) allowing four mismatches (-m 4). Aligned bam files were sorted with samtools v1.6 (Li et al., 2009) and duplicate reads were removed with picard tools v2.17.11[11] MarkDuplicates function.

Deduplicated bam files were used to call DNA methylation levels using the "bam2cgmap" function within CGmaptools (Guo et al., 2018) with default options to generate ATCGmap and CGmap files for each sample. The ATCGmap file format summarizes mapping information for all covered nucleotides on both strands, and is specifically designed for BS-seq data; while the CGmap format is a more condensed summary providing sequence context and estimated methylation levels at any covered cytosine in the reference genome.

Hypermethylated and hypomethylated regions were determined for each sample using methpipe v3.4.3 (Song et al., 2013). Specifically, CGmap files for each sample were reformatted for the methpipe v3.4.3 workflow using custom awk scripts. The methpipe symmetric-cpgs program was used to merge individual methylation levels at symmetric CpG pairs. Hypomethylated and hypermethylated regions were determined using the hmr program within methpipe, which uses a hidden Markov model using a Beta-Binomial distribution to describe methylation levels at individual CpG sites, accounting for the read coverage at each site.

Visualization of the individual CpG site methylation levels with a minimum read depth cut-off of 10x coverage was done using Gviz package v.1.28.3 (Hahne and Ivanek, 2016).

Comparative Analysis of Annotated and "Novel" TSS with WGBS Methylation Information

We expected that reproducible TSS, either annotated or novel, would overlap with hypomethylated regions of the genome (Yamashita et al., 2005; Yagi et al., 2008). To

[7]https://data.faang.org/api/fire_api/assays/USDA_SOP_DNA_Extraction_From_WholeBloodandLiver_20200611.pdf

[8]https://data.faang.org/api/fire_api/assays/AGR_SOP_WGBS_AgR_Libary_prep_20200610.pdf

[9]https://data.faang.org/api/fire_api/analysis/AGR_SOP_WGBS_AgR_data_analysis_20200610.pdf

[10]https://github.com/FelixKrueger/TrimGalore

[11]https://broadinstitute.github.io/picard/

test whether this was true for those identified in our analysis, both annotated and novel TSS from the CAGE BED tracks were intersected with WGBS hypomethylation profiles using bedtools v.2.29.2 (Quinlan and Hall, 2010) and the following script: *bedtools intersect -b WGBS_HypoCpG.bed -a Novel_or_ Annotated.bed > Novel_or_annotated_HypoCpG.bed.* Any annotated and novel TSS (within a ± 50 bp window of the promoter) that intersected hypomethylated regions of DNA in each tissue, were verified as reproducible TSS and the remainder as "noise". The overlay of these regions was visualized as a genomic track using the Gviz package v.1.28.3 (Hahne and Ivanek, 2016).

Visualization of the Annotated TSS, mRNA-Seq, and WGBS Tracks in the Ovine Genome

To confirm the simultaneous expression of mRNA, CAGE tags corresponding to an active TSS and a hypomethylated region of DNA, a genomic track on which all three datasets could be visualized, were generated. This visualization consists of the following tracks: (1) uni-directional CAGE tag clusters (TSS), (2) bi-directional CAGE tag clusters (TSS-Enhancers), (3) WGBS hypomethylation score (bp resolution), (4) transcript level expression (mRNA-Seq [TPM]), (5) the transcript models, and (6) the gene model. Areas of the genome where TSS or TSS-Enhancer regions overlapped regions with a high hypomethylation score, within the 5′ end of an actively expressing transcript (TPM score), were considered reproducible TSS for that tissue. This process was performed using eight tissues with matching mRNA-Seq, CAGE, and WGBS data. The Gviz package v.1.28.3 was used to visualize these tracks (Hahne and Ivanek, 2016).

Validation of Tissue-Specific Expression Profiles

mRNA Sequencing

Total RNA for mRNA-Seq from 32 tissues (**Figure 1**) was prepared, as previously for the CAGE samples, by USMARC, and for 26 tissues by Baylor College of Medicine (BCM) using the MagMAX mirVana total RNA isolation kit (Thermo Fisher Scientific, Waltham, MA, United States) according to the manufacturer's instructions. Paired-end polyA selected mRNA-Seq libraries were prepared and sequenced on an Illumina NextSeq500 at USMARC or the Illumina HiSeq2000 at BCM using the Illumina Tru-Seq Stranded mRNA Library Preparation Kit. For each tissue, a set of expression estimates, as transcripts per million (TPM), were obtained using the transcript quantification tool Kallisto v0.43.0 (Bray et al., 2016). The mRNA-Seq analysis pipeline is accessible via the FAANG Data Coordination Centre.[12] A pairwise distance matrix (multiple correlation coefficient based) was produced using MI values for all tissues and a dendrogram of tissues was created to visualize grouping patterns of tissues with similar mRNA expression profiles, and for comparison with the CAGE dataset.

[12]https://data.faang.org/api/fire_api/analysis/ROSLIN_SOP_RNA-Seq_analysis_pipeline_20200610.pdf

Comparative Analysis of Tissue-Specific Expression Profiles Using Information From CAGE and mRNA-Seq

We assessed whether TSS expression profiles from the CAGE dataset were biologically meaningful using the mutual information (MI) sharing algorithm (Joe, 1989). Tissues with the same function and physiology should have similar TSS expression profiles. The CTPM expression level was binned (n = 10) using the bioDist package v.1.56.0 (Ding et al., 2012) and mutual information (MI) for each pair of tissue samples was calculated as in Joe (1989).

$$\delta = \left(1 - \exp\left(-2 \times \delta\right)\right)^{0.5}$$

$$MI\ distance = 1 - \delta$$

A pairwise distance matrix (multiple correlation coefficient based) was produced using MI values for all tissues and a dendrogram of tissues created to visualize grouping patterns of tissues with similar TSS expression profiles. If the expression profiles were meaningful, then tissues with similar function and physiology would group together in clades within the dendrogram. These tissue-specific groupings were then further validated by comparison with mRNA-Seq data for the same samples, using the MI sharing algorithm and dendrogram approach.

RESULTS

Library Size and Annotation Metrics

The mean CAGE library depth based on uniquely mapped CAGE reads was 4,862,957 reads. A detailed explanation of the attrition of reads at each stage of the analysis pipeline is included in **Supplementary File 1**, **Section 1.1**. Library depth varied across tissues. Tissues with low depth were not related to any specific barcodes and were evenly spread over the two sequencing runs (**Supplementary Figure 1** and **Supplementaty Table 1**), suggesting random variation rather than systematic differences due to specific barcodes or sequencing runs. The RIN^e values were also consistently >7 for all tissues with low counts, indicating RNA integrity was also unlikely to be affecting library depth. Differences in tag numbers are therefore more likely to relate to variation in efficiency between individual libraries or tissue-specific differences related to the physiology of the tissue.

CAGE Tag Clustering and Annotation by Genomic Regions

We used a newly developed software package to annotate TSS in the Rambouillet Benz2616 genome (Thodberg and Sandelin, 2019; Thodberg et al., 2019) which clustered the CAGE tags as (1) uni-directionally into predicted TSS or (2) bi-directionally into correlated TSS and enhancer (TSS-Enhancer) clusters (**Figure 3**). The clustered CAGE tags were filtered to

remove any clusters with a minimum expression level of <10 tag counts. The mean (±SD) and median number of tissues per cluster was 3.68 ± 4.78 and 2, respectively. Application of the two-thirds representation criteria (i.e., a minimum of 37/56 tissues had to express the tag cluster) and filtering out of tag clusters with <10 TPM resulted in an average of 8,219 uni-directional TSS clusters, from a total of 5,450,864 (pre-filtering), for downstream analysis. A detailed description of the cluster metrics at each stage of filtering is included in **Supplementary File 1**, **Sections 1.1** and **1.3**. Although direct comparisons are difficult due to differences in methodology and the relative "completeness" of the reference annotation used, the level of retained CAGE sequencing datasets (0.5% retained clusters with two-thirds tissue representation) is somewhat lower than reported for other mammalian promoter-level expression atlas projects. In the FANTOM5 project, for example, approximately 5% of clusters were retained (Forrest et al., 2014). To further validate the two-thirds tissue representation criteria we chose, we also investigated the number of transcripts captured in the poly-A enriched mRNA-Seq dataset. Poly-A enriched mRNA-Seq data were available for 52 matching tissues and captured a smaller number of transcripts (n = 32,852) with TPM > 10 in comparison with CAGE CTPM > 10 (n = 53,507). Direct comparison of expression for CAGE tags (27 nt) and paired-end RNA-Seq (75 nt) reads could result in technology-dependent bias. Taking this into consideration, the CAGE dataset with the two-thirds representation criteria applied provided annotation for 31,113 transcripts with minimum CPTM > 10. When the same criteria were applied to the mRNA-Seq dataset, only 3,908 transcripts with TPM > 10 were annotated. The expression (as TPM) of transcripts for each tissue and the TPM threshold metrics are included in **Supplementary File 1** and **Supplementary Table 4**.

Bi-directional TSS-enhancer clusters were far fewer in number, although retention was higher with over 23% meeting the same two-thirds representation criteria 741 from a total of 3,131. Though fewer in number, these bi-directional (including TSS-enhancer) clusters are functionally important in the regulation of expression of their target genes (Andersson et al., 2014; Thodberg and Sandelin, 2019), consistent with finding them in over two-thirds of tissues. The co-expression of leading enhancer RNA (eRNA) which is captured by CAGE sequencing can provide a map to enhancer families in the genome and the genes under their regulation (Andersson et al., 2014).

The locations of both uni-directional TSS and bi-directional TSS-Enhancer clusters were identified in *Oar rambouillet v1.0* and the proportion of TSS clusters located within or near annotated gene features was estimated (**Figure 4**). The custom BSgenome and TxDB objects created from the GFF3 file format provide detailed calculated coordinates for the following sections: intergenic (>1,000 bp before 5′UTR or after the end of 3′UTR), proximal (1,000 bp upstream of the 5′UTR), promoter (± 100 bp from 5′UTR), and the standard gene model (5′UTR, exon, intron, and 3′UTR). The genomic region class with the highest number of uni-directional clusters (39.25%) was the promoter regions (± 100 bp from 5′UTR) (**Figure 4A**), with a relatively even distribution within the other regions of the genome, including 6% mapping proximally to the 5′UTR. The majority of bi-directional TSS-Enhancer clusters were also located in promoter regions (70.1%) with a smaller proportion (25.6%) located in proximal regions (**Figure 4B**). The lack of bi-directional TSS-Enhancer clusters in other regions is a consequence of the operation of the CAGEfightR algorithm, which only considers bi-directional clusters within a 400–1,000 bp window of a TSS CAGE tag cluster (Thodberg and Sandelin, 2019; Thodberg et al., 2019). This approach also reduced the total count compared with unidirectional clusters (28,148 uni-directional clusters relative to 741 bi-directional TSS-Enhancer clusters across tissues) (Thodberg et al., 2019).

Capturing Metrics of CAGE Tag Clusters per Gene

During the clustering process, we also determined the proportion of annotated genes and transcripts in the *Oar rambouillet v1.0* NCBI annotation that we did not capture using our dataset. When the two-thirds representation filtering criteria were applied, 44.7% of transcripts (25,195) and 54.6% of genes (16,950) were not captured by our CAGE TSS clusters. When the two-thirds representation filtering criteria were removed, and presence of the CAGE tag in only one tissue out of 56 considered sufficient, the proportion that we did not capture was reduced to 7% of genes and 5% of transcripts. To investigate whether some genes posessed multiple putative TSS, we also estimated the number of CAGE TSS clusters per gene. The median and also the highest frequency of TSS cluster per gene was 1 (mean 1.8) (13,912 genes/31,3113 transcripts annotated using our dataset), indicating that the vast majority of genes annotated in the *Oar rambouillet v1.0* reference genome have only one TSS, and genes with more than five TSS were rare (**Supplementary Figure 5** and **Supplementary Table 5**).

Distribution of CAGE Tag Clusters in *Oar rambouillet v1.0* Relative to *Oar_v3.1*

The new reference sheep genome assembly (*Oar rambouillet v1.0*) is more contiguous than the earlier draft genome sequence *Oar_v3.1* (Jiang et al., 2014) with contig N50 values of 2,572,683 and 40,376 bp, respectively, and would be expected to provide a better template for annotation of gene models and other genomic features. As a proxy for testing this assumption, we investigated how mapped CAGE tag clusters were distributed across the two genome assemblies (**Supplementary Figure 2**). The percentage of uni-directional CAGE tag clusters mapping to intergenic regions, which usually occurs due to missing gene model information, was greater for *Oar_v3.1* (33.9%) relative to *Oar rambouillet v1.0* (8%). The percentage of uni-directional CAGE tag clusters mapping to annotated promoter regions was greater for *Oar rambouillet v1.0* (39.25%) compared with *Oar_v3.1* (14.94%), indicating the proportion of accurate gene models in *Oar rambouillet v1.0* was greater. Of the 28,148 unidirectional TSS clusters mapped to *Oar rambouillet v1.0*, 87.74% mapped to 13,868 unique genes (31,729 transcripts). In comparison, of the 23,829 unidirectional TSS clusters mapped to *Oar_v3.1*, 49.1% mapped to 6,549 genes (9,914 transcripts). A larger number of TSS-Enhancer CAGE clusters were detected in Oar_v3.1 (1121)

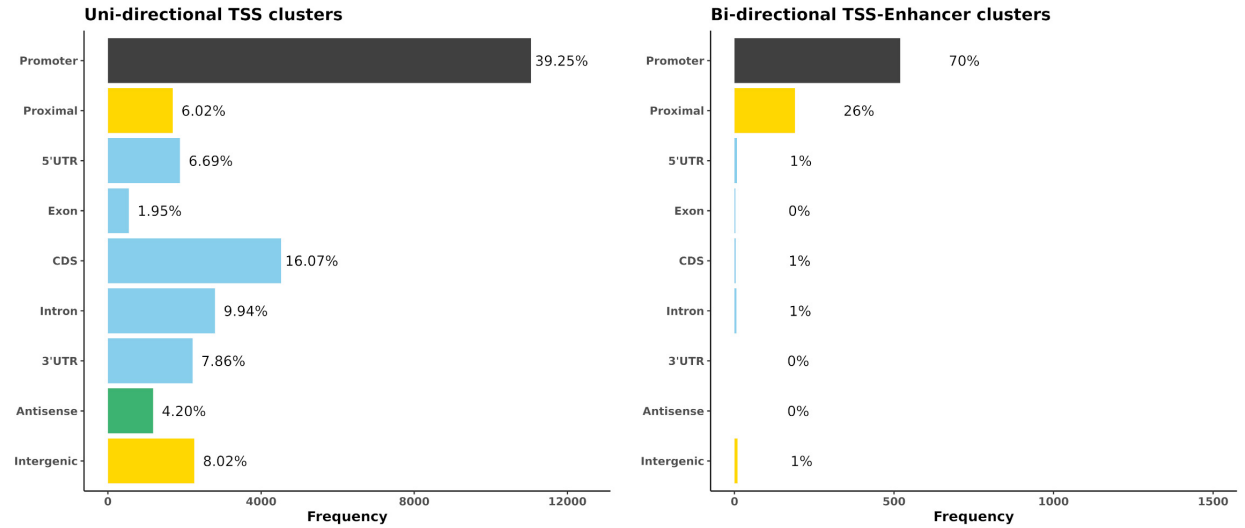


FIGURE 4 | The genomic region distribution of CAGE tag clusters mapped against *Oar rambouillet v1.0* assembly and gene annotation. The counts were averaged across tissues. **(A)** Uni-directional TSS clusters with the highest proportion in promoter region (± 100 bp of the 5 UTR beginning at the [TSS]). **(B)** Bi-directional TSS-Enhancer clusters with the highest proportion in the proximal region (1,000 bp upstream of the 5′UTR beginning at the [TSS]).

in comparison with *Oar rambouillet v1.0* (741) mapping to 1371 and 2598 unique genes, respectively. A detailed comparison of mapping of the CAGE tags with the two reference assemblies is included in **Supplementary File 1**, **Sections 2** and **3**.

Mapping of CAGE Tags Shared Across All Tissue Samples

Correlation-based and mutual information (MI) distance matrices were used to evaluate the occurrence of TSS and enhancer TSS across tissues. The mean ± SD number of tissues in which each cluster passed the two-thirds criteria (expressed in 37/56 tissues) was 47.73 ± 6.03. Uni-directional TSS clusters (n = 28,148 TSS regions) that were shared across tissues and detected in at least 37/56 tissues are visualized in **Figure 5**. Each chord in **Figure 5** represents the presence of an expressed uni-directional TSS cluster shared across tissues. The majority of the uni-directional TSS that were shared across tissues mapped to promoters (39.25%) and were shared evenly across the tissues sampled (**Figure 5**). Some tissues, e.g., mammary gland, pituitary gland, and urinary bladder, had more uni-directional TSS mapping to intergenic regions, which might indicate evidence of alternative splicing or differential TSS usage across tissues (**Figure 5**). Alternative splicing events and differential TSS usage, captured by CAGE, are often not included in reference gene prediction models (Berger et al., 2019).

Bi-directional TSS-Enhancer CAGE clusters were far fewer in number but were shared in a similar pattern across tissues as the uni-directional TSS clusters (**Figure 6**). The majority (70.1%) of the TSS-Enhancer clusters mapped to promoters (n = 520) while 25.6% mapped to "proximal" regions as expected according to the 400−1,000 bp detection window for TSS-Enhancer clusters from the center of the promoter (**Figure 6**). For some tissues including abomasum, spleen, and heart right atrium, the proportion of bi-directional TSS-Enhancer clusters mapping to proximal regions was greater indicating more enhancer families could be present within these tissues (**Figure 6**).

Mapping of Tissue-Specific CAGE Tags

The application of the two-thirds criteria provided a high level of confidence in assigning TSS and TSS-Enhancer elements, but eliminated the ability to observe potential tissue-specific CAGE tags or TSS clusters. Tissue-specific tags, i.e., those observed in only one of the 56 tissues, were examined to evaluate the ability to distinguish tissue-specific clusters from the background. A total of 3,228,425 tags were observed in only one tissue, and a much higher proportion (80.0%) of these tags mapped to intergenic and intronic regions compared with tags found across tissues, suggesting they do not represent true TSS (**Supplementary Table 2**). Only 0.8% of the tissue-specific CAGE tag clusters mapped to promoter or proximal regions (**Supplementary Table 2**). The cecum (n = 1554), cerebellum (n = 601), and longisimus dorsi muscle (n = 477) had the highest number of tissue-specific predicted unidirectional TSS. The greatest number of expressed TSS (>1 CTPM) was detected in ceberellum (84/601) as shown in **Supplementary Figure 3A**. However, the expression level of tissue-specific CAGE tag clusters was very low (<2 CTPM), which combined with the small sample size (n = 1) for each tissue meant that analysis of tissue-specific TSS was not particularly meaningful using this dataset. The analysis was repeated for tissue-specific TSS-Enhancer clusters which is detailed in **Supplementary Figure 3B**.

Proportion of "Novel" TSS Within the CAGE Dataset for Each Tissue

Cap Analysis Gene Expression tag clusters were annotated initially using the *Oar rambouillet v1.0* gene models from NCBI.

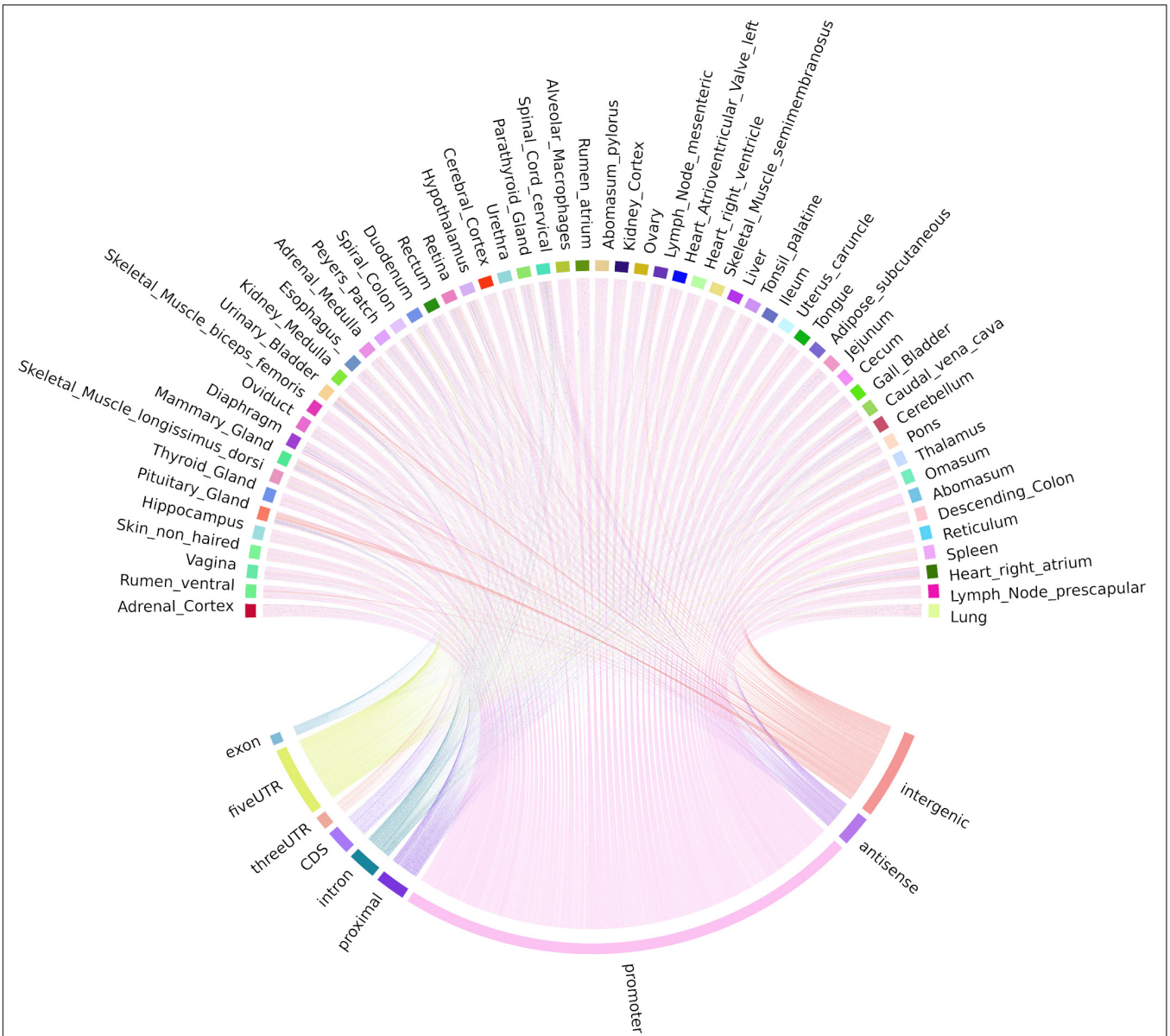


FIGURE 5 | Chord diagram of expression level (TPM) of CAGE tag clusters (uni-directional TSS) across all the tissues collected from Benz2616. Shared CAGE tag clusters are common to at least two-thirds of the tissues (37/56).

A tissue-by-tissue annotation was performed using the same gene models to identify any CAGE tag clusters within a 50 bp window of the promoter boundaries of every gene. From a total of 23,994 ± 518 TSS (the average number of TSS per tissue ± SE), we found 11,349 ± 170 (49.8% ± 0.01) were located within 50 bp of the promoter. The CAGE tag clusters were annotated using the NCBI *Oar rambouillet v1.0* GFF3 gene track file (version 103) and a TxDB object created in the GenomicFeatures package (version 1.36.4) in R. CAGE tag clusters within 50 bp (short range) or 400 bp (long range) of the promoter were defined as annotated. **Supplementary File 2** includes BED files for these CAGE tag clusters. The percentage of "novel" previously un-annotated, but likely to be reproducible, CAGE tag clusters for each tissue within 50 bp (short range) and 400 bp (long range) from the promoter are detailed in **Table 1**.

Comparative Analysis of CAGE and WGBS to Validate "Novel" TSS

True TSS and TSS-Enhancer elements are very likely to be associated with areas of hypomethylation (Yamashita et al., 2005; Yagi et al., 2008). The assessment of hypomethylation of regions where "novel" TSS were identified thus provides a means to support or question their designation as true TSS. The methylation status of putative TSS regions for eight of the tissues used for CAGE analysis was examined at single nucleotide

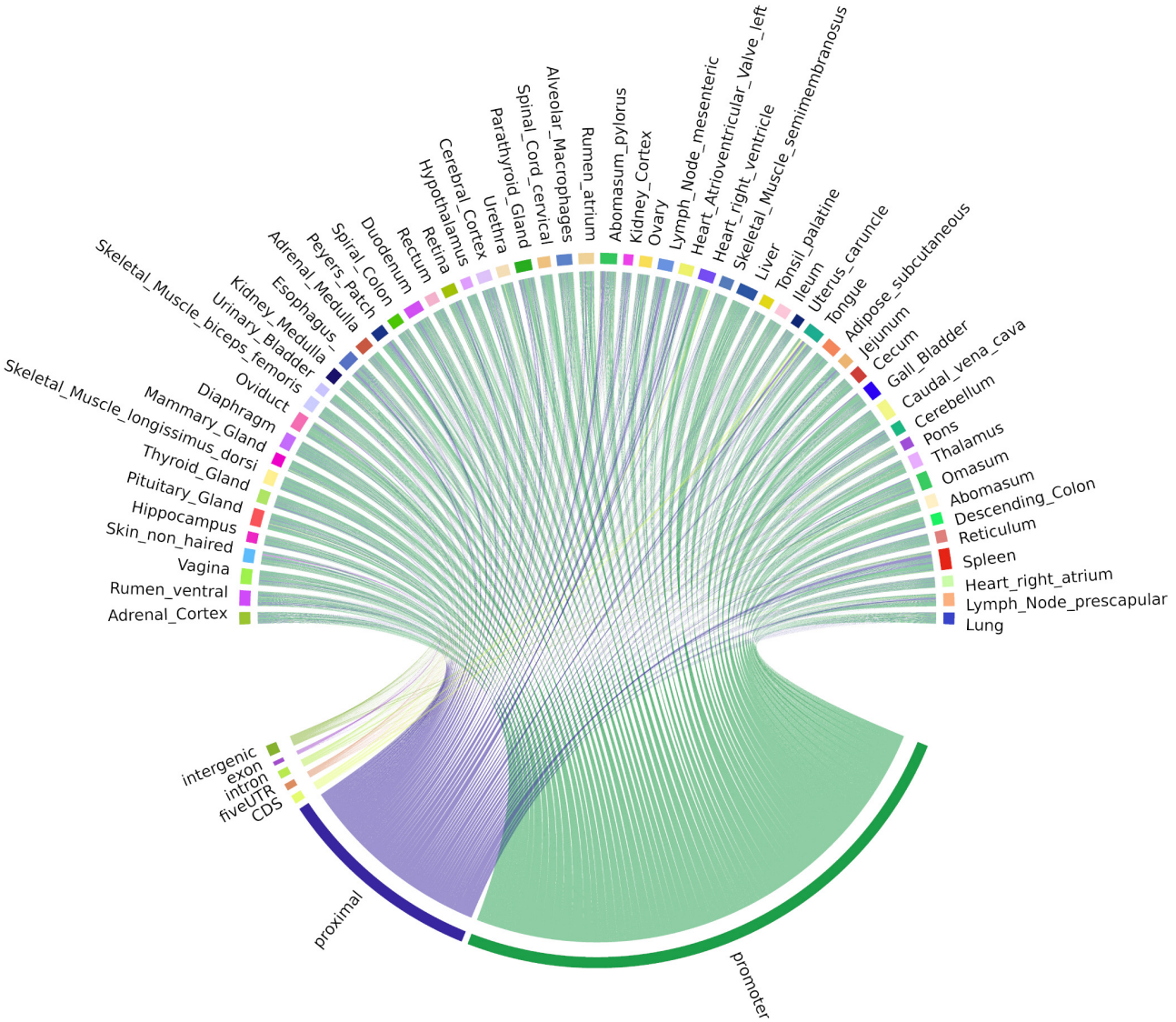


FIGURE 6 | Chord diagram of expression level (TPM) of CAGE tag clusters (bi-directional TSS-Enhancer) across all the tissues collected from Benz2616. CAGE tag clusters expressed (>10 CTPM) by at least two-thirds of the tissues (37/56).

resolution using WGBS. Each WGBS library was pooled before sequencing and multiplexed across eight lanes of the HiSeq X platform. After trimming of the raw reads, the sequenced libraries produced an average of 103 Gbp of clean data. The average mapping rate of the reads was 78.8%. A small proportion (8.5%) of mapped reads were identified as PCR or optical duplicates and were removed before downstream analysis. The average read depth of the filtered libraries was 20× coverage (**Supplementary Table 3**). Only cytosines with a minimum of 10 reads were retained for the subsequent comparative analysis with CAGE data to ensure a high level of confidence in the methylation level estimates, as per published recommendations (Doherty and Couldrey, 2014; Ziller et al., 2015). We would expect that reproducible TSS, either annotated or novel, would overlap with hypomethylated regions of the genome (Yamashita et al., 2005; Yagi et al., 2008). Comparative analysis of the CAGE data with the WGBS methylation levels from eight tissues from Benz2616 was used to investigate methylation levels at the TSS in comparison with gene body and UTR regions. For the majority of genes, the methylation level was much lower around the transcriptionally active TSS or regulatory enhancer candidate regions compared with the gene body (e.g., for gene *IRF2BP2*, **Figure 7**). We overlaid the WGBS hypomethylated regions and the CAGE uni-directional TSS clusters (annotated and "novel") within 50 bp of the promoter. For the eight matching tissues, 88.7% of the annotated TSS clusters and 32.2% of the "novel" TSS were hypomethylated (**Figure 8**). The combined evidence of the hypomethylation and TSS support the conclusion that 32.2% are in fact novel TSS clusters, whereas 67.8% of the novel TSS clusters lack this confirmation.

TABLE 1 | The total number and percentage of "novel" CAGE tag clusters for each tissue within 50 bp (short range) and 400 bp (long range) from the promoter.

Tissue	% Novel	Clusters within 50 bp	Clusters within 400 bp	Total
Abomasum	49.38	8,161	8,688	16,584
Abomasum pylorus	49.89	12,339	13,074	25,963
Adipose subcutaneous	51.74	12,336	13,074	26,970
Adrenal cortex	51.79	12,285	13,019	26,859
Adrenal medulla	52.86	11,520	12,210	25,604
Alveolar macrophages	51.19	12,008	12,731	25,845
Caudal vena cava	37.75	10,937	11,578	18,179
Cecum	51.68	12,070	12,801	26,261
Cerebellum	48.59	8,393	8,936	16,796
Cerebral cortex	51.19	12,199	12,917	26,327
Descending colon	51.80	11,830	12,539	25,810
Diaphragm	52.53	10,367	11,016	22,733
Duodenum	52.34	11,243	11,932	24,620
Esophagus	49.90	10,016	10,625	20,741
Gall bladder	47.78	11,870	12,578	23,852
Heart atrioventricular valve left	50.90	12,268	13,000	26,330
Heart right atrium	52.96	10,996	11,666	24,444
Heart right ventricle	50.47	12,260	12,987	26,082
Hippocampus	53.40	12,142	12,878	27,451
Hypothalamus	52.7	11,105	11,786	24,527
Ileum	52.45	12,352	13,094	27,411
Jejunum	31.67	10,810	11,418	16,361
Kidney cortex	52.04	12,317	13,057	27,076
Kidney medulla	51.07	10,946	11,618	23,365
Liver	49.35	12,255	12,981	25,459
Lung	52.91	11,644	12,339	25,995
Lymph node mesenteric	46.34	12,132	12,838	23,742
Lymph node prescapular	53.56	11,533	12,228	26,096
Mammary gland	49.75	10,048	10,688	20,774
Omasum	39.89	9,167	9,708	15,692
Ovary	50.79	12,334	13,073	26,434
Oviduct	53.29	11,563	12,260	25,957
Parathyroid gland	53.5	11,577	12,272	26,231
Peyer's patch	52.41	11,881	12,578	26,240
Pituitary gland	47.25	6,918	7,362	13,400
Pons	40.69	11,622	12,296	20,506
Rectum	53.55	12,002	12,723	27,192
Reticulum	53.39	12,185	12,911	27,589
Retina	53.54	11,805	12,537	26,691
Rumen atrium	50.69	12,335	13,077	26,363
Rumen ventral	40.20	7,109	7,567	12,165
Skeletal muscle biceps femoris	50.23	12,151	12,872	25,715
Skeletal muscle longissimus dorsi	53.67	11,356	12,060	25,748
Skeletal muscle semimembranosus	51.15	12,262	12,993	26,471
Skin non-haired	52.4	11,907	12,629	26,337
Spinal cord cervical	51.47	11,376	12,050	24,508
Spiral colon	53.25	11,937	12,662	26,813
Spleen	53.46	12,161	12,892	27,568
Thalamus	41.61	11,426	12,079	20,404
Thyroid gland	53.6	11,894	12,615	27,012
Tongue	39.57	9,639	10,244	16,512
Tonsil palatine	46.57	12,178	12,875	23,978
Urethra	52.76	11,387	12,087	25,292
Urinary bladder	51.68	11,163	11,840	24,174
Uterus caruncle	48.33	12,199	12,917	24,857
Vagina	52.30	11,600	12,300	25,543
Average	49.80	11,349	12,032	23,994

Total number represents the number of CAGE tag clusters (from the total of 28,148) that are present in each of the 56 tissues. The clusters for each tissue were then annotated by proximity to promoter regions either 50 or 400 bp (the latter includes the count of the former). The % Novel represents the count of the clusters falling outside of 400 bp vicinity of any current promoter region.

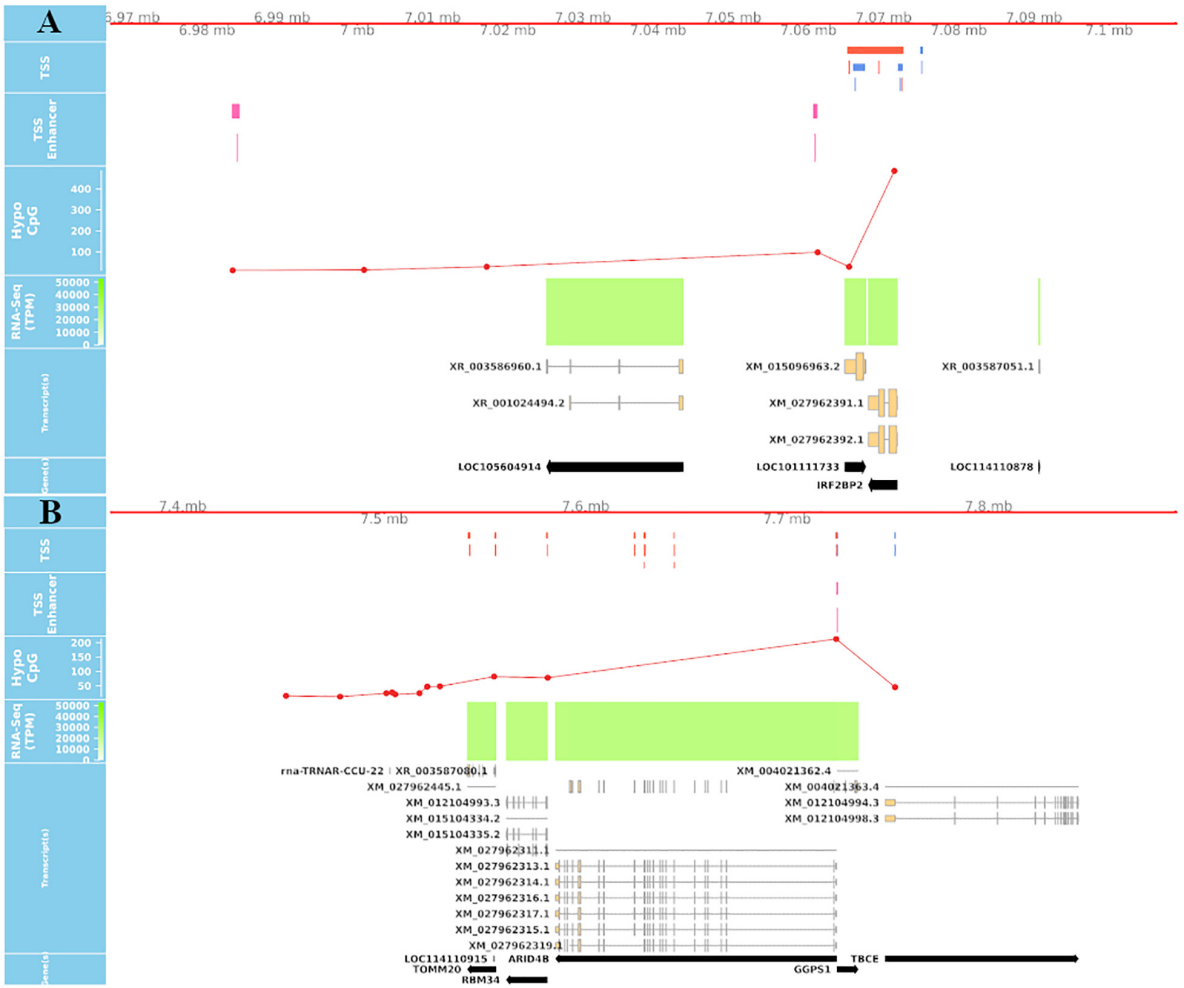


FIGURE 7 | Overlay of CAGE, mRNA-Seq, and WGBS data tracks centered using the genomic coordinates of genes *IRF2BP2* and *ARID4B*. **(A)** Shows a hypomethylated area overlapping multiple uni- and bi-directional CAGE tag clusters at 5′UTR of IRF2BP2. **(B)** Predicted CAGE tag clusters with no verifying hypomethylation island within the middle of ARID4B gene, which are likely to be "noise".

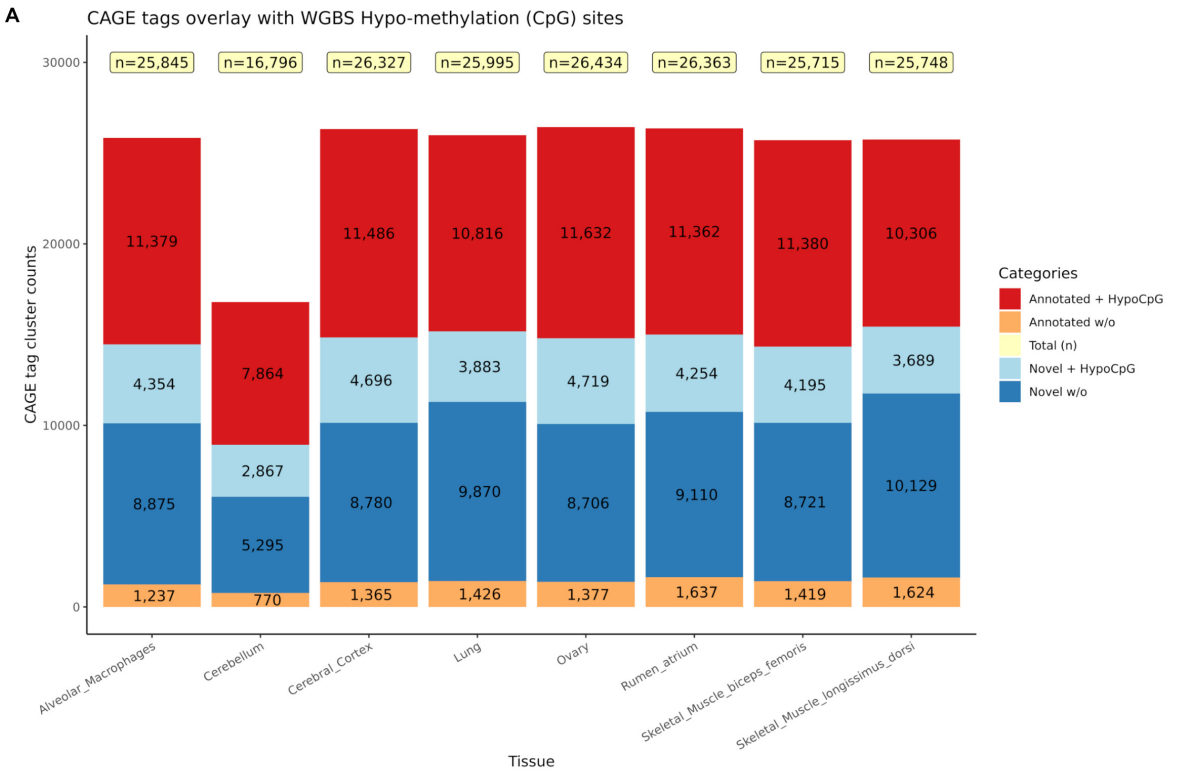
Validation of Tissue Expression Profiles Using mRNA-Seq

The tissue samples from Benz2616 were collected for the purpose of annotating her genome and as such $N = 1$ in all cases. As an alternative strategy to having multiple biological replicates, we validated the expression profiles for each tissue by comparing the CAGE data (CTPM) and mRNA-Seq (TPM) in 52 matching tissues. The transcript expression TPM was significantly correlated with the CAGE tag cluster CTPM values (correlation coefficient 0.19, Pearson p-value $< 1 \times 10^{-8}$) and visualized as a heatmap (**Supplementary Figure 4**).

The similarity of tissue expression profiles for the uni-directional TSS clusters was estimated to determine if tissues with similar physiology and function formed distinct groups as expected. Similarity (distance) analysis showed a partial grouping based on tissue type and organ system as shown in **Figure 9A**. Physiologically similar tissues including nervous system and muscle tissues grouped closely together. This grouping was also present in the mRNA-Seq data from tissue-matched samples (**Figure 9B**), indicating good correlation between the two datasets. Similar groupings based on organ system and tissue type were observed for multiple tissues and cell types generated for the sheep gene expression atlas using mRNA-Seq (Clark et al., 2017).

Comparative Visualization of the Datasets

An interactive visualization interface was developed to make these datasets accessible and usable for the livestock genomics community. The genomic browser incorporates the bp resolution hypomethylation data, the CTPM expression of TSS and TSS-Enhancer regions, and the mRNA-Seq TPM expression at transcript level. These tracks are also overlaid using the coordinates provided by the TxDB objects for transcripts and gene models as shown in **Figure 10**. This form of overlaid view allows for confirmation of transcript expression and the exact coordinate of the corresponding TSS in each tissue. For



B

Tissue	Annotated + HypoCpG	Annotated w/o	Novel + HypoCpG	Novel w/o
Alveolar_Macrophages	90%	10%	33%	67%
Cerebellum	91%	9%	35%	65%
Cerebral_Cortex	89%	11%	35%	65%
Lung	88%	12%	28%	72%
Ovary	89%	11%	35%	65%
Rumen_atrium	87%	13%	32%	68%
Skeletal_Muscle_biceps_femoris	89%	11%	32%	68%
Skeletal_Muscle_longissimus_dorsi	86%	14%	27%	73%

FIGURE 8 | Numbers of CAGE TSS that were hypomethylated according to the WGBS data to distinguish between "novel" reproducible (+HypoCpG) TSS and "noise" (w/o). **(A)** Shows the distribution of CAGE clusters as novel and annotated with or without HypoCpG. **(B)** Percentage of CAGE clusters in each category for each of the eight tissues.

validation purposes, the promoter region should be under a hypomethylated CpG island on the DNA track for a proportion of actively transcribed gene in each tissue. The detailed bigBED format tracks for all the tissues are available online.[13,14]

These visualization tools were used to identify any co-expressed enhancers within the proximity of a TSS. We were able to identify 741 TSS-Enhancer clusters across the 56 tissues. An example of these bi-directional clusters is shown in **Figure 10** as a pink box. The pairwise CTPM levels of co-expression of the bi-directional clusters and those of the uni-directional TSS clusters were compared using the Kendal correlation function in CAGEfightR (Thodberg and Sandelin, 2019). There were 5,383 significant co-expression pairs between uni-directional clusters (28,148) and bi-directional clusters (741). An example of a co-expressed TSS-Enhancer is shown in **Figure 10** as a black line connecting the significant start positions of the co-expression pairs.

The co-expression range of bi-directional clusters, in some cases, can span beyond the 10-Kbp distance, as shown in the *IK* gene example (**Figure 10**). The expression of enhancer RNA (eRNA) with the promoter expression level of their target genes has been reported before (Tippens et al., 2018). This layer of annotation provides a foundation for enhancer target mapping in the sheep genome. The detailed list and annotated target

[13]https://trackhubregistry.org/search/view_trackhub/TW3SmXMBjGhbrAAjJGTU

[14]https://data.faang.org/api/fire_api/trackhubregistry/hub.txt

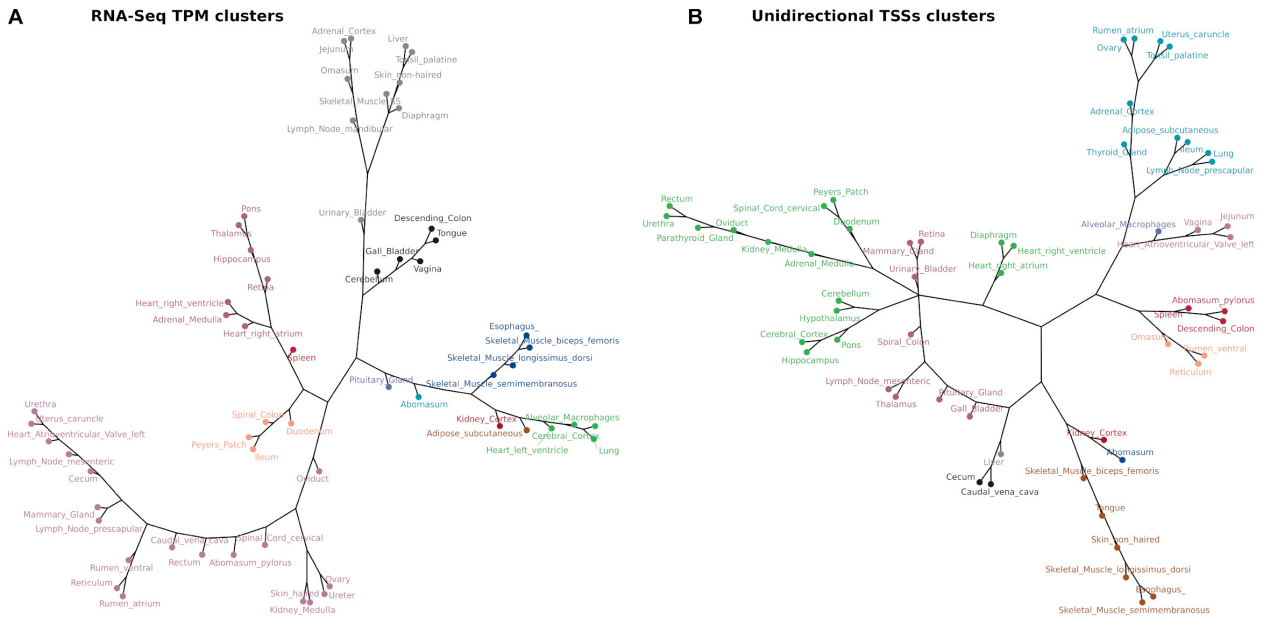


FIGURE 9 | Network analysis of tissue TSS and gene expression profiles in 52 matched samples from Benz2616. The clustering algorithm was based on MI distance of each tissue given the expressed **(A)** mRNA-Seq transcript level TPM and **(B)** CAGE tag clusters (TSSs).

transcripts of these co-expression clusters can be found in **Supplementary File 2**.

FANTOM5 Mammalian and Avian CAGE TSS

The FANTOM5 project also used CAGE to annotate TSSs in mammalian and avian genomes (Andersson et al., 2014; Forrest et al., 2014; Imada et al., 2020). The FANTOM5 data release contained putative TSSs for human, mouse, chicken, rhesus monkey, and dog.[15] We performed a comparative analysis of the number of TSSs captured by these datasets with the CAGE dataset we generated for sheep (**Table 2** and **Supplementary Table 6**). The number of genes annotated by CAGE uni-directional TSS clusters in this study was greater than chicken, rhesus monkey, and dog produced as part of the FANTOM5 project; however, TSS annotation for sheep was still consistently less robust than for murine and human genomes.

DISCUSSION

High-quality reference genomes are now available for many farmed animal species including domestic sheep (*Ovis aries*). The earlier draft genome sequence (Jiang et al., 2014) has been superseded by a more contiguous genome assembly (*Oar rambouillet v1.0*[16]). Annotation of this genome sequence, however, is currently limited to gene and transcript models. There is a lack of information on regulatory sequences and the complexity of the transcriptome is underestimated. For example, promoters and TSS are not well annotated and alternative promoters and transcripts are poorly characterized. The overall aim of the Ovine FAANG project was to provide a comprehensive annotation of *Oar rambouillet v1.0*. To contribute to this aim, we generated a high-resolution global annotation of transcription start sites (TSS) for sheep. After removal of CAGE tags with <10 read counts, 39.3% of TSS overlapped with 5′ ends of transcripts, as annotated previously by NCBI. A further 14.7% mapped to within 50 bp of annotated promoter regions. Intersecting these predicted TSS regions with annotated promoter regions (±50 bp) revealed 46% of the predicted TSS were "novel" and previously un-annotated. Using WGBS from the same tissues, we were able to determine that a proportion of these "novel" TSS were hypomethylated (32.2%), indicating that they are likely to be reproducible rather than "noise". The number of NCBI transcript/gene models for which there was no associated CAGE tag cluster was relatively small (7%) when we removed the strict filtering criteria, indicating the usefulness of CAGE data for genome annotation. However, the "noisy" nature of CAGE data, proportion of multi-mappers and duplicated reads, resulted in a considerable attrition of raw reads. We also chose to use strict filtration criteria, requiring the CAGE tags to be present in two-thirds of tissues. This resulted in a relatively modest number of high confidence CAGE clusters. This strict filtering could be relaxed for future analysis of the data. The global annotation of TSS in sheep we present will significantly enhance the annotation of gene models in the new ovine reference assembly (*Oar rambouillet v1.0*).

The quality of the annotation of reference genomes for livestock species is improving rapidly with reductions in the cost of sequencing and generation of new datasets from multiple different functional assays (Giuffra and Tuggle, 2019).

[15]https://fantom.gsc.riken.jp/5/data/

[16]https://www.ncbi.nlm.nih.gov/assembly/GCF_002742125.1/

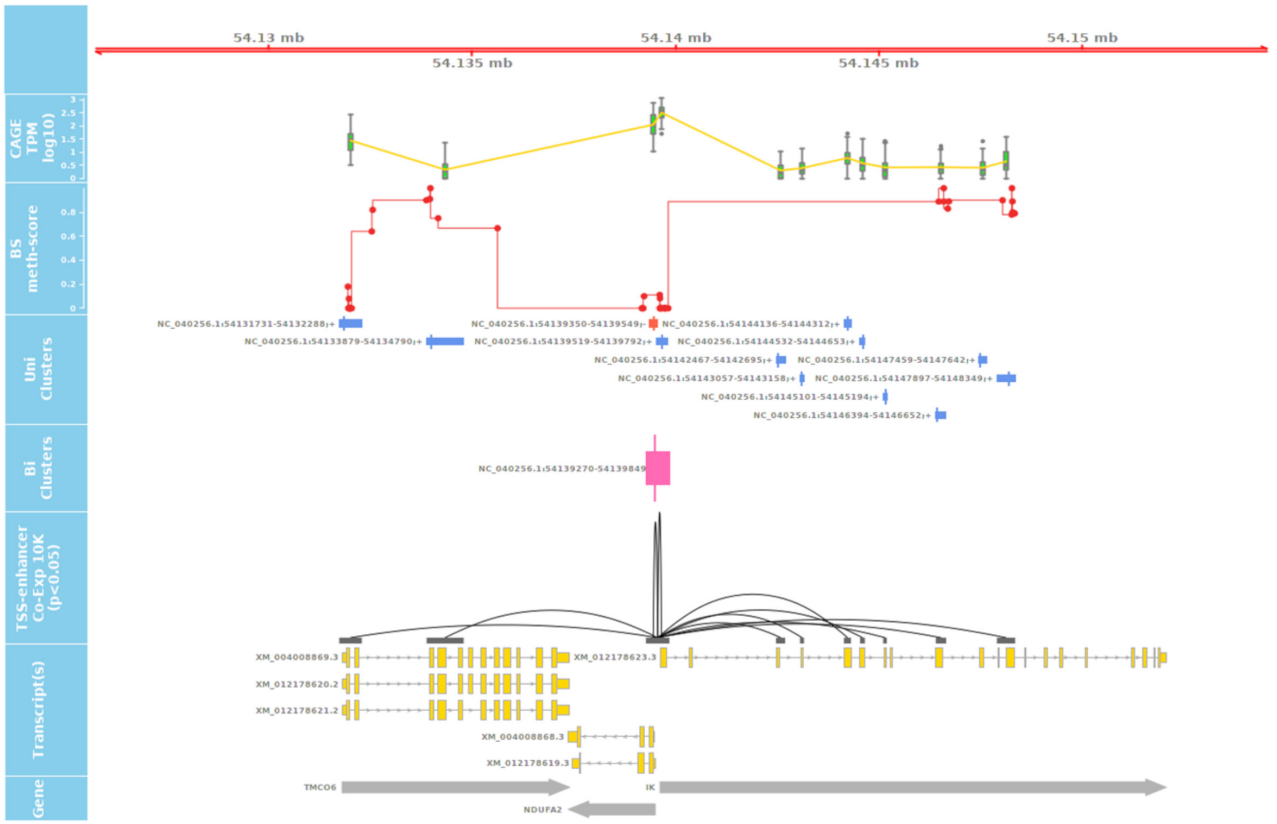


FIGURE 10 | Long-range correlation of single enhancer site with multiple promotors of several genes. The track shows the significant correlation of a leading/primary enhancer site highly co-expressed with several TSS sites of different genes in a relatively long coding frame (± 10,000 Kb). The 3rd track from the top also shows the level of methylation at CpG sites at DNA level of Benz2616 overlaying the same coordinates of the IK gene and ± 10 Kbp.

Oar rambouillet v1.0 superseded the Texel reference assembly (*Oar_v3.1*) Jiang et al. (2014). *Oar_v3.1* is still widely utilized by the sheep genomics community and the Ensembl annotation[17] also includes sequence variation information. We compared how mapped CAGE tag clusters were distributed across genomic features in *Oar rambouillet v1.0* and *Oar_v3.1* (Jiang et al., 2014) and found that the proportion of CAGE tag clusters mapping to promoter regions was greater for *Oar rambouillet v1.0* (39%) than *Oar_v3.1* (15%). This may be because *Oar_v3.1* was built using short-read technology (Jiang et al., 2014), which had a significant bias to GC-rich regions, and therefore did not robustly capture the 5′ ends of many genes (Chen et al., 2013). In comparison, the *Oar rambouillet v1.0* assembly was generated using long-read technology, which dramatically improves the ease of assembly resulting in increased contiguity (Contig N50: *Oar_v3.1* 0.07 Mb and *Oar rambouillet v1.0* 2.57 Mb). Other recent high-quality reference genome assemblies for livestock, e.g., goat (Bickhart et al., 2017; Worley, 2017) and water buffalo (Low et al., 2019), have been built using long-read sequencing technology in combination with optical mapping for scaffolding.

Highly annotated genomes are powerful tools that can help us to understand the mechanisms underlying complex traits in livestock (Georges et al., 2018; Giuffra and Tuggle, 2019) and mitigate future challenges to food production (Rexroad et al., 2019). GWAS results, for example, can be integrated with functional annotation information to identify causal variants enriched in trait-linked tissues or cell types (reviewed in Cano-Gamez and Trynka, 2020). Using enrichment analysis (Finucane et al., 2018) showed that heritable disease associated variants from GWAS were enriched in enhancer regions in relevant tissues and cell types in humans. The TSS and TSS-Enhancer clusters identified in this study could be utilized in a similar way for SNP

TABLE 2 | Metrics comparison of CAGE atlases from 7 species.

Species	Genome	TSS	Genes
Human	hg38	209,911	31,184
Mouse	mm10	164,672	30,501
Chicken	galGal5	32,015	7,759
Sheep	Oar rambouillet v1.0	28,148	13,912
Rhesus monkey	rheMac8	25,869	8,047
Dog	canFam3	23,147	5,288

The total number of TSSs identified using CAGE methodology and the number of corresponding genes. Data for all the species other than sheep were accessed via the FANTOM5 data portal.

[17]https://www.ensembl.org/Ovis_aries/Info/Index

enrichment analysis of GWAS variants in sheep. Using ChIP-Seq data, Naval-Sanchez et al. (2018) found that selective sweeps were significantly enriched for proximal regulatory elements to protein coding genes and genome features associated with active transcription. A high-quality set of variants for sheep, generated using whole-genome sequencing information for hundreds of animals across multiple breeds, is available through the Sheep Genomes Database (2020). This dataset could be used to identify functional SNPs enriched in the TSS and TSS-Enhancer clusters for multiple tissues and cell types that we have annotated in the *Oar rambouillet v1.0* assembly. High-throughput functional screens using gene editing technologies are now possible to validate these functional variants (reviewed in Tait-Burkard et al., 2018). New iPSC lines for livestock species also now offer the potential to do this in relevant cell types (Ogorevc et al., 2016).

Our high-resolution atlas of TSS complements other available large-scale RNA-Seq datasets for sheep (e.g., Clark et al., 2017). The analysis we present here includes tissues representing all major organ systems. However, we were unable to generate CAGE libraries for a small number of difficult to collect or problematic tissues, and as such may have missed transcripts specific to these tissues. We were also only able to generate CAGE libraries from one isolated cell type, alveolar macrophages. As demonstrated by the FANTOM5 (Forrest et al., 2014), ENCODE (Birney et al., 2007) and FragENCODE (Foissac et al., 2019) projects, including a diversity of immune cell types, in both activated and inactivated states, in future work would capture additional transcriptional diversity. New technologies, such as single cell sequencing, will allow annotation of cell-specific expressed and regulatory regions of the genome at unprecedented resolution (Papatheodorou et al., 2019). C1 CAGE now offers the opportunity to detect TSS and enhancer activity at single-cell resolution (Kouno et al., 2019).

We have also generated full-length transcript information using the Iso-Seq method, for a small subset of tissues from Benz2616. Integrating mRNA-Seq and Iso-Seq datasets has been used successfully to improve the annotation of the pig genome (Beiki et al., 2019). By merging the Iso-Seq data with the CAGE and mRNA-Seq datasets, we will be able to measure differential transcript usage across tissues and improve the resolution of the *Oar rambouillet v1.0* transcriptome further. Our analysis indicated that although the vast majority of transcripts had one TSS, some genes had multiple putative TSS which could be validated with the additional resolution provided by the Iso-Seq data. As such, the study we present here represents just the first step in demonstrating the power and utility of the different datasets generated for the Ovine FAANG project, which will provide one of the highest resolution annotations of transcript regulation and diversity in a livestock species to date.

MEMBERS OF THE OVINE FAANG PROJECT CONSORTIUM (LISTED BY INSTITUTION)

Brenda Murdoch, University of Idaho, Moscow, ID, United States; Kimberly M. Davenport, University of Idaho, Moscow, ID, United States; Stephen White, USDA, ARS, Washington State University, Pullman, WA, United States; Michelle Mousel, USDA, ARS ADRC, Clay Center, NE, United States; Alisha Massa, Washington State University, Pullman, WA, United States; Kim Worley, Baylor College of Medicine, Houston, TX, United States; Alan Archibald, The Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom; Emily Clark, The Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom; Brian Dalrymple, University of Western Australia, Perth, WA, Australia; James Kijas, CSIRO, Canberra, ACT, Australia; Shannon Clarke, AgResearch, Mosgiel, New Zealand; Rudiger Brauning, AgResearch, Mosgiel, New Zealand; Timothy Smith, USDA, ARS MARC, Clay Center, NE, United States; Tracey Hadfield, Utah State University, Logan, UT, United States; Noelle Cockett, Utah State University, Logan, UT, United States.

CODE AVAILABILITY

All the code base for the analytical pipeline in this study are available at https://msalavat@bitbucket.org/msalavat/rnaseqwrap_public.git for RNA-Seq analysis; https://msalavat@bitbucket.org/msalavat/cagewrap_public.git for the CAGE mapping, annotation, and metrics pipeline; and https://msalavat@bitbucket.org/caultona/wgbswrap_public.git for WGBS pipeline.

DATA AVAILABILITY STATEMENT

All the raw sequence data and analysis BAM files for this study are publicly available via the OAR_USU_Benz2616 NCBI BioProject: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA414087 and via the European Nucleotide Archive (ENA): https://www.ebi.ac.uk/ena/browser/view/PRJEB34864 (CAGE), https://www.ebi.ac.uk/ena/data/view/PRJEB35292 (mRNA-Seq) and http://www.ebi.ac.uk/ena/data/view/PRJEB39178 (WGBS). Details of all 100 samples collected from Benz 2616 are included in the BioSamples database under submission GSB-7268, group accession number SAMEG329607 (https://www.ebi.ac.uk/biosamples/samples/SAMEG329607). The datasets are accessible via the FAANG data portal and were submitted according to FAANG sample and experimental metadata requirements (Harrison et al., 2018). Oar rambouillet v1.0 is now available on the Ensembl Genome Browser http://www.ensembl.org/Ovis_aries_rambouillet/Info/Index.

ETHICS STATEMENT

The animal study was reviewed and approved by the Utah State University. IACUC approval: #2826, expiration date 21st of February 2021.

AUTHOR CONTRIBUTIONS

RC and IG performed CAGE library, optimization, preparation, and sequencing. MS performed all bioinformatic and data

analyses, with the exception of the WGBS data, which was analyzed by AC. MS and AC generated the GViz tracks. TS coordinated generation of the mRNA-Seq data at US-MARC. KW coordinated generation of the *Oar rambouillet v1.0* reference assembly and mRNA-Seq data at BM. SC coordinated the generation and analysis of the WGBS with AC. EC and AA coordinated the CAGE components of the study. NC and KW planned and coordinated the sample collection at USU. BM was coordinator of the Ovine FAANG project with AA, SW, BD, JK, RB, NC, SC, KW, EC, and TS who designed the overall project and acquired the funding to support the work. EC wrote the article with MS and AC. All authors contributed to editing and approved the final version of the article.

FUNDING

This work was supported by the National Institute of Food and Agriculture, United States Department of Agriculture awards USDA-NIFA-2017-67016-26301 and USDA-NIFA-2013-67015-21228. Sample collection was funded by the NRSP-8 Sheep Genome Coordinator Funding, Project UTA-1172. EC, AA, and MS were partially supported by the Institute Strategic Program grants awarded to the Roslin Institute "Farm Animal Genomics" (BBS/E/D/2021550) and "Prediction of genes and regulatory elements in farm animal genomes" (BBS/E/D/10002070). EC was supported by a University of Edinburgh Chancellors Fellowship. The Edinburgh Clinical Research Facility was funded by the Wellcome Trust. AC was supported by the University of Otago Ph.D. scholarship and granted the Marjorie McCallum travel award to visit the Roslin Institute and her Ph.D. research was funded by the NZ MBIE C10X1906. TS was supported by the USDA-ARS Project No. 3040-31000-100-00D. MS and EC were partially supported by funding from the Bill & Melinda Gates Foundation and with United Kingdom aid from the UK Foreign, Commonwealth and Development Office (Grant Agreement OPP1127286) under the auspices of the Centre for Tropical Livestock Genetics and Health (CTLGH), established jointly by the University of Edinburgh, SRUC (Scotland's Rural College), and the International Livestock Research Institute. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation nor the UK Government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

ACKNOWLEDGMENTS

We would like to thank the Human Genome Sequencing Center staff for the *Oar rambouillet v1.0* reference genome assembly and for long-read and short-read mRNA and miRNA sequencing. HGSC contributors include the genome assembly and analysis team of Y. Liu, R.A. Harris, and X. Qin, led by K. Worley and production team members including M.-C. Gingras and L. Perez for RNA and DNA preparation, V. Vee, Y. Han. V. Korchina, S. Dugan-Perez for sequencing, Q. Meng, H. Doddapaneni, M. Wang for library production, the system support and LIMs teams, D.M. Muzny the HGSC Director of Operations, R.A. Gibbs the HGSC Director. We thank S. Sullivan and I. Liachko of Phase Genomics for PGA genome scaffolding. We would also like to thank William Thompson for isolation of RNA and Jacky Carnahan for isolation of DNA at USMARC. We are grateful to Lucas Lefevre, Rachel Young, and Heather Finlayson for performing the initial optimization to establish the CAGE protocol at the Roslin Institute and Sara Clohisey and Lee Murphy for advice on data analysis and assistance in establishing the protocol at the Edinburgh Clinical Research Facility. CAGE libraries were sequenced at the Centre for Genomic Research, University of Liverpool. We are also grateful for the support of the FAANG Data Coordination Centre (http://data.faang.org) in the upload and archiving of the sample data and metadata, and hosting of the genome tracks. We would also like to thank the tissue collection design team, Noelle Cockett and Kim Worley, who coordinated the team, James Kijas, Brian Dalrymple, Tracy Hadfield, Kara Thornton, Tom Baldwin, Shuna Jones, Bob Lee, Sue Hauver, Christy Kelley, Jessica Eisenhauer, Mike Heaton, William Thompson, Timothy Smith, Stephen White, Michelle Mousel, Alisha Massa; Brian Sayre, and Brenda Murdoch who all contributed in planning and designing the collection. We would also like to thank all those who contributed to the FAANG tissue collection at USU: Noelle Cockett, Tracy Hadfield, Tom Baldwin, Rusty Stott, Arnaud Van Wettere, Holly Mason, Jaqueline LaRose, Dave Forrester, Corey Wareham, Sarah Behunin, Kara Thornton, Gordon Hullinger (VMES), Alisha Massa, Maria Herndon, Brenda Murdoch, Brian Sayre, Caylee Birge, Codie Durfee, Michelle R. Mousel, Rachael Christianson, Nicole Ineck, Angie Robinson, Dallin Wengert, Kerry Rood, Erica Moscoso, Rickie Warr, Dustin Kinney, Abbey Benninghoff, Sumira Phatak, Kevin Contreras, Braden Abercrombie, and Misha Regouski. This article has been released as a pre-print at bioRxiv (Salavati et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.580580/full#supplementary-material

Supplementary Figure 1 | CAGE library size for each of the 56 tissues analyzed.

Supplementary Figure 2 | The percentage of CAGE tags mapped to each genomic region for Oar rambouillet v1.0 **(A)** and Oar_v3.1 **(B)** reference genome assemblies. The counts were averaged across tissues prior to annotation.

Supplementary Figure 3A | The distribution of tissue specific TSS in 56 tissues of Benz2616. The bar shows the count of tissue specific TSS in each tissue with the proportion being expressed with CTPM > 1 colored in red.

Supplementary Figure 3B | The distribution of tissue specific TSS-Enhancers across the 56 tissues from Benz2616. The bars show the count of tissue specific TSS in each tissue with the proportion being expressed with CTPM > 1 colored in red.

Supplementary Figure 4 | Heatmap of mRNA-Seq and CAGE expression profiles (TPM and CTPM). The correlation was calculated over 52 matched tissues and 5732 transcripts—TSS expressed in all tissues.

Supplementary Figure 5 | Distribution of uni-directional CAGE TSS clusters per annotated gene. **(A)** The histogram of the TSS cluster per gene. **(B)** Detailed table of TSS per gene data underlying the histogram and percentage per total TSS clusters.

Supplementary Table 1 | Details of 5′ linker barcodes and pool ID assigned to each tissue sample.

Supplementary Table 2 | Percentage of tissue-specific CAGE tags mapping to genomic features.

Supplementary Table 3 | Summary of WGBS sequencing and mapping results.

Supplementary Table 4 | Comparison of mRNA-Seq dataset with matched tissues within the CAGE dataset with regards to tissue support criteria.

Supplementary Table 5 | Summary of minimum tissue support calculations for TSSs per gene in each scenario. Tissue support thresholds of 1, 5, 18, 28, and 37 tissues out of total 56 were analyzed.

Supplementary Table 6 | Comparison of this study with other CAGE datasets available as part of the FANTOM5 consortium data release.

Supplementary File 1 | Section 1: Attrition of raw reads at each stage of the analysis pipeline, rationale for selecting the two-thirds representation threshold for mapped CAGE tags and clustering metrics. **Sections 2, 3**: Detailed comparison of mapping of the CAGE tags to the two reference assemblies *Oar_v3.1* and *Oar rambouillet v1.0* and analysis workflow.

Supplementary File 2 | Expression data frames from uni-, bi-directional, long-range linked co-expression clustering and transcript level mRNA-Seq from all 56 tissues (two-thirds representation rule applied).

REFERENCES

An, X., Ma, H., Han, P., Zhu, C., Cao, B., and Bai, Y. (2018). Genome-wide differences in DNA methylation changes in caprine ovaries between oestrous and dioestrous phases. *J. Anim. Sci. Biotechnol.* 9:58. doi: 10.1186/s40104-018-0301-x

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project. *Genome Biol.* 16:57. doi: 10.1186/s13059-015-0622-4

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

Baillie, J. K., Arner, E., Daub, C., De Hoon, M., Itoh, M., Kawaji, H., et al. (2017). Analysis of the human monocyte-derived macrophage transcriptome and response to lipopolysaccharide provides new insights into genetic aetiology of inflammatory bowel disease. *PLoS Genet.* 13:e1006641. doi: 10.1371/journal.pgen.1006641

Beiki, H., Liu, H., Huang, J., Manchanda, N., Nonneman, D., Smith, T. P. L., et al. (2019). Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. *BMC Genomics* 20:344. doi: 10.1186/s12864-019-5709-y

Berger, M. R., Alvarado, R., and Kiss, D. L. (2019). mRNA 5′ ends targeted by cytoplasmic recapping cluster at CAGE tags and select transcripts are alternatively spliced. *FEBS Lett.* 593, 670–679. doi: 10.1002/1873-3468.13349

Bertin, N., Mendez, M., Hasegawa, A., Lizio, M., Abugessaisa, I., Severin, J., et al. (2017). Linking FANTOM5 CAGE peaks to annotations with CAGEscan. *Sci. Data* 4:170147. doi: 10.1038/sdata.2017.147

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., et al. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* 49, 643–650. doi: 10.1038/ng.3802

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816. doi: 10.1038/nature05874

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi: 10.1038/nbt.3519

Cano-Gamez, E., and Trynka, G. (2020). From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.* 11:424. doi: 10.3389/fgene.2020.00424

Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., and Hwang, C.-C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One* 8:e62856. doi: 10.1371/journal.pone.0062856

Clark, E. L., Bush, S. J., McCulloch, M. E. B., Farquhar, I. L., Young, R., Lefevre, L., et al. (2017). A high resolution atlas of gene expression in the domestic sheep (*Ovis aries*). *PLoS Genet.* 13:e1006997. doi: 10.1371/journal.pgen.1006997

Cordier, G., Cozon, G., Greenland, T., Rocher, F., Guiguen, F., Guerret, S., et al. (1990). In vivo activation of alveolar macrophages in ovine lentivirus infection. *Clin. Immunol. Immunopathol.* 55, 355–367. doi: 10.1016/0090-1229(90)90124-9

Ding, B., Gentleman, R., and Carey, V. (2012). *bioDist: Different Distance Measures. R Package Version 1.28.0*.

Doherty, R., and Couldrey, C. (2014). Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment. *Front. Genet.* 5:126. doi: 10.3389/fgene.2014.00126

Edinburgh (2020). *Edinburgh Compute and Data Facility*. Available online at: https://www.ed.ac.uk/is/research-computing-service (Accessed July 6, 2020).

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

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., et al. (2019). Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol.* 17:108. doi: 10.1186/s12915-019-0726-5

Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., De Hoon, M. J. L., Haberle, V., et al. (2014). A promoter-level mammalian expression atlas. *Nature* 507, 462–470. doi: 10.1038/nature13182

Georges, M., Charlier, C., and Hayes, B. (2018). Harnessing genomic information for livestock improvement. *Nat. Rev. Genet.* 20:1. doi: 10.1038/s41576-018-0082-2

Giuffra, E., and Tuggle, C. K. (2019). Functional annotation of animal genomes (FAANG): current achievements and roadmap. *Annu. Rev. Anim. Biosci.* 7, 65–88. doi: 10.1146/annurev-animal-020518-114913

Guo, W., Zhu, P., Pellegrini, M., Zhang, M. Q., Wang, X., and Ni, Z. (2018). CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics* 34, 381–387. doi: 10.1093/bioinformatics/btx595

Hahne, F., and Ivanek, R. (2016). "Visualizing Genomic Data Using Gviz and Bioconductor BT," in *Statistical Genomics: Methods and Protocols*, eds E. Mathé and S. Davis (New York, NY: Springer), 335–351. doi: 10.1007/978-1-4939-3578-9_16

Hannon Lab (2017). *FASTX-Toolkit FASTQ/A Short Reads Pre-Processing Tools*. Available online at: http://hannonlab.cshl.edu/fastx_toolkit/ (accessed July 6, 2020).

Harrison, P. W., Fan, J., Richardson, D., Clarke, L., Zerbino, D., Cochrane, G., et al. (2018). FAANG, establishing metadata standards, validation and best practices for the farmed and companion animal community. *Anim. Genet.* 49, 520–526. doi: 10.1111/age.12736

Ibeagha-Awemu, E. M., and Zhao, X. (2015). Epigenetic marks: regulators of livestock phenotypes and conceivable sources of missing variation in livestock improvement programs. *Front. Genet.* 6:302. doi: 10.3389/fgene.2015.00302

Imada, E. L., Sanchez, D. F., Collado-Torres, L., Wilks, C., Matam, T., Dinalankara, W., et al. (2020). Recounting the FANTOM CAGE–associated transcriptome. *Genome Res.* 30, 1073–1081. doi: 10.1101/gr.254656.119

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J. F., Faraut, T., et al. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* 344, 1168–1173. doi: 10.1126/science.1252806

Joe, H. (1989). Relative entropy measures of multivariate dependence. *J. Am. Stat. Assoc.* 84, 157–164. doi: 10.1080/01621459.1989.10478751

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* 13, 484–492. doi: 10.1038/nrg3230

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 26, 2204–2207. doi: 10.1093/bioinformatics/btq351

Kouno, T., Moody, J., Kwon, A. T.-J., Shibayama, Y., Kato, S., Huang, Y., et al. (2019). C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nat. Commun.* 10:360. doi: 10.1038/s41467-018-08126-5

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lassmann, T. (2015). TagDust2: a generic method to extract reads from sequencing data. *BMC Bioinformatics* 16:24. doi: 10.1186/s12859-015-0454-y

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9:e1003118. doi: 10.1371/journal.pcbi.1003118

Lev Maor, G., Yearim, A., and Ast, G. (2015). The alternative role of DNA methylation in splicing regulation. *Trends Genet.* 5, 274–280. doi: 10.1016/j.tig.2015.03.002

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lizio, M., Deviatiiarov, R., Nagai, H., Galan, L., Arner, E., Itoh, M., et al. (2017). Systematic analysis of transcription start sites in avian development. *PLoS Biol.* 15:e2002887. doi: 10.1371/journal.pbio.2002887

Low, W. Y., Tearle, R., Bickhart, D. M., Rosen, B. D., Kingan, S. B., Swale, T., et al. (2019). Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat. Commun.* 10:260. doi: 10.1038/s41467-018-08260-0

Murdoch, B. M. (2019). The functional annotation of the sheep genome project. *J. Anim. Sci.* 97:16. doi: 10.1093/jas/skz122.029

Naval-Sanchez, M., Nguyen, Q., McWilliam, S., Porto-Neto, L. R., Tellam, R., Vuocolo, T., et al. (2018). Sheep genome functional annotation reveals proximal regulatory elements contributed to the evolution of modern breeds. *Nat. Commun.* 9:859. doi: 10.1038/s41467-017-02809-1

Ogorevc, J., Orehek, S., and Dovč, P. (2016). Cellular reprogramming in farm animals: an overview of iPSC generation in the mammalian farm animal species. *J. Anim. Sci. Biotechnol.* 7:10. doi: 10.1186/s40104-016-0070-3

Pages, H. (2020). *BSgenome: Software Infrastructure for Efficient Representation of Full Genomes and Their SNPs. R Package Version* 1.56.0.

Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A. M.-P., George, N., Fexova, S., et al. (2019). Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* 48, D77–D83. doi: 10.1093/nar/gkz947

Priness, I., Maimon, O., and Ben-Gal, I. (2007). Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics* 8:111. doi: 10.1186/1471-2105-8-111

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Reshef, D. N., Reshef, Y. A., Sabeti, P. C., and Mitzenmacher, M. (2018). An empirical study of the maximal and total information coefficients and leading measures of dependence. *Ann. Appl. Stat.* 12, 123–155. doi: 10.1214/17-AOAS1093

Rexroad, C., Vallet, J., Matukumalli, L. K., Reecy, J., Bickhart, D., Blackburn, H., et al. (2019). Genome to phenome: improving animal health, production, and well-being – a new USDA blueprint for animal genome research 2018–2027. *Front. Genet.* 10:327. doi: 10.3389/fgene.2019.00327

Salavati, M., Caulton, A., Clark, R., Gazova, I., Smith, T. P. L., Worley, K. C., et al. (2020). Global analysis of transcription start sites in the new ovine reference genome (Oar rambouillet v1.0). *bioRxiv* [Preprint]. doi: 10.1101/2020.07.06.189480

Sheep Genomes Database (2020). Available online at: https://sheepgenomesdb.org/ (accessed July 6, 2020).

Song, Q., Decato, B., Hong, E. E., Zhou, M., Fang, F., Qu, J., et al. (2013). A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* 8:e81148. doi: 10.1371/journal.pone.0081148

Tait-Burkard, C., Doeschl-Wilson, A., McGrew, M. J., Archibald, A. L., Sang, H. M., Houston, R. D., et al. (2018). Livestock 2.0 – genome editing for fitter, healthier, and more productive farmed animals. *Genome Biol.* 19:204. doi: 10.1186/s13059-018-1583-1

Takahashi, H., Kato, S., Murata, M., and Carninci, P. (2012). "CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks," in *Methods in Molecular Biology*, ed. J. M. Walker (Totowa, NJ: Humana Press), 181–200. doi: 10.1007/978-1-61779-292-2_11

Thodberg, M., and Sandelin, A. (2019). A step-by-step guide to analyzing CAGE data using R/Bioconductor. *F1000Research* 8:886. doi: 10.12688/f1000research.18456.1

Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R., and Sandelin, A. (2019). CAGEfightR: analysis of 5′-end data using R/bioconductor. *BMC Bioinformatics* 20:487. doi: 10.1186/s12859-019-3029-5

Tippens, N. D., Vihervaara, A., and Lis, J. T. (2018). Enhancer transcription: what, where, when, and why? *Genes Dev.* 32, 1–3. doi: 10.1101/GAD.311605.118

Worley, K. C. (2017). A golden goat genome. *Nat. Genet.* 49, 485–486. doi: 10.1038/ng.3824

Yagi, S., Hirabayashi, K., Sato, S., Li, W., Takahashi, Y., Hirakawa, T., et al. (2008). DNA methylation profile of tissue-dependent and differentially methylated regions (T-DMRs) in mouse promoter regions demonstrating tissue-specific gene expression. *Genome Res.* 18, 1969–1978. doi: 10.1101/gr.074070.107

Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. (2005). Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* 350, 129–136. doi: 10.1016/j.gene.2005.01.012

Ziller, M. J., Hansen, K. D., Meissner, A., and Aryee, M. J. (2015). Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat. Methods* 12, 230–232. doi: 10.1038/nmeth.3152

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.

Copyright © 2020 Salavati, Caulton, Clark, Gazova, Smith, Worley, Cockett, Archibald, Clarke, Murdoch and Clark. 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: 07 January 2021
doi: 10.3389/fgene.2020.612031

Genome-Wide Histone Modifications and CTCF Enrichment Predict Gene Expression in Sheep Macrophages

*Alisha T. Massa[1], Michelle R. Mousel[2,3], Maria K. Herndon[1], David R. Herndon[2], Brenda M. Murdoch[4,5] and Stephen N. White[1,2,5]**

[1] Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, WA, United States, [2] Animal Disease Research Unit, Agricultural Research Service, United States Department of Agriculture, Pullman, WA, United States, [3] Paul G. Allen School for Global Animal Health, Washington State University, Pullman, WA, United States, [4] Department of Animal and Veterinary Science, University of Idaho, Moscow, ID, United States, [5] Center for Reproductive Biology, Washington State University, Pullman, WA, United States

Alveolar macrophages function in innate and adaptive immunity, wound healing, and homeostasis in the lungs dependent on tissue-specific gene expression under epigenetic regulation. The functional diversity of tissue resident macrophages, despite their common myeloid lineage, highlights the need to study tissue-specific regulatory elements that control gene expression. Increasing evidence supports the hypothesis that subtle genetic changes alter sheep macrophage response to important production pathogens and zoonoses, for example, viruses like small ruminant lentiviruses and bacteria like *Coxiella burnetii*. Annotation of transcriptional regulatory elements will aid researchers in identifying genetic mutations of immunological consequence. Here we report the first genome-wide survey of regulatory elements in any sheep immune cell, utilizing alveolar macrophages. We assayed histone modifications and CTCF enrichment by chromatin immunoprecipitation with deep sequencing (ChIP-seq) in two sheep to determine *cis*-regulatory DNA elements and chromatin domain boundaries that control immunity-related gene expression. Histone modifications included H3K4me3 (denoting active promoters), H3K27ac (active enhancers), H3K4me1 (primed and distal enhancers), and H3K27me3 (broad silencers). In total, we identified 248,674 reproducible regulatory elements, which allowed assignment of putative biological function in macrophages to 12% of the sheep genome. Data exceeded the FAANG and ENCODE standards of 20 million and 45 million useable fragments for narrow and broad marks, respectively. Active elements showed consensus with RNA-seq data and were predictive of gene expression in alveolar macrophages from the publicly available Sheep Gene Expression Atlas. Silencer elements were not enriched for expressed genes, but rather for repressed developmental genes. CTCF enrichment enabled identification of 11,000 chromatin domains with mean size of 258 kb. To our knowledge, this is the first report to use immunoprecipitated CTCF to determine putative topological domains in sheep immune cells. Furthermore, these data will empower phenotype-associated mutation discovery since most causal variants are within regulatory elements.

Keywords: insulator, promoter, enhancer, innate immunity, alveolar macrophage, sheep, epigenetics, ChIP-seq

OPEN ACCESS

Edited by:
Christa Kuehn,
Leibniz Institute for Farm Animal Biology (FBN), Germany

Reviewed by:
Alex Clop,
Centre for Research in Agricultural Genomics (CRAG), Spain
Thomas Hall,
University College Dublin, Ireland

***Correspondence:**
Stephen N. White
stephen.white@usda.gov

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

Received: *30 September 2020*
Accepted: *30 November 2020*
Published: *07 January 2021*

Citation:
Massa AT, Mousel MR, Herndon MK, Herndon DR, Murdoch BM and White SN (2021) Genome-Wide Histone Modifications and CTCF Enrichment Predict Gene Expression in Sheep Macrophages. Front. Genet. 11:612031. doi: 10.3389/fgene.2020.612031

INTRODUCTION

Increasingly, livestock researchers are identifying functional variants outside of genes as associated with valuable production traits, supporting the need to molecularly annotate regulatory elements (Ibeagha-Awemu and Zhao, 2015; Zhao et al., 2015; Wang et al., 2017). DNA regulatory elements are sequences associated with a reproducible biological function that can control gene expression through epigenetic modifications (Birney et al., 2007). Human studies consistently document the importance of variants within CRE sequences to critical phenotypic traits. Several groups estimated that over 90% of causal mutations that explain phenotypic variation laid outside of genes within regulatory elements (Hindorff et al., 2009; Maurano et al., 2012; Albert and Kruglyak, 2015). Currently, little is known regarding *in vivo* tissue annotation of regulatory elements in livestock species (Villar et al., 2015; Zhao et al., 2015; Wang et al., 2017, 2018; Naval-Sanchez et al., 2018; Nguyen et al., 2018; Fang et al., 2019; Hall et al., 2020; Kingsley et al., 2020). Therefore, the FAANG consortium recognized this need and formed a global network of researchers for epigenetic discovery in food animal species (Andersson et al., 2015; Tuggle et al., 2016; Giuffra and Tuggle, 2019).

Types of *cis*-acting, DNA regulatory elements (CREs) that control gene expression include active promoters and enhancers, primed enhancers, silencers, and insulators (Dunham et al., 2012). Although CTCF has also been associated with *trans*-acting regulation (Handoko et al., 2011). Promoters are stretches of DNA located at the TSS of genes and serve as scaffolding for promotion, assembly, and initiation of transcription (Birney et al., 2007). Enhancers act more distally and in an orientation independent fashion to activate gene transcription (Banerji et al., 1981). However, enhancers classically must be within the same three-dimensional chromatin domain as their target gene (Schaffner, 2015). Chromatin immunoprecipitation and sequencing (ChIP-seq) (Barski et al., 2007; Johnson et al., 2007) of the post-translational modification marks histone 3 lysine 27 acetylation (H3K27ac) and histone 3 lysine 4 trimethylation (H3K4me3) allowed genome wide identification of active enhancers and active promoters, as demonstrated in early ChIP-seq assays (Heintzman et al., 2007; Won et al., 2008). In addition, H3K27ac often overlapped H3K4me3 regions in active promoters of highly expressed genes (Wang et al., 2008). Potential enhancers that are epigenetically primed but not fully active are marked by histone 3 lysine 4 monomethylation (H3K4me1) alone (Heintzman et al., 2007). H3K4me1, in conjunction with H3K27ac, is found at active distal enhancers (Jambhekar et al., 2019). Lastly, histone 3 lysine 27 trimethylation (H3K27me3) marks broad regions that are transcriptionally repressed or silenced (O'Geen et al., 2011) as the modification is established by the activity of polycomb complexes that help to supercoil the heterochromatin (Dunham et al., 2012; Aranda et al., 2015). Uniquely, some regions are marked simultaneously by methylation at H3K4 and H3K27 termed bivalent regulatory chromatin. Bivalent histone modifications (the combination of H3K4me3 and H3K27me3) were reported to responsively shift gene expression from a poised or primed state to active transcription, most widely studied in embryonic stem cells (Vastenhouw and Schier, 2012). Tissue resident macrophages share features with embryonic stem cells in that they retain the ability to replenish local cell populations (Sieweke and Allen, 2013).

Since regulatory element functions are dependent on three-dimensional chromatin structure within the nucleus, we also sought to define the boundaries of chromatin loops. Chromosomes are compartmentalized into physically interacting segments called TADs (Dixon et al., 2012; Nora et al., 2012) also known as chromatin loops (Rao et al., 2014; Bonev and Cavalli, 2016) that have shared function. Chromatin immunoprecipitation of CCCTC-binding factor (CTCF), denotes insulator regions which anchor domain boundaries (Zhou et al., 2010). The function of *cis*-acting regulatory elements, including those marked by H3K4me3, H3K27ac, and H3K4me1, is generally constrained to genes within the same domain. While histone post translational modifications serve as predictive signals of specific types of regulatory elements, and functions are conserved across species, the exact sequence of the regulatory element is generally not well conserved (Birney et al., 2007; Villar et al., 2015). Therefore, experimental determination of regulatory elements within a variety of tissues is necessary to fully understand unique gene regulatory networks within food animal species.

Host regulatory element variation likely plays a significant role in macrophage immune response to infections. Immunity-related gene regulatory variation has potential to affect production efficiency by altering both the global and tissue-specific transcriptome (Rauw, 2012; Lavin et al., 2014). For example, recent work showed that macrophages can develop trained immunity or innate immune memory which provides non-specific enhanced protection after exposure to pathogens. This non-adaptive immunological memory is reversibly retained in the epigenome of macrophages (Saeed et al., 2014). Trained immunity may be dependent on genetic variants in genes separate from those involved in classical immunological memory (van der Heijden et al., 2018). Furthermore, Salavati et al. (2019) found that sheep immune-related tissues including macrophages have moderate to extreme allele-specific expression. Allele specific expression is commonly attributed to *cis*-acting regulatory variation which provides an understandable mechanism for parent-of-origin or tissue-specific gene expression since *cis*-acting regulatory elements are physically linked to a single allele copy. We have chosen to study alveolar macrophages from sheep lungs both for their tissue-specific gene expression

Abbreviations: ChIP-seq, chromatin immunoprecipitation with sequencing; CREs, *cis*-regulatory elements; DPBS, Dulbecco's phosphate buffered saline; ENCODE, Encyclopedia of DNA elements Project Consortium; FAANG, Functional Annotation of Animal Genomes Consortium; FDR, false discovery rate; FRiP, fraction of reads in peaks; GO, gene ontology; H3K4me1, histone three lysine four monomethylation; H3K4me3, histone three lysine four trimethylation; H3K27ac, histone three lysine 27 acetylation; H3K27me3, histone three lysine 27 trimethylation; kb, kilobases; N-ChIP, native chromatin immunoprecipitation with sequencing; NSC, normalized strand correlation; RSC, relative strand correlation; TAD, topologically associating domain; TPM, transcripts per million; TSS, transcription start site; X-ChIP, cross-linked chromatin immunoprecipitation with sequencing.

and as a representative cell type to identify immunity related regulatory elements.

Macrophages, as part of the immune system, are a core tissue identified by the FAANG consortium for epigenetic studies (Andersson et al., 2015). Macrophages are professional phagocytes that function in cell-mediated innate immunity at interfaces of the body with the environment and in adaptive immunity as professional antigen presenting cells. Alveolar macrophages in the lungs serve as infection surveillance against airborne pathogens. They also participate in homeostasis in their local tissue microenvironment, a function of specialized tissue-resident macrophages in essentially every organ in the body (Lavin et al., 2014). Macrophages can be hijacked by pathogens like *Mycobacterium bovis*, Ovine lentivirus, *Coxiella burnetii*, *Mycoplasma ovipneumoniae*, *Brucella melitensis*, and *Salmonella enterica*, that cause zoonotic and economically important diseases in sheep: tuberculosis, ovine progressive pneumonia, Q-fever, atypical pneumonia (part of respiratory disease complex), brucellosis, and salmonellosis, respectively (Gendelman et al., 1986; Niang, 1992; Niang et al., 1997; Shannon and Heinzen, 2009; Blacklaws, 2012; Hall et al., 2020). Many of these infectious agents are intracellular organisms that can sequester within host macrophages from the full force of the immune system and manipulate antigen processing and presentation. Elucidation of variation within DNA regulatory elements will aid detection of disease resistant animals that reduce infectious burden within flocks. Genetic determination of resistance and susceptibility can be a crucial tool for disease eradication from individual animal, herd health, and One Health perspectives (Sundberg and Schofield, 2009).

Our objectives for this experiment were to develop a catalog of core histone modifications and of CTCF enriched boundaries in sheep macrophages to locate and functionally annotate regulatory elements. Since CREs compose a far greater portion of the genome than protein coding genes (Moore et al., 2020) lack of annotation in the sheep represents a critical knowledge gap. To the authors' knowledge, this work is the first epigenetic analysis based on ChIP-seq in any sheep immune cell. We chose native ChIP-seq for greater enrichment and reproducibility of signal (David et al., 2017). As a method of validation, we compared genes near discovered regulatory element regions to RNA-seq data in alveolar macrophages from the Sheep Gene Expression Atlas (Clark et al., 2017). These data presented here will serve as functional epigenetic annotation in sheep immune cells to aid future work on phenotypic-associated variation for important food production, fiber, and immunity related traits.

MATERIALS AND METHODS

Alveolar Macrophage Cell Collection

Animals were cared for and handled according to protocols approved by the Institutional Animal Care and Use Committee at Washington State University under Animal Subject Approval Form 4618. Sheep were humanely euthanized with intravenous sodium pentobarbital and lungs were removed firstly during routine postmortem examination by a veterinarian. No gross lesions were detected in the sheep. Alveolar macrophages were collected from the lungs of 2, 1-year-old, clinically healthy, crossbred (Suffolk, Polypay, and Targhee) ewes using methods modified from those previously described (Gendelman et al., 1984; Cordier et al., 1990; Clark et al., 2017). Briefly, bronchoalveolar lavage fluid was collected by serial lavages with sterile DPBS (Mg^{2+} Ca^{2+} free). Cells were isolated from collected lavage fluid by centrifugation (400 × *g* for 10 min) and washed with DPBS at room temperature. Erythrocytes within the pellet were lysed by suspension in sterile water for 30 s. The harvested cells were confirmed to be morphologically consistent with macrophages on cytological evaluation as others have reported (Sheehan et al., 2005). Cells were stained with trypan blue to assess membrane integrity then counted with an automated cytometer (Nexcelom Bioscience, Lawrence, MA, United States). Aliquots of 5×10^7 live macrophages were suspended in cryopreservation medium (CryoStor CS10, BioLife Solutions, Bothell, WA, United States) and slowly frozen to −80°C in isopropyl alcohol baths (Mr. Frosty, Thermo Fisher Scientific, Waltham, MA, United States) for short term storage.

Chromatin Immunoprecipitation and Sequencing

Isolation of Native Chromatin

Native chromatin isolation and immunoprecipitation was modified from methods published previously for tissues (Wagschal et al., 2007; Maunakea et al., 2010; David et al., 2017; Naval-Sanchez et al., 2018). Additional protocol details are included in **Supplementary Methods 1** and provided on the FAANG data portal (see **Supplementary Methods 1**[1]). Cells and buffers were maintained on ice during all steps. Nuclei were isolated from approximately 5×10^7 unfixed, thawed cells firstly by incubation on ice in hypotonic buffer [0.3 M sucrose, 60 mM KCl, 15 mM NaCl, 5 mM MgCl2, 0.1 mM EGTA, 15 mM Tris-HCl, pH 7.5, and HALT protease inhibitor cocktail (Thermo Fisher Scientific)]. Sodium butyrate 5 mM was included to inhibit histone deacetylases during processing. Next, 0.2% IGEPAL CA-630 detergent (Sigma-Aldrich, St. Louis, MO, United StatesA) was added to the suspension with gentle Dounce homogenization using a tight pestle. The nuclei suspension was then carefully layered onto 8 mL of buffer containing 1.2 M sucrose and centrifuged at 4,000 × *g* for 20 min at 4°C. Detergent layers were removed carefully from the nuclei pellet, then the pellet was resuspended in micrococcal nuclease digestion buffer with protease inhibitors. The pellet was briefly vortexed and then 60 Kunitz units of micrococcal nuclease (M0247S, New England Biolabs, Ipswich, MA, United States) was added for 12 min incubation at 37°C to digest the chromatin into mono- and di-nucleosomes. Addition of 20 mM EGTA quenched the digestion reaction and soluble chromatin fragments were recovered in the supernatant by probe-free, cup horn sonication for 2 × 30 s on ice at high power (260 watts). A sample of purified digested chromatin was checked for adequate fragmentation on an agarose gel and on a fragment bioanalyzer (Agilent,

[1]https://data.faang.org/api/fire_api/assays/WSU_SOP_Native_ChIP-seq_Protocol_2019.pdf

Santa Clara, CA, United States) to ensure oligonucleosome fragment lengths within 100–450 base pairs. Average chromatin fragment size was approximately 150 bp in both biological replicates. Chromatin concentration was then measured by fluorescence quantification using the Qubit dsDNA HS kit (Thermo Fisher Scientific).

Immunoprecipitation of Chromatin

Input nucleosomal DNA for each ewe were used as negative controls (no addition of antibody or magnetic beads). Chromatin for immunoprecipitation was pre-cleared by incubation with protein G coupled magnetic beads (Dynabeads, Invitrogen, Waltham, MA, United States). Antibodies to the following targets were used for each chromatin immunoprecipitation: five microliters of anti-H3K4me3, anti-H3K27ac, anti-H3K27me3, anti-H3K4me1, and 10 microliters of anti-CTCF (see **Supplementary Table 2** for catalog numbers and lots). Antibodies were pre-bound to magnetic beads at 4°C then the antibody-bead complexes were added to the diluted (50 mM NaCl, 50 mM Tris-HCl pH 7.5, and 5 mM EDTA), fragmented chromatin for overnight incubation in one milliliter volumes with rotation at 4°C. Enriched chromatin was harvested by magnetic bead pulldown, washed with increasing salt buffers (75–175 mM NaCl) to remove non-specific chromatin, and DNA was purified with the iPure kit (Diagenode, Liege, Belgium) as per manufacturer's recommendation, excluding the cross-linking reversal step. Total amount of immunoprecipitated DNA obtained for each sample was determined by Qubit dsDNA HS analysis.

Library Preparation and Sequencing

Sequencing libraries were prepared from 7.5 ng of immunoprecipitated or input control DNA using Truseq ChIP Sample Prep kit (Illumina, San Diego, CA, United States) following the manufacturer's protocol with 15 PCR cycles to minimize duplication bias and size selection of 250–600 bp to include the bulk of immunoprecipitated fragments ligated to adapters. Preparation for multiplexing was accomplished by utilizing indexing adapters included in the kit. ChIP library size was assessed by Fragment Analyzer (Advanced Analytical Technologies, Ankeny, IA, United States) with the High Sensitivity NGS Fragment Analysis Kit (Agilent, Ankeny, IA, United States), and library concentration was determined by StepOnePlus Real-Time PCR System (Thermo Fisher Scientific) with the KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, MA, United States). Each library was diluted to 4 nM with RSB (10 mM Tris-HCl, pH 8.5), followed by denaturation with 0.1 M NaOH, and 20 pM was clustered in a high-output flow cell using HiSeq Cluster Kit v4 on a cBot (Illumina). After cluster generation, the flow cell was loaded onto HiSeq 2500 for sequencing using HiSeq SBS kit v4 (Illumina). DNA was sequenced with a read length of 50 bp from a single end generating between 41.96 million and 80.15 million filter-passed reads for each library. These were derived from a total of 644,923,132 reads for the experiment that passed initial sequencing quality filters (97% pass-rate) (**Table 1** and **Supplementary Data 15**).

TABLE 1 | Summary of read counts from ChIP-seq assays.

ChIP-seq target	Ewe A		Ewe B	
	Total reads	Usable fragments	Total reads	Usable fragments
Input Control	43,513,683	33,901,350	44,568,055	34,245,907
H3K4me3	41,963,508	23,545,621	43,705,079	23,796,392
H3K27ac	46,534,416	33,599,474	46,205,646	34,309,656
H3K27me3	72,339,834	45,918,247	68,038,026	47,130,715
H3K4me1	78,646,889	59,258,883	80,153,553	48,412,817
CTCF	44,138,652	28,139,080	42,575,226	28,717,487

Total reads include all raw data from sequencing. Usable fragments are defined consistent with ENCODE standards as reads that map to a single best location (quality filter -q 5), with optical duplicates removed as flagged by MACS2. Additional mapping statistics are in the ***Supplementary Data 15****.*

Analysis of ChIP-Seq Data

Sheep ChIP-seq sequencing files generated for this article are publicly available in the ENA database and FAANG data portal under project accession PRJEB40528 (ERP124181). Optional parameters used for all bioinformatics tools and detailed bioinformatics protocol are included in **Supplementary Table 3**.

Sequencing data bcl files were converted to fastq format and adaptor sequences were trimmed using bcl2fastq2 (Illumina). Reads were quality checked with FastQC software (Andrews and Babraham, 2016) with attention to duplication rate (**Supplementary Data 15**). Sequence reads were mapped to the unmasked Rambouillet sheep genome (Oar_rambouillet_v1.0, GCA_002742125.1, Worley, 2017; Salavati et al., 2020), that excludes the mitochondrial genome, with BWA v0.7.17 (Li and Durbin, 2009) (see **Supplementary Data 15** and **Supplementary Data 4** supplementary results for additional mapping details). Reads were sorted and indexed with Picard v2.9.2[2]. Reads were filtered for quality and unique mapping with SAMtools v1.9 (Li et al., 2009). Peaks for histone modifications and CTCF were found for each animal individually using MACS2 v2.1.1 at FDR cut-offs of less than 5% (Zhang et al., 2008; Feng et al., 2012). Effective genome size of the sheep was specified as 2.62×10^9 bp based on the Golden Path Length from ENSEMBL. The broad peak calling option in MACS2 was enabled to calculate both narrow peaks and broad block binding of the H3K27me3 and H3K4me1 datasets. A third set of peaks were called from pooled reads from both animal replicates to maximize sensitivity; these pooled peaks were subsequently filtered for only those called in both individual animals. Overlap between all three peak sets, each individual animal and the pooled reads, were determined with bedtools v2.26.0 and bedops v2.4.38 to create the reproducible consensus peaks (Quinlan and Hall, 2010; Neph et al., 2012). These reproducible consensus peaks were used for all downstream analysis of regulatory elements. Regulatory elements were categorized into active promoters (H3K4me3-enriched regions, with or without overlapping H3K27ac enrichment), active enhancers (all regions enriched for H3K27ac, and

[2]http://broadinstitute.github.io/picard/

regions with H3K27ac only), primed enhancers (H3K4me1-enriched regions), silencers (H3K27me3-enriched), and insulator chromatin domain boundaries (CTCF-enriched).

The called peaks were annotated with the nearest gene and genomic feature type using the annotatePeaks.pl program in HOMER v4.10.4 (Heinz et al., 2010) and the NCBI *Ovis aries* Refseq Annotation Release 103 (O'Leary et al., 2016) for the Oar_Rambouillet_v1.0 genome (GCF_002742125.1). The definition of genomic promoter features was manually adjusted to regions within 2 kb of any gene TSS in the Annotation Release. GO analysis was completed from ChIP-seq target associated gene lists with PANTHER (Mi et al., 2019). HOMER findMotifs.pl was used to scan consensus peaks for transcription factor protein binding motifs (see **Supplementary Methods 1** for further details). Correlation analysis and conversion of BAM files to normalized, input-subtracted bigwig files for visualization was completed with deepTools v3.3.0 (Ramírez et al., 2016) (**Supplementary Methods 1**).

Comparison of ChIP-Seq Data to Public RNA-Seq Data

Processed gene expression data in sheep alveolar macrophages from publicly available mRNA-seq datasets were obtained from supplementary files provided by Clark et al. "**Supplementary Dataset 1**. Gene expression level atlas as TPM (unaveraged)" available at https://doi.org/10.1371/journal.pgen.1006997.s004 (Clark et al., 2017). The authors also provided the processed data available for download through the University of Edinburgh DataShare portal at http://dx.doi.org/10.7488/ds/2112. This processed data was derived from paired end alveolar macrophage transcriptomic RNA-seq from two adult, females: a Texel × Scottish Blackface available at the ENA database under study accession number PRJEB19199[3] at sample accession SAMEA5535418 run accession ERR2074323 (Clark et al., 2017) and a Texel from study accession PRJEB6169[4], published previously (Jiang et al., 2014).

The processed RNA-seq data from female alveolar macrophages was filtered by genes expressed equal to or greater than 1.0 TPM in at least one animal. Mitochondrial genes were removed as ChIP-seq data is from nuclear chromatin only. This yielded a list of 12,042 genes expressed in alveolar macrophages from either individual female at TPM $\geq$ 1. These genes were then ranked by average TPM for comparison to ChIP-seq peak enrichment. Consensus BED files for each histone modification were annotated with ChIP-seq read count per peak region from the pooled mapped reads of both Crossbred ewes, then ranked from highest to lowest by count. Rank of peaks by total read count and their corresponding nearest gene were compared to rank of genes from RNA-seq TPM with Spearman's Rho correlation test since the data was non-parametric. Unidentified *LOC* open reading frames that were mapped to Oar_v3.1 without gene names and that could not be converted to an open reading frame in Oar_Rambouillet_v1.0 with NCBI Genome Remap were removed before comparisons.

[3]http://www.ebi.ac.uk/ena/data/view/PRJEB19199

[4]http://www.ebi.ac.uk/ena/data/view/PRJEB6169

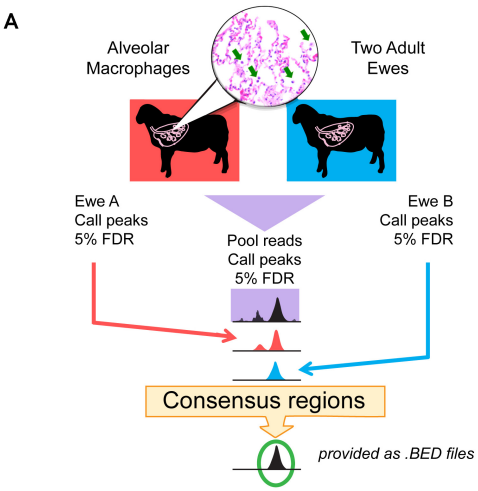
RESULTS

Summary of Quality Metrics

Chromatin immunoprecipitation and sequencing for four histone modification marks: H3K4me3, H3K27ac, H3K4me1, H3K27me3, and CTCF were completed on two animal replicates to identify regulatory elements in sheep alveolar macrophages. Negative controls consisted of input fragmented chromatin for each animal sequenced to a similar depth. Mean mapping rate for raw reads was 98.58% to the Rambouillet genome assembly (Oar_rambouillet_v1.0) (see **Supplementary Data 15** for detailed mapping rates). Non-duplicated fractions of reads were high between 0.80 and 0.94 indicating good library quality. Usable fragments exceeded 23 million reads for all narrow marks and 45 million reads for broad marks (**Table 1**). Correlation of mapped filtered reads for all ChIP-seq datasets sorted by each chromatin mark rather than by individual animal (**Supplementary Data 15** and **Supplementary Figure 5**). This confirmed reproducibility of antibody enrichment between the two animal replicates (Pearson's correlation coefficient: 0.94–0.99) (see **Supplementary Figures 5, 6** for additional animal replicate comparisons). Cumulative enrichment "fingerprint" plots showed significant enrichment above the background, particularly for narrow marks such as H3K4me3 (**Supplementary Figure 7**). NSC and RSC values confirmed significant enrichment in immunoprecipitated datasets compared to input controls, exceeding 1.05 and 2.17, respectively, in all datasets (**Supplementary Data 15**).

Regulatory Element Region Characteristics, GC Content, and Genome Coverage

Regulatory elements were defined by regions of ChIP-seq signal enrichment along the genome for each of the five chromatin marks; total analysis included ten antibody-enriched, epigenomic datasets from alveolar macrophages. Significant regions were called at 5% FDR in both individual animals and then in pooled reads. In total 491,635 and 446,798 regulatory elements were defined in each individual animal (**Supplementary Figure 8** and **Supplementary Data S15-Table 3**). Together regulatory elements covered between 8.79 and 8.23% of the genome in individual animals. We then filtered the set of significantly enriched regions in the pooled reads to select only those that were also significant in both individual animals (see **Figure 1A** for study design). We termed these reproducible regions of signal enrichment as consensus regulatory elements. This yielded 248,674 consensus regulatory elements in sheep alveolar macrophages (**Figure 1B**). Consensus regions were putatively assigned to regulatory element classes. Active *cis*-acting regulatory elements include 71,933 regions marked by H3K4me3 classified as promoters and 68,818 marked by H3K27ac grouped as enhancers or highly active regions. Regions enriched for H3K4me1, considered primed and active enhancers were discovered at 31,800 genomic locations that included both broad and narrow regions of signal enrichment. Silencers were regions with signal enrichment for H3K27me3 found at 53,879 broad regions that cover long stretches of DNA. Lastly, 22,244



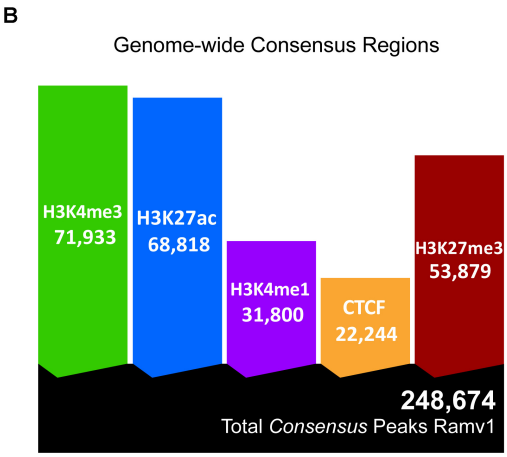


FIGURE 1 | Study design and total consensus regulatory elements identified in alveolar macrophages. **(A)** Schematic overview of the study design for determining consensus regulatory element regions, statistically reproducible with a FDR cut-off of 5% in all three datasets: pooled reads, ewe A, and ewe B. **(B)** Native ChIP-seq yielded many reproducible consensus regions for all chromatin marks. Ramv1 is Oar_rambouillet_v1.0 genome.

very narrow regions marked by CTCF were identified that denoted genomic locations of insulators indicative of chromatin domain boundaries. Consensus regulatory elements were used for further analysis since there was acceptable agreement amongst animal replicates (**Supplementary Datas 5, 6, 15**).

Altogether, consensus regulatory elements from macrophages cover 11.77% of the sheep genome. Promoter signal enrichment covered 2.77% of the genome and regions were narrow, short stretches of DNA, with a median length of 0.81 kb (**Table 2**). Promoter regions had higher GC nucleotide content compared with other regulatory elements and the background GC content of the sheep genome (**Table 2**). Active enhancers (H3K27ac) comprised a slightly longer portion of the genome than promoters. Active enhancer regions also had increased GC nucleotide content compared to the genomic average but less so than promoters. Broad and narrow primed enhancers

TABLE 2 | Consensus regulatory element region details.

Chromatin mark	Total genome coverage %	Median region length in bp (Q1–Q3)		Average GC base content %
H3K4me3	2.77	818	(534–1,320)	47.9
H3K27ac	2.85	843	(542–1,399)	43.2
H3K4me1	3.76	2,419	(1,435–4,059)	40.9
CTCF	0.71	764	(534–1,089)	39.4
H3K27me3	6.05	2,488	(1,542–3,962)	38.4

The Oar_Rambouillet_v1.0 genome assembly was used to calculate coverage and GC content for regions captured by each ChIP-seq mark. The assembly is 2,869.9 megabases (Mb) with an estimated effective genome size of 2,620 Mb and calculated 41.9% GC base content. Q1–Q3 is quartile 1 and quartile 3 for interquartile range.

marked by H3K4me1 covered larger regions of the genome (**Table 2**) and had neutral to mildly depleted GC content. Insulator regions enriched for CTCF were fewer and narrowest with a median length of 0.76 kb occupying the least percentage of the genome. Silencers marked by H3K27me3, covered the largest portion of the genome at 6.05% and displayed broad blocks of signal enrichment at 2.5 kb median length. Silencer regions were markedly depleted of GC nucleotide content compared with other regulatory elements and the genome background content.

Regulatory Elements Have Combinations of Multiple Chromatin Marks

Detected regulatory element regions had either a single type of chromatin mark or a combination of enrichment from multiple marks in that stretch of DNA (**Figure 2**). Most combinations of marks were between those associated with active gene expression (H3K4me3, H3K27ac, and H3K4me1) whereas the repressive mark H3K27me3 had much fewer regions with overlap by another mark. Generally, H3K4me3 and H3K27ac active marks had greater numbers of overlapping regions and H3K27me3 silencer regions had few overlapping regions with either active mark. Boundary regions between these two types of chromatin were often marked by CTCF and H3K4me1 enrichment. Shown in **Figure 2**, promoters and enhancers captured 54–65% of the same regions. However, not all promoters marked by H3K4me3 appeared to be active as only 40,112 of them were also marked by H3K27ac. In fact, H3K4me3 regions were also occasionally enriched for the repressive mark H3K27me3. Promoter regions that did not have H3K27ac signal were mostly enriched for only H3K4me3 (22,036; 70% self only) and did not have overlap with other chromatin marks.

We also analyzed enhancer associated overlap of marks to elucidate primed, transitional, and active enhancers used in sheep macrophages. Overall, 87,458 putative enhancer regions were identified as those enriched for H3K27ac or H3K4me1, or a combination of both. H3K27ac enrichment was found at 28% of the regions marked by H3K4me1 which yielded 8,932 putative highly active enhancers (**Supplementary Data 15**). However, we also found that 19,569 putative active enhancers were exclusively marked by H3K27ac (self-only) and were not overlapped by enrichment for H3K4me1 suggestive of regions with a different

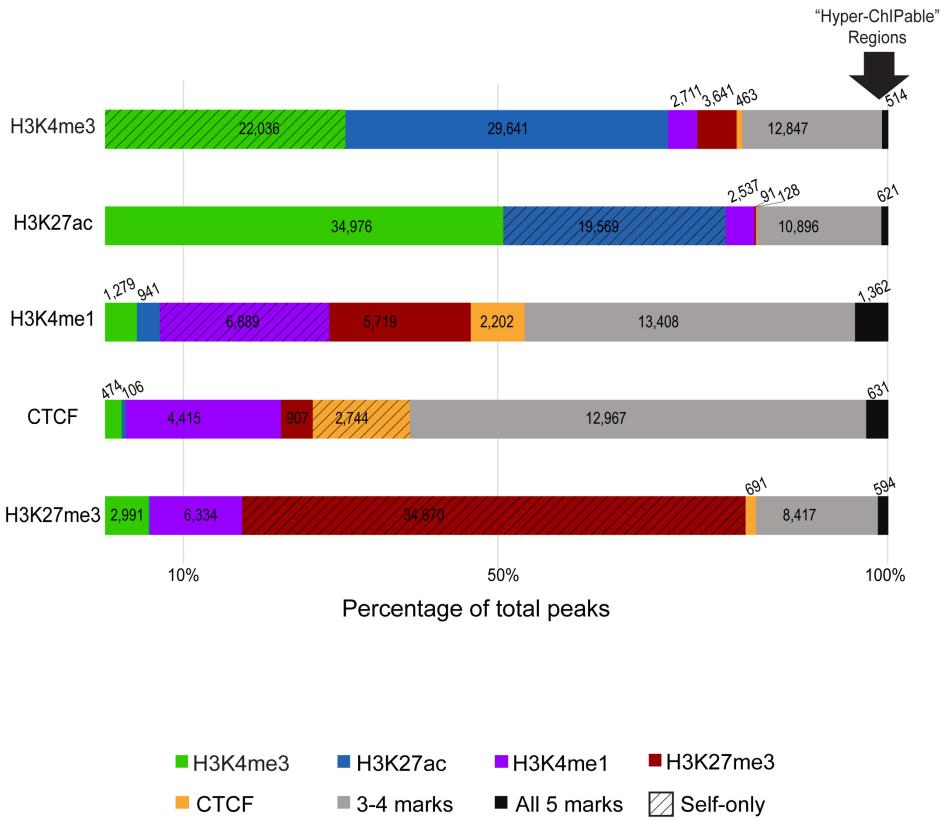


FIGURE 2 | Overlaps of chromatin marks with one another as a percentage of consensus regions for each ChIP-seq target. The segmented bar graphs show regions enriched in each respective ChIP-seq mark color coded by overlap with other marks, the number of peaks from that respective mark that overlap with peaks from other marks are noted in each color-coded segment. Textured bars are regions exclusive to that mark (only marked by self) that did not have enrichment for other marks. Gray bars are regions that have three or four overlapping chromatin marks in that region. On the far-right, the arrow indicates black bars representative of regions with significant signal enrichment in all immunoprecipitated datasets with no specificity for the ChIP-seq antibody target, these were considered putative "hyperChIPable" regions.

cis-acting regulatory function. Primed enhancers, with H3K4me1 signal enrichment, and CTCF -enriched insulators shared notable overlap since approximately 80% of total CTCF regions were also marked by H3K4me1 (**Supplementary Data 15**). Both CTCF and H3K4me1 signal enriched regions had greater overlap by multiple marks (42–58%, **Figure 2** gray bars) than in other immunoprecipitated datasets.

Most silencer regions, 64%, are only enriched for H3K27me3 signal consistent with the expected prediction of heterochromatin that would exclude the other ChIP-seq targets we assayed (**Figure 2**, red hashed bar). Approximately 15% of silencer regions have some overlap with H3K4me1 indicative of primed or transitional regulatory elements at boundary regions. Acetylation and trimethylation of H3K27 are essentially never found in the same regions (0.1%, **Supplementary Data 15**), except where regions were enriched in most or all immunoprecipitated datasets. Interestingly, 1520 genomic regions had significant signal enrichment in all histone modification and CTCF datasets (**Figure 2**, black bars).

Regulatory Element Annotations and Genomic Localization

Each regulatory region was binned into a genomic category (promoter, intron, exon, or intergenic) and annotated with the nearest Refseq gene. The majority of H3K4me3 enriched regions were located within genes (intron and exon) or near the 5′ end of genes within 2 kb of TSS annotated as promoter regions (**Figure 3A** and **Supplementary Figure 9**). Twenty-eight percent of all H3K4me3 enriched regions were within 2 kb of the annotated TSSs of genes and pseudogenes. Nearly half (49%) of the regions distal to the TSS were within the first intron or first exon of genes. The pattern of H3K4me3 signal around gene TSSs was bimodal with high enrichment

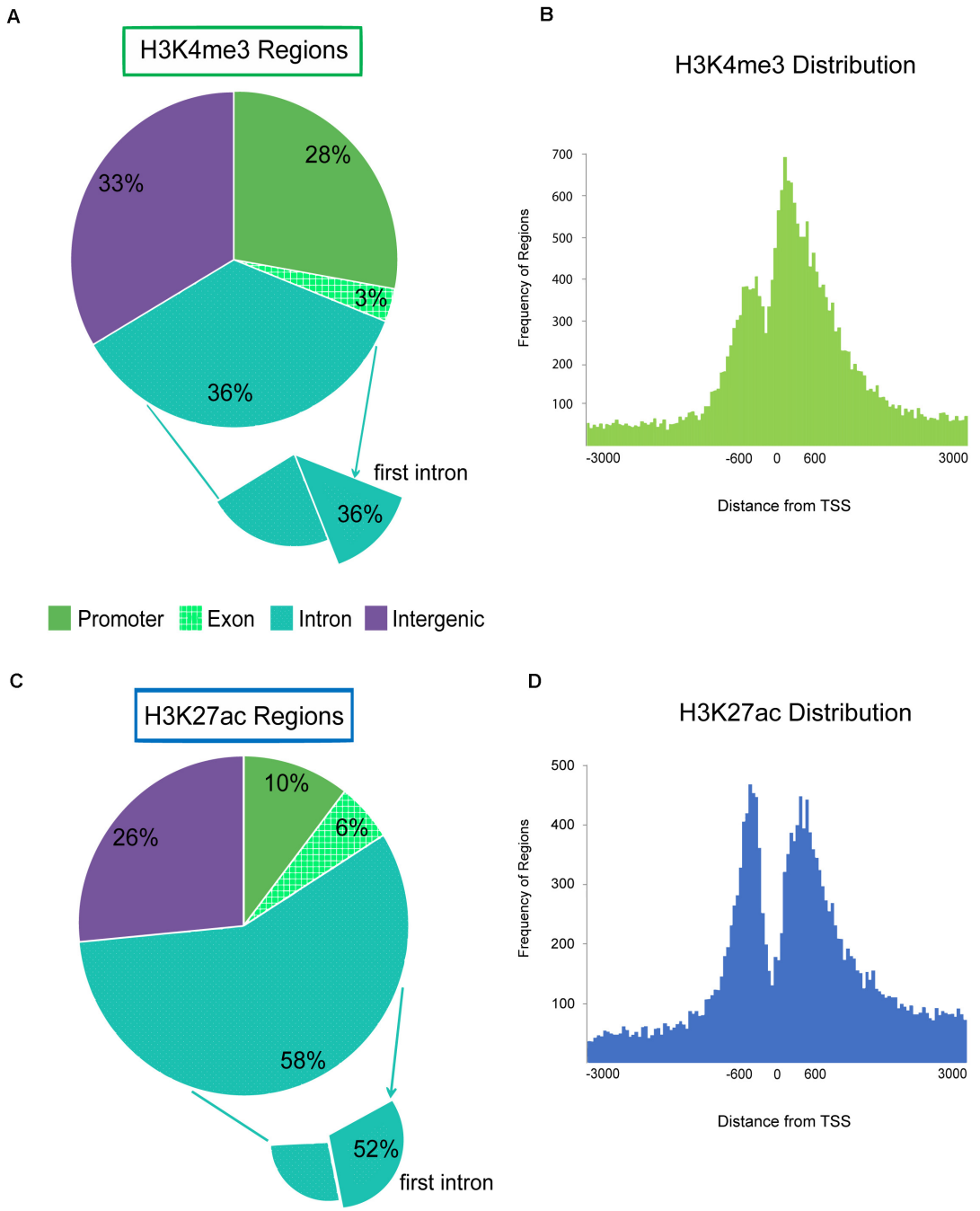


FIGURE 3 | Genomic localization of active promoters and enhancers, H3K4me3 and H3K27ac. Pie charts display the genomic feature (promoter, exon, intron, or intergenic) that regions from each active ChIP-seq target fell within. Promoter is defined as features that fall within 2 kb of the 5′ end of genes. Histograms display the distribution of ChIP-seq regions around the transcription start site (TSS) 0 at the 5′ end of genes by distance in bp. ChIP-seq regions were binned into number per 50 bp segment. **(A)** Genomic locations of H3K4me3 enriched regions by feature. The subset that were found within the first intron is also shown. **(B)** Distribution of H3K4me3 regions around the TSS of genes. **(C)** Genomic locations of H3K27ac enriched regions by feature; and **(D)** H3K27ac regions around gene TSS.

of promoters regions 500 base pairs upstream of the gene and a maximal enrichment at 200 base pairs downstream of the TSS with severe depletion of signal at the TSS (**Figure 3B**). Inspection of H3K4me3 regions that were within 2 kb of an annotated TSS, revealed 11 were associated with miRNAs in sheep and 295 were associated with tRNA. Overall, 3.6% of active enhancer and active promoter regions were associated with tRNA genes.

More than half of H3K27ac enhancer regions were annotated within introns (**Figure 3C**). Half of those regions were within

the first intron which indicated enrichment for active enhancers near the 5′ end of genes. Active enhancer signal had a similar bimodal distribution around the TSS of genes as promoters, but the maximal signal was located approximately 450 bp upstream of genes, with a second peak of signal approximately 450 bp downstream of the TSS (**Figure 3D**). However, on average, H3K27ac-exclusively enriched enhancer regions were more distal, 36 kb from the nearest TSS with 21% greater than 50 kb away from the nearest gene. Only 4% of the H3K27ac-exclusive regions were within 2 kb of a TSS. Silencers exclusively marked by H3K27me3 were further from genes than other ChIP-seq targets at an average of 53.5 kb from TSS with 34% greater than 50 kb away.

Promoters Predicted Actively Expressed Genes From RNA-Seq

Exactly half of all genes annotated in the Rambouillet reference (Refseq Annotation release 103) were associated with an H3K4me3 enriched regulatory element. Regulatory elements with H3K4me3 signal were identified with gene-rich regions of the genome, 77% were within 20 kb of the nearest gene. Regions with ChIP enrichment for active promoters and enhancers, H3K4me3, H3K27ac, and H3K4me1 were at constitutively expressed housekeeping genes including *POL3D*, *ACTB* (**Figure 4A**), and *GAPDH* and many had moderate signal enrichment. Active promoters were also found at macrophage tissue-specific genes like *PPARG* (**Supplementary Figure 10**) and at environment-specific genes like *ITGAX* that are highly expressed in RNA-seq (**Figure 4B**). Several tissue-specific highly expressed genes were associated with high ChIP-seq signal (**Figure 4B**). Lineage-specific genes that are not expressed in alveolar macrophages such as *GATA6* (**Figure 4C**) had enrichment of H3K4me3 at promoter regions but lacked distal enhancers (H3K4me1 and H3K27ac) and were not enriched for H3K27ac at promoters. Developmental genes which are not expressed in adult alveolar macrophages had broad regions of enrichment for H3K27me3 (**Figure 4D**).

Promoter regions enriched for H3K4me3 were then filtered for those within only 2 kb of annotated genes, as these were most likely to be a correct match between regulatory element and TSS. Approximately 73% of genes with active H3K4me3 enrichment within 2 kb of the TSS were also expressed in RNA-seq data from sheep alveolar macrophages in the Sheep Gene Expression Atlas. Analysis of regions with H3K4me3 and H3K27ac enrichment showed 78% were associated with gene expression from RNA-seq regardless of distance from that annotated gene. Maximum signal enrichment of H3K4me3 at promoter peaks within 2 kb of TSS was positively correlated with gene expression TPM when compared to RNA-seq data ($r = 0.28$, $P = 10^{-130}$). The top third H3K4me3 enriched regions corresponded to genes from the RNA-seq data with an average expression of 77 TPM (**Figure 5A**). Within the bottom third of H3K4me3-enriched regions the average expression of nearest genes was 40 TPM. Signal at promoter regions with enrichment for both H3K4me3 and H3K27ac did not have a quantitatively stronger correlation with gene expression than did H3K4me3 signal alone ($r = 0.12$).

Gene ontology analysis of genes associated with both H3K4me3 and H3K27ac enrichment that were expressed in alveolar macrophages had significant ($P < 0.05$), greater than two fold overrepresentation for biological processes like viral protein processing, positive regulation of antigen receptor-mediated signaling pathway, type I interferon-mediated signaling, regulation of autophagy, and mitotic spindle assembly checkpoint, among others. Antigen processing and presentation via MHC class II was overrepresented at 1.76-fold (additional GO annotation in **Supplementary Data 15**). Motif analysis of promoter sequences identified many known bindings sites for transcription factor proteins ELF4, ETS, and interferon-regulatory factors (IRF1-3 and IRF8) within H3K4me3 regions. Additional *de novo* motifs had highest similarity to binding sites for transcription factor proteins SFPI1, MYB family, and CEBP family.

Further analysis of promoters revealed a small subset of 3,641 regions with signal enrichment for both H3K4me3 and the repressive mark H3K27me3. Annotation with the nearest gene revealed 45% of these regions were within 2 kb of a TSS. These 1,630 regions were considered bivalent promoters. Regions were associated with 1,166 protein-coding genes. Gene ontology analysis was largely enriched for genes involved in molecular functions for transcriptional regulation, transcription factor activity, DNA and RNA binding, and RNA polymerase II regulation (**Figure 5B**). Biological processes discovered in GO overrepresentation analysis involved cell and tissue differentiation, stimulus response, and cell movement, among others (**Supplementary Data 15**). Motif analysis revealed that *de novo* motifs were more significant ($P = 10^{-58}$) than known motifs ($P = 10^{-30}$) within bivalent promoter sequences. A *de novo* motif with similarity to the binding motif for yeast protein STB1 was found in 38% of regions. Known motif analysis revealed bivalent promoter sequences were enriched for "CCCGC" and "CGCGCG" sequences and the motif for the *Drosophila* GAGA factor protein.

Enhancers Were Enriched for *de novo* Binding Motifs

Total enhancer regions with either H3K4me1, H3K27ac, or both were more numerous than H3K4me3 promoter regions. We found that genes have multiple enhancer regions. On average five significant regions enriched for enhancer signal were associated with each unique gene. In fact, we found that 69% of active enhancer regions marked with H3K27ac were found in clusters of two to seven (average of 3.2) significant regions around the same gene. Multiple regions meant that enhancers are further from the genes they control. H3K27ac regions were a mean of 1.9 kb further from their genes than H3K4me3 regions. Our data revealed that 56% of genes may be controlled by multiple active distal enhancers (e.g., H3K27ac regions further than 2 kb from genes) in macrophages.

DNA sequences from active enhancer regions marked by H3K27ac were scanned for motifs with HOMER. Enhancer

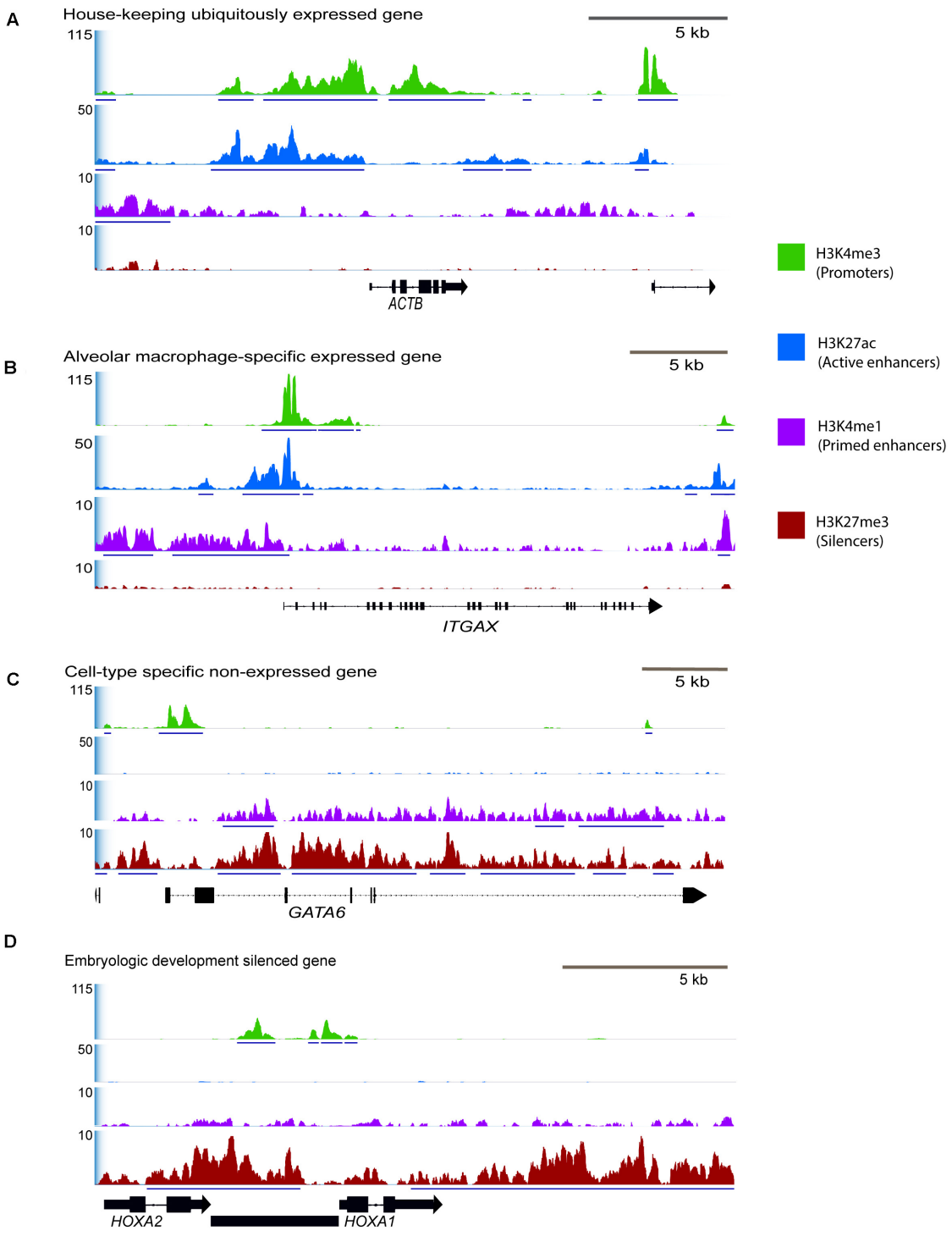


FIGURE 4 | Selected consensus regions of ChIP-seq signal at active and repressed genes in macrophages. Signal enrichment along the *Y*-axis is displayed as average reads per genomic content (RPGC) normalized for sequencing depth to 1× genome coverage. Input control signal was subtracted from the profiles to remove noise. The *X*-axis represents the chromosomal location with size bar given in kb. **(A)** Region displaying a housekeeping gene that we would be constitutively expressed in all cells and have active promoter and enhancer peaks, *ACTB* (actin-beta) from chromosome 24. **(B)** Tissue-specific gene actively expressed in alveolar macrophages, *ITGAX*, chromosome 24, and **(C)** tissue specific gene that is not expressed in alveolar macrophages, *GATA6*, chromosome 23. **(D)** Developmental genes that should be silenced with broad H3K27me3 signal, *HOXA1* and *HOXA2* on chromosome 4 that are not expressed in fully differentiated macrophages. In this region there is also an unannotated gene (predicted lncRNA) that has a bivalent promoter enriched by both H3K27me3 and H3K4me3. See **Supplementary Figure 10** for signal around additional immune-related genes.

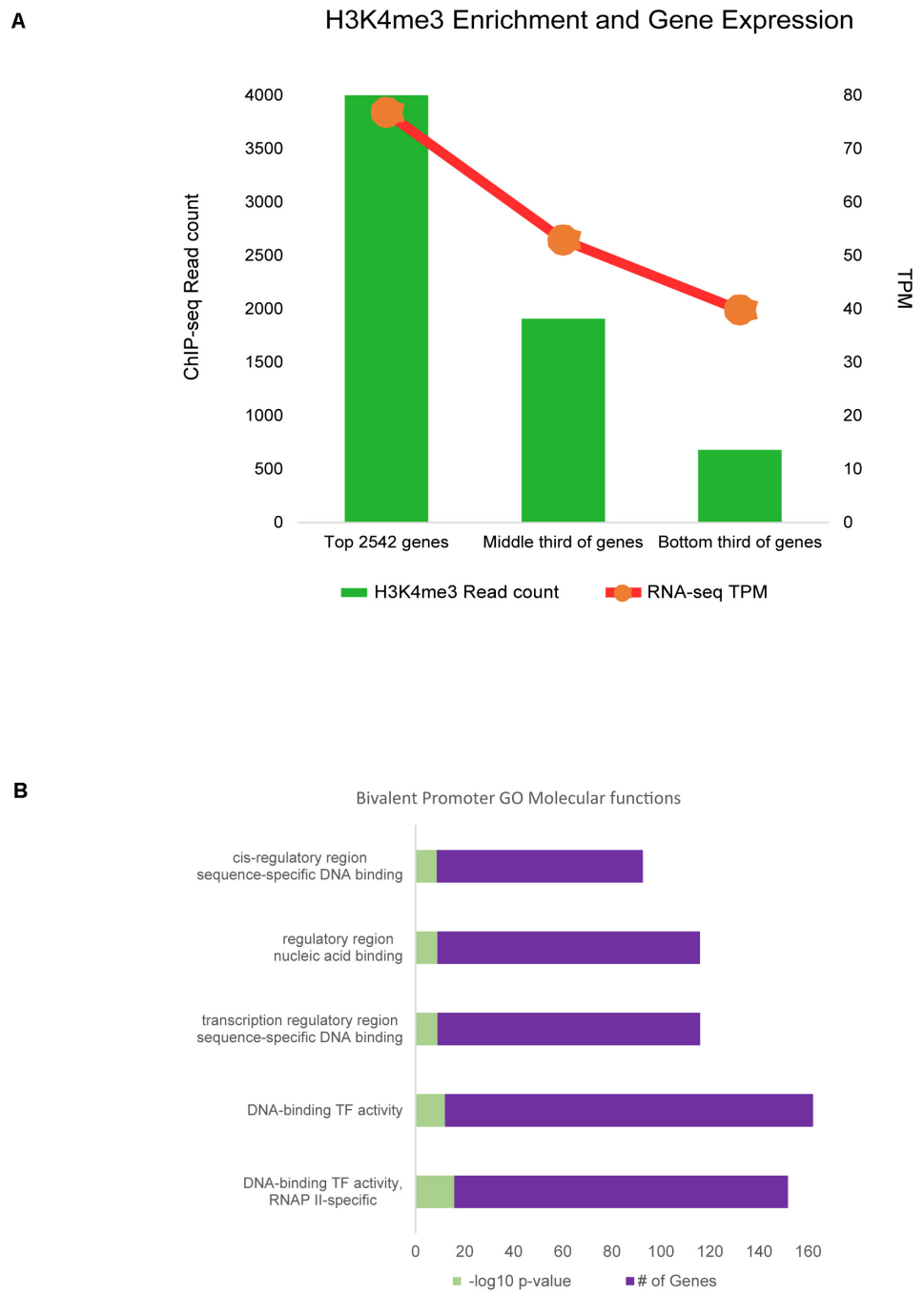


FIGURE 5 | Promoter regulatory elements and gene expression. **(A)** RNA-seq gene expression was associated with signal enrichment of active promoters marked by H3K4me3. RNA expression is shown as transcripts per million (TPM). **(B)** Bivalent promoter (H3K4me3 and H3K27me3) associated GO analysis annotations for categories of molecular functions.

sequences contained similar central bases but often shorter consensus motifs with degenerate bases at the flanking sequences (**Figure 6A**) to known human macrophage-specific binding sites. Within the top three most significant *de novo* motifs within H3K27ac regions was a 15-bp sequence with 0.89 match score similarity to the known binding motif for the protein peroxisome proliferator activated receptor gamma (PPARG). PPARG is a transcription factor specific to macrophages within the lung microenvironment. The *de novo* PPARG motif was discovered in approximately 10% of active enhancers. Primed enhancers denoted by H3K4me1 were generally enriched for known lineage specific and

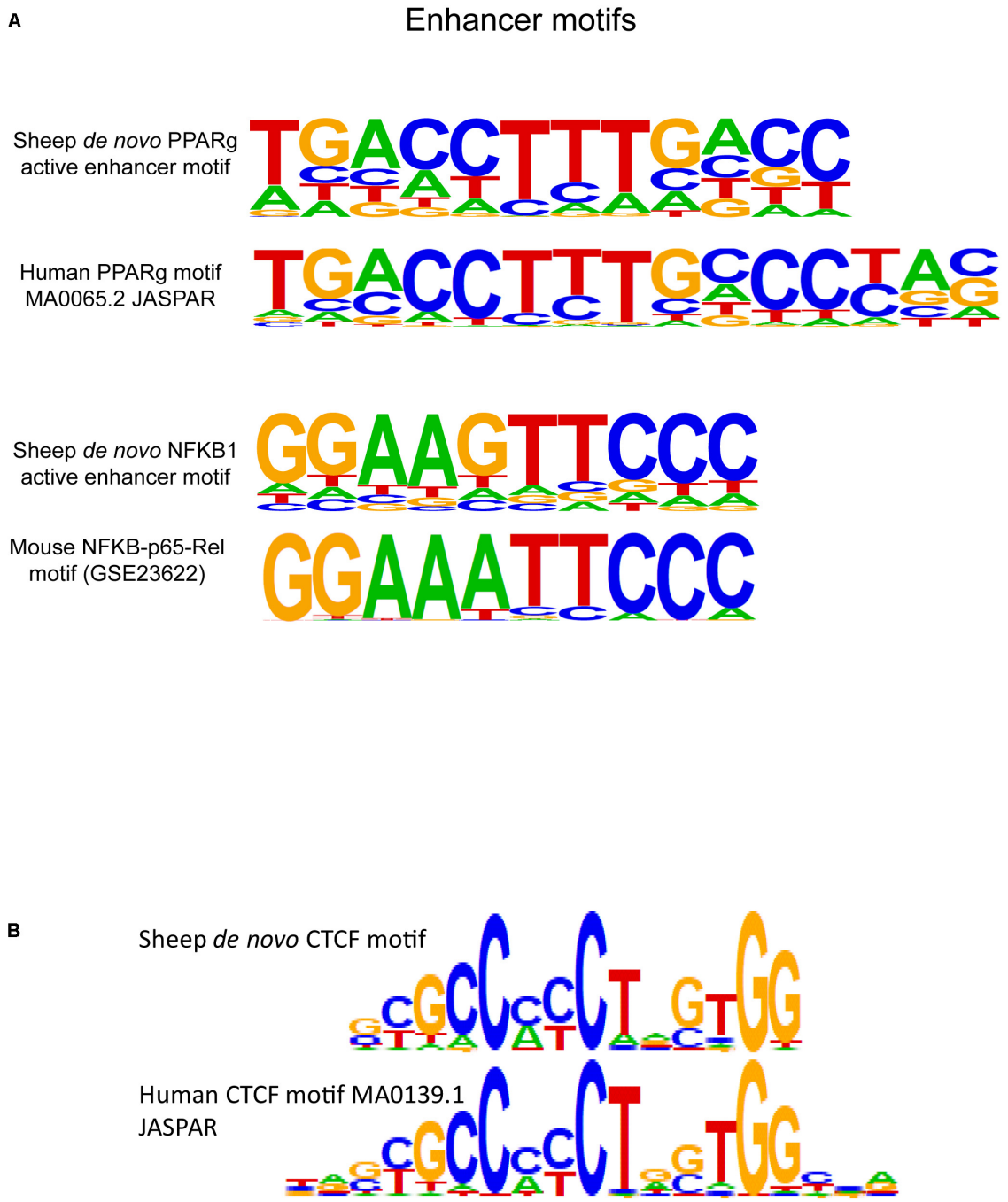


FIGURE 6 | Regulatory element motif enrichment in sheep alveolar macrophages with comparison to known motifs from model species. **(A)** *De novo* motifs from sheep macrophages were discovered for the transcription factor protein PPARG in active enhancers (H3K27ac-enriched regions), this protein is a known tissue-specific regulatory protein in alveolar macrophages. The third motif is a *de novo* motif that had a 0.88 match score to NFKB1 protein binding motifs identified in mice. **(B)** HOMER scanning of insulator sequences identified by CTCF immunoprecipitation revealed enrichment in 50.5% of regulatory regions by the *de novo* motif in sheep displayed here. The motif is similar but not identical to the known CTCF motif from humans. Additional CTCF motifs discovered in sheep are shown in **Supplementary Figure 12**.

pioneering factor motifs. Approximately half of H3K4me1 regions contained the known motif for the transcription factor protein PU.1. Primed enhancers in sheep additionally contained binding motifs for the transcription factor proteins CCAT enhancer binding protein beta ("C/EBP-beta," CEBPB) and CEBPC, SpiB, and SpiC, and a *de novo* motif for NFKB1 reflecting presence of immunity related binding sites (**Figure 6A**).

Silencers Were Associated With Lack of Gene Expression

Silencer regulatory elements denoted by H3K27me3 signal were generally not near genes on a genome-wide basis, since 69% of significant regions were in intergenic regions (see **Supplementary Figure 11** for genomic localizations). Less than a third of H3K27me3 regions were within introns, and 1–2% of regions were within exons or near the 5′ end of genes. Under 7% of the genes nearest to H3K27me3 regions were expressed in alveolar macrophage RNA-seq data. Silencer regions were located an average of greater than 29 kb from the nearest gene. Motif analysis identified binding sites for two significant motifs ($P < 10^{-51}$), to the proteins Zfp281 and NFKB1. Several additional C2H2 zinc finger protein binding sites were discovered with borderline significance ($P = 10^{-42}$) for transcription factor proteins ZKSCAN1, ZNF467, and ZNF165. Silencer regions were found near homeobox transcription factor genes (*HOX* family) involved in embryologic development. The closest genes to H3K27me3 regions in the sheep genome included 46% of all homeobox genes know in humans.

CTCF Insulators Motif Analysis and TAD Anchoring

Insulator element CTCF peaks were scanned with HOMER to determine the top *de novo* binding motifs in sheep and top known motifs (**Figure 6B**). CTCF and CTCF like motifs were the most significant motifs discovered amongst known motifs with ($P = 10^{-1482}$) and amongst closest matches to *de novo* motifs ($P = 10^{-4338}$) (**Supplementary Figure 12**). Significant motifs with match score greater than 0.75 were used to scan the DNA sequence of all CTCF regions and revealed 50.5% of consensus regions contained CTCF motifs. Next, we calculated the genomic distance between pairs of nearest CTCF enriched regions on each chromosome since these insulators form domains with paired anchors. We found that the genome could be divided into approximately 11,100 predicted TADs based on pair-wise counts per chromosome (see **Supplementary Data 15**). These regions were calculated to be an average of 258 kb in length. Based on gene content per chromosome insulators delimited an average of three genes per chromatin domain in sheep.

HyperChIPable Regions Found in Sheep Macrophages

In our comparative analyses between ChIP-seq targets we discovered enriched signal in 1,520 regions of the genome that were significant compared with the control in all immunoprecipitated datasets. The regions did not have specificity for any antibody used for pulldowns. These "hyperChIPable regions" spanned a small fraction (0.43%) of the genome. However, they were broad at 8 kb, longer than regions found in any individual target. Although the majority of hyperChIPable regions were found within intergenic regions (59%), there was moderate but significant enrichment at promoters ($P = 8.9 \times 10^{-8}$) indicating a pattern to their location rather than pure noise. Like H3K4me3 regions, hyperChIPable regions were found within the first intron of genes or at the TSS (39%) (see **Supplementary Figure 13**).

DISCUSSION

Overview

We identified repressive and active regulatory elements and validated regions by comparison to public RNA-seq data on alveolar macrophages from the Sheep Gene Expression Atlas (Clark et al., 2017). These data met the benchmarks for acceptable quality set forth by the ENCODE project consortium and the livestock FAANG consortium exceeding sequencing depth of 20 million usable fragments for narrow marks and 45 million usable fragments for broad marks with production of complex libraries (Dunham et al., 2012; Andersson et al., 2015). We identified the promoters and distal *cis*-acting regulatory elements for housekeeping genes, genes associated with macrophage differentiation, and tissue-specific alveolar macrophage genes. We found bivalent regulatory elements at the promoters of few genes and annotated GO processes that varied from the processes found in more typical promoters. We also identified significant regions bound by CTCF with N-ChIP that revealed insulators and allowed the first preliminary estimates of chromatin domains in sheep alveolar macrophages. Our collective data assigned a putative biological regulatory function in macrophages to nearly 12% of the sheep genome.

Alveolar macrophages served as a biologically interesting tissue given their myriad local functions and importance to zoonotic intracellular pathogens. We were able to identify GO overrepresentation in regulatory element associated genes involved in pathways reflective of tissue-resident macrophage main functions in homeostasis such as protein catabolism, autophagy, and nitrogen metabolism (Lavin et al., 2014). Both innate immune functions like interferon signaling, and adaptive immune functions like antigen processing and MHC class II presentation were identified as significant, reflecting macrophages unique role in both branches (Schmidt et al., 2016). We chose native ChIP-seq (N-ChIP) instead of formaldehyde cross-linked (X-ChIP) because it has been reported to preserve antibody epitopes leading to increased enrichment of signal and less back ground noise (O'Neill and Turner, 2003; Wagschal et al., 2007; Villar et al., 2015; David et al., 2017; Fang et al., 2019). Cells were frozen and stored short term prior to processing for N-ChIP since this method has been demonstrated to maintain sensitivity and reproducibility (Brind'Amour et al., 2015). While a possible limitation is that some protein-chromatin interactions may be lost in freezing. Optimization of shearing was effective with highly reproducible micrococcal nuclease digestion in our hands, that reliably yielded mononucleosomes. Native ChIP has compounded advantages in that no large protein-chromatin complexes are created that have been shown to inhibit shearing, and the endo- and exo-nuclease activity of micrococcal nuclease allows excellent resolution of the ChIP-seq target regions (Skene and Henikoff, 2015).

Individual Chromatin Modifications in Sheep Macrophages Reflect Expectations

We found evidence in ChIP-seq enriched active regions (H3K4me3 and H3K27ac) of increased guanosine and cytosine content, as others have shown in various cell types that promotes open chromatin (Glass and Natoli, 2016). Heterochromatin and silencer sequences were shown to be depleted for GC content (Glass and Natoli, 2016) like our H3K27me3-enriched regions (see **Table 2** for GC content summary). We found a clear bimodal distribution of active promoters and enhancers around the TSS (**Figure 3**) as has been reported by others for a variety of tissue types (Kingsley et al., 2020). At the TSS there was a slight depression in signal reflecting this nucleosome depletion that would allow for the positioning of the initiation complex and RNA polymerase along the chromatin.

Indicative of promoter regions, H3K4me3 enrichment should be detected at one-half to two-thirds of all genes in a cell including 60% of silenced genes (Barski et al., 2007). We were able to find reproducible peaks in both animal replicates that account for 50% of the annotated genes or pseudogenes in the sheep genome yielding putative active and primed (poised) promoters used in sheep alveolar macrophages. Promoters primed by H3K4me3 have also been demonstrated in macrophages notably at *TLR4* promoters and immediate early genes that help to induce rapid expression after exposure to foreign and injurious stimuli like LPS (Escoubet-Lozach et al., 2011). We produced an annotation resource for active regulatory elements at key immune genes including *TLR4, TLR8, TLR6*, MHC class II genes, *BHLHE40,* and *BHLHE41* that are highly expressed in alveolar macrophages (**Supplementary Figure 10**). The transcription factors BHLHE40 and BHLHE41 are master regulators that repress expression of lineage-inappropriate genes in alveolar macrophages and govern self-renewal. In fact, BHLHE40 inhibits H3K27ac in regulatory elements to control gene expression (Rauschmeier et al., 2019). *BHLHE41* was not annotated in the previous reference annotation for Oar_v3.1. Therefore its gene expression profile is absent from the original analysis in the Sheep Gene Expression Atlas, but it was associated with very high signal enrichment for ChIP-seq active marks so we would predict it is also highly expressed in sheep alveolar macrophages. Our analysis revealed regulatory elements for core tissue-specific genes like the transcription factor *PPARG* (**Supplementary Figure 10**) which regulates homeostasis and surfactant catabolism (Lavin et al., 2014) and also detected enrichment for its protein binding motif in active regulatory regions.

Nearly two-thirds of H3K4me3 regions in sheep macrophages were found further than 2 kb from annotated genes or within the first intron or first exon. This perhaps suggests that alternative start sites exist in macrophages or that gene and transcript annotation is incomplete for the sheep genome. Previous work showed that immunity related genes are enriched for tissue specific allelic expression (Salavati et al., 2019) so alveolar macrophages may express unique isoforms. The FR-AgENCODE project found similar results in immune cells of goats, where 37% of coding transcripts were determined to be alternative and multi-exonic (alternative splicing) compared to the reference annotations and many extensions of annotated genes were also discovered (Foissac et al., 2019). In fact, the ENCODE Project consortium originally detected a similar trend in humans and reported that the sole use of Refseq based annotations led to dramatically overestimated distance of regulatory elements from expected promoter locations at TSS (Birney et al., 2007). The annotation build for Oar_rambouillet_v1.0 included few immune tissues for gene prediction. Generation of experimental data to improve TSS annotation is one of the objectives defined by the Ovine FAANG Project and cap analysis gene expression (CAGE) data on 55 tissues and alveolar macrophages was recently published (Salavati et al., 2020). CAGE is an excellent method to confirm function and location of promoters and enhancers (Andersson et al., 2014) for validation of ChIP-seq (Wood et al., 2020).

Total enhancer regions were more numerous than active promoter regions as each gene can be controlled by multiple enhancers but generally a single promoter. Differentiation between active and primed distal regulatory elements was possible in our data as H3K27ac (Creyghton et al., 2010) had a clear association with sheep macrophage promoters and predicted gene expression from the Sheep Gene Expression Atlas. The mark H3K4me1 functions to prime enhancer regions disallowing recruitment of histone deacetylases and was less predictive of gene expression in sheep. Rather, we found these primed enhancer regions were highly enriched for canonical PU.1 binding motifs in sheep. So called pioneer factors, PU.1 can bind partially compact chromatin and help open chromatin for additional transcription factors (Bernstein et al., 2002). PU.1 is also a lineage-determining transcription factor highly active in macrophages (Glass and Natoli, 2016; Soares et al., 2017). We found PU.1 motifs in 50% of enhancers, Lavin et al. (2014) reported motifs in 30–40% of murine macrophage enhancers with X-ChIP. Sheep enhancers contained binding motifs for enhancer binding proteins CEBPB and CEBPD that are known to regulate genes involved in immunity including cytokines, chemokines, and proinflammatory factors (Wang et al., 2019). CEBP proteins also mediate acetylation of H3K27 through coactivators which prevent methylation at this residue, priming the region for further activation. Next, we identified insulators that can modulate enhancer function.

Our data was able to identify greater than 50,000 CTCF enriched regions in the genome of each individual animal and approximately 22,000 common to both animals. This matches the estimate of 40,000–50,000 CTCF occupied sites obtained in individual cell types from the wealth of ENCODE data (Ghirlando and Felsenfeld, 2016). This was an interesting experiment as relatively few studies use native chromatin for CTCF immunoprecipitation, and we may not have captured transient CTCF regions. In ChIA-PET studies, many insulators are transiently bound by CTCF with low correlation of occupancy and other regions form more permanent contacts, which may explain the lower percentage of overlapping sites we saw between the two sheep compared to other marks (Handoko et al., 2011; Guo et al., 2015). Native ChIP-seq is reported to be successful for CTCF since its binding affinity to chromatin is far greater

than other transcription factors (Nakahashi et al., 2013). In fact, our dataset may be enriched for predominantly "non-exchangeable" CTCF sites that have the highest binding affinity and generally denote the largest structural chromosome loops. In the future, comparison with X-ChIP and Hi-C from sheep may be helpful to elucidate localized transient chromatin loops. We were able to estimate the average size of chromatin loops from our CTCF sites by empirically assuming pairs will form contact domains. This yielded a mean estimated domain size of 258 kb comparable to human contact domains determined from Hi-C data of median 185 kb (Rao et al., 2014). Literature surveys report 1,000–1,000,000 loops per genome (Fullwood et al., 2009; Jin et al., 2013; Sanborn et al., 2015). Our data yielded an estimated 11,000 pairs (22,000 regions) that could form loops in the sheep genome. Hi-C assays in goats yielded 8,990 TADs in goats with a similar size of 220 kb (Foissac et al., 2019).

We defined the first *de novo* CTCF motif in sheep macrophages. The central core of the 19-bp canonical motif is maintained between human and sheep, however, the flanking nucleotides on either end of the motif displayed heterogeneity compared to the core human motif in the JASPAR database (MA0139.1). The *de novo* motifs were found in approximately half of the sheep insulator regions, meaning half do not contain recognizable motifs or may have bindings sites adjacent to immunoprecipitated regions since CTCF binds chromatin in large protein complexes. However, from Rao et al. (2014), only 54% of CTCF-bound regions contain CTCF motifs, paralleling the 50% motif content we found in sheep. This can create difficulty in calling pairs of CTCF that form the anchors for TADs and localized sub-TADs.

Combinatorial Patterns of Mark Overlap in Regulatory Elements

We found complex patterns of overlapping histone modifications across the regulatory element landscape. The "histone code" precisely titrates gene expression at multiple levels and is better assayed by analysis of multiple ChIP-seq targets together as we saw in sheep (**Figure 2**). This "cross-talk" reinforces the chromatin state by either supporting activation or attenuation of gene expression and may provide mechanisms for redundancy and epigenetic memory (Fischle et al., 2003; Wang et al., 2008). Epigenetic memory serves a key role in macrophages as it is the proposed mechanism behind trained immunity, that can reversibly recalibrate responses to pathogens, non-specifically. For all our ChIP-seq targets, regulatory elements contained at least some degree of overlap. Patterns of overlap in histone modifications compartmentalize the genome into euchromatin and heterochromatin, i.e., active versus repressed transcription. Subsequently, we saw relatively little overlap of H3K27me3 with the other marks tested as this is the only distinctly repressive mark we examined. In these sheep, the subset of active elements with both H3K4me3 and H3K27ac were better predictors of highly expressed genes that either mark alone, found at well-annotated genes as they were more frequently at the TSS. Conversely, we determined H3K27ac-exclusive regions were consistent with distal CREs (true enhancers) as reported in many species (Villar et al., 2015).

Some promoter regions in sheep macrophages were found to paradoxically have both H3K4me3 active marks and H3K27me3 repressive marks. These bivalent promoters signify unique genes that have highly variable and responsive gene expression (Vastenhouw and Schier, 2012). GO revealed different functions from those with active promoters with the caveat that both gene lists may contain noise from the RNA-seq data being from different animals than the ChIP-seq data. Alveolar macrophages are known to maintain tissue homeostasis when quiescent but once activated in response to invading pathogens or tissue injury can begin cytokinesis and phagocytosis (Lavin et al., 2014; Glass and Natoli, 2016; Schmidt et al., 2016). Bivalent promoters play essential roles in myeloid differentiation and when macrophage progenitors lose H3K27me3 repression at certain bivalent sites it can contribute to development of cancers like acute myeloid leukemia (Thalheim et al., 2017). Motif analysis revealed very few know transcription factor binding sites and several motif sequences of low complexity and high GC enrichment comparable to bivalent promoters of mammalian embryonic stem cells (Mantsoki et al., 2015).

In contrast, overlap of H3K27me3 and H3K27ac histone modifications are antagonistic to one another and not found in the same regions (Tie et al., 2009). Accordingly, we did not find enrichment of these two marks together in the same regions. Overall, silencer elements, H3K27me3, were found in broadly different locations than active elements captured by H3K27ac, H3K4me3 and H3K4me1. We were able to find enrichment of 6% of the sheep genome in alveolar macrophages with the silencer mark H3K27me3. This likely represents the bulk of this compartment in the sheep genome since H3K27me3 corresponds to regions of heterochromatin estimated to comprise 8% of the human genome as 92% is euchromatin (Consortium International Human Genome Sequencing, 2004; Rao et al., 2014). In our data, H3K4me1 and CTCF showed some overlap with one another and with H3K27me3 regions near boundary zones between heterochromatin and euchromatin. We found in sheep macrophages as Barski et al. (2007) found in human T-cells, that locations with CTCF enrichment also were enriched for multiple histone methylation marks found at domain boundaries.

Lastly, because we produced data for multiple marks, we were able to elucidate putative hyperChIPable regions in the sheep genome. These regions were found in all immunoprecipitated datasets and were not specific for any one target or antibody (**Figure 3B**). HyperChIPable regions were slightly more likely to be found at promoter regions and within the first intron of genes near TSSs (**Supplementary Figure 13**). This active promoter effect has been reported in the past for biologically hyperChIPable regions in human and mouse (Wreczycka et al., 2019). There may be a biological reason that these regions appeared in all immunoprecipitated fractions or are perhaps more efficiently sequenced. Enrichment of these non-specific sites near promoters may also be an artifact of the experimental protocol as micrococcal nuclease digestion is more efficient at euchromatin than heterochromatin, so a larger portion of fragment ends available for sequencing will naturally occur

around open chromatin. HyperChIPable regions may also be caused by artifacts in the reference assembly. Regions containing repetitive elements are troublesome for genome assembly and may be collapsed, therefore natural copy number variation would create the appearance of falsely elevated signal in the region (Amemiya et al., 2019). We have provided these putative hyperChIPable regions for sheep in the public OSF repository (see section "Data Availability Statement"). As these regions were not known previously in sheep and not yet validated, we have not removed them from our ChIP-seq datasets. However, once validated in additional sheep tissues these regions can be included on a "block list" of sites to be removed from future experiments, like the ENCODE consortium created for model organisms, since they do not represent signal from the protein target of interest (Carroll et al., 2014).

Regulatory Element Locations and Signal Enrichment Associate With Gene Expression

Generally, gene expression can be quantitatively predicted by the signal enrichment of histone modifications. However, specific gene expression is highly contingent on cell type and the usage of specific regulatory elements is cell type dependent, especially in immune genes (Lavin et al., 2014). Thus, it was critical for us to experimentally determine histone modifications in primary macrophages most representative of *in vivo* conditions rather than from cell culture conditions to identify the regulatory elements that are uniquely used by the immune system. We found modest, positive correlation that was highly statistically significant ($P = 10^{-130}$), between signal enrichment of H3K4me3 in promoters and gene expression in alveolar macrophages determined in the Sheep Gene Expression Atlas (Clark et al., 2017). This correlation served as a "proof-of-concept" validation of our ChIP-seq regions. Importantly, we found that overlap of both H3K27ac and H3K4me3 had a stronger predictive value for gene expression than H3K4me3 alone as nearly 80% of genes with enrichment for both were expressed from RNA-seq. In sum these active regulatory elements were at the TSS of approximately 7,600 protein coding genes that were actively expressed in alveolar macrophages. Quantitative correlation between our ChIP-seq signal and RNA-seq expression was limited since the data was obtained from different animals, of different breeds, raised on different continents, and the RNA-seq data were obtain from two individual female animals in separate experiments. More complex regression analysis could improve correlation between ChIP-seq signal and RNA-seq data (Angelini and Costa, 2014), however, we opted for a simple analysis as proof-of-concept for this data resource. We would expect ChIP-seq signal to have improved quantitative correlation with RNA-seq data if generated from the same animals at the same time points. We envision the ChIP-seq data presented here being used as foundational annotation of CREs in quiescent macrophages from healthy sheep and these data will allow identification of target regions for further study. Future work may expand upon the multiple functions of macrophages by examining activated or infected macrophages and yield both epigenomic data and transcriptional data from the same macrophage populations. These types of studies have potential to capture epigenetic modifications caused by response to exogenous agents or orchestrated by infectious agents at regions identified in resting cells and at additional genomic regions (Hall et al., 2020; Herrera-Uribe et al., 2020).

We also captured transcriptional activators for a variety of types of RNA that could not be correlated to gene expression from mRNA-seq. For example, we identified the promoter for several members of the *let-7* microRNA precursor family. In human and murine macrophages, *let-7* has been shown to post-transcriptionally control cytokine production in innate immune responses by repressing production of interleukin-10 (IL-10), IL-6, and TLR4 until pathogens are detected (Schulte et al., 2011). Annotation of short RNA elements, which is largely missing from the sheep genome annotation, could be defined by combining RNA-seq methods with more stable DNA based methods like ChIP-seq to find short regions of active transcription. Our data indicated several regulatory elements that displayed the pattern of bona fide active promoters but are not near any currently annotated genes or regulatory RNA; we hypothesize these regions may control expression of either novel tissue-specific, short regulatory, or weakly expressed transcripts which are difficult to annotate. Deep sequencing RNA experiments in sheep and goats have indeed found lncRNA had shorter transcripts and weaker expression which explains difficulty in annotation of these types of functional elements (Clark et al., 2017).

Conclusion

In summary, we generated ChIP-seq data for four core histone modifications and chromatin domain defining CTCF locations for the first time in sheep primary alveolar macrophages. We have shown that active enhancer and promoter signal enrichment was predictive of gene expression in sheep macrophages. We also provided annotations of novel hyperChIPable regions that may represent biological or non-specific experimental artifacts and potentially should be included on a "block list" to be removed from future ChIP-seq experiments in sheep. The data generated here are publicly available for researchers and will be valuable for comparative and ovine immunology studies as well as fine mapping to improve marker assisted selection for infectious disease resilience. ChIP-seq defined promoters may help to annotate TSSs of genes, especially those that are not well or widely expressed. We also put forth novel binding motifs found within regulatory elements in sheep macrophages. Understanding the epigenetic control and response mechanism of the immune system is very important not only for animal health and infectious agent eradication but also for numerous economically important production traits. The immune response in sheep has energy resource costs despite the health outcome, and this ultimately affects efficiency of meat and milk production for human consumption. Therefore, genetic, and epigenetic improvement of infectious disease resistance or tolerance is important to increasing production efficiency in sheep. Use of regulatory element annotation data to develop marker-assisted or genomic selection tools has advantages over traditional methods to control infectious diseases as it promotes selection of hardier

animals prior to the introduction of pathogens and avoids antibiotic resistance altogether.

These data, as part of FAANG, can be readily incorporated into the reference genome annotation or viewed as custom tracks. Generation of these data on a macrophage immune cell type will allow future work on mutations and epigenetic variations that cause differences in sheep immune response, zoonoses transmission, and immunological effects on production efficiency.

DATA AVAILABILITY STATEMENT

The original contributions generated for this study are publicly available. The animal metadata are deposited to the EMBL-EBI BioSamples database (http://www.ebi.ac.uk/biosamples) under accession numbers SAMEA7423844 and SAMEA7423845 and samples under SAMEA7423964 and SAMEA7423965. Raw DNA sequencing data in .fastq.gz format is available on the FAANG data portal and within European Nucleotide Archive (ENA) project accession PRJEB40528 (https://www.ebi.ac.uk/ena/browser/view/PRJEB40528). Analysis BED files of the peak calls in each animal, the consensus regions for each mark, bivalent promoters regions and hyperChIPable regions are available on the Open Science Framework website upon request and will be made public, https://osf.io/8rq4c/?view_only=1937b862ea874a14b85fff2043a9b522. Publicly available datasets from the Sheep Gene Expression Atlas (Roslin Sheep Atlas) were analyzed in this study. This data can be found here: University of Edinburgh DataShare portal at http://dx.doi.org/10.7488/ds/2112 (Clark et al., 2017). The sheep reference genome Oar_rambouillet_v1.0 is available at NCBI under accession GCA_002742125.1 (https://www.ncbi.nlm.nih.gov/assembly/GCF_002742125.1/) (Worley, 2017).

ETHICS STATEMENT

The animal study was reviewed and approved by Institutional Animal Care and Use Committee at Washington State University.

AUTHOR CONTRIBUTIONS

Experimental protocols were designed by AM, MM, MH, BM, and SW. Cell collection was completed by AM, MM, and MH. Experiments, data generation, and initial data analysis were completed by AM. The bioinformatics pipeline was designed by AM and DH. Final analysis was completed by AM, MM, MH, DH, BM, and SW. AM, MM, BM, and SW acquired funding. AM wrote the first draft of the manuscript and created visualizations. All authors contributed to editing the manuscript and visualizations. All authors read and approved the final manuscript.

FUNDING

This work was supported by USDA-ARS 2090-32000-036-00D and by USDA-NIFA Grant 2017-67016-26301. Thank you to the Seattle Chapter of ARCS and Connie Niva for stipend support and the American Sheep Industry Heritage grant and the NSF Research Coordination Network grant provided by Carl Schmidt (University of Delaware) for training/travel support of ATM. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

Thank you to Emily Clark and colleagues at the Roslin Institute for publication and provision of data for the Sheep Gene Expression Atlas. Thank you to Codie Durfee for excellent laboratory support and bioinformatics assistance. Thank you to Kaneesha Hemmerling for assistance in tissue collection. Thanks to Caylee Birge for laboratory support, Emma Karel and WSU Animal care staff for dedicated animal care, WSU Spokane Genomics Core especially Ben Liu and Jing Wang for library preparation, sequencing, and assistance with troubleshooting. Thank you to Derek Pouchnik and the WSU Laboratory for Biotechnology and Bioanalysis for fragment analysis, sonicator access, and troubleshooting. Thank you to Mazdak Salavati and Emily Clark for discussion on sheep genome annotation. Thank you to Kim Davenport and Mingrui Duan for initial discussions on bioinformatic analysis and ChIP. Thank you to FAANG members for discussions and providing public protocols and data. Thank you to the FAANG Data Coordination Centre for assistance with data submission and for archiving sample data and metadata.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.612031/full#supplementary-material

REFERENCES

Albert, F. W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16:197. doi: 10.1038/nrg3891

Amemiya, H. M., Kundaje, A., and Boyle, A. P. (2019). The ENCODE blacklist: identification of problematic regions of the Genome. *Sci. Rep.* 9:9354. doi: 10.1038/s41598-019-45839-z

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol* 16:57. doi: 10.1186/s13059-015-0622-4

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

Andrews, S., and Babraham, B. (2016). *FastQC: a quality control tool for high throughput sequence data.* (v0.11.3). Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed January 11, 2016).

Angelini, C., and Costa, V. (2014). Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. *Front. Cell and Dev. Biol.* 2:51. doi: 10.3389/fcell.2014.00051

Aranda, S., Mas, G., and Di Croce, L. (2015). Regulation of gene transcription by Polycomb proteins. *Sci. Adv.* 1:e1500737. doi: 10.1126/sciadv.1500737

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27(2 Pt. 1), 299–308. doi: 10.1016/0092-8674(81)90413-x

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837. doi: 10.1016/j.cell.2007.05.009

Bernstein, B. E., Humphrey, E. L., Erlich, R. L., Schneider, R., Bouman, P., Liu, J. S., et al. (2002). Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc. Natl. Acad. Sci. U.S.A.* 99, 8695–8700. doi: 10.1073/pnas.082249499

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816. doi: 10.1038/nature05874

Blacklaws, B. A. (2012). Small ruminant lentiviruses: immunopathogenesis of visna-maedi and caprine arthritis and encephalitis virus. *Comp. Immunol. Microbiol. Infect. Dis.* 35, 259–269. doi: 10.1016/j.cimid.2011.12.003

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. *Nat. Rev. Genet.* 17:661. doi: 10.1038/nrg.2016.112

Brind'Amour, J., Liu, S., Hudson, M., Chen, C., Karimi, M. M., and Lorincz, M. C. (2015). An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. *Nat. Commun.* 6:6033. doi: 10.1038/ncomms7033

Carroll, T. S., Liang, Z., Salama, R., Stark, R., and de Santiago, I. (2014). Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front. Genet.* 5:75. doi: 10.3389/fgene.2014.00075

Clark, E. L., Bush, S. J., McCulloch, M. E. B., Farquhar, I. L., Young, R., Lefevre, L., et al. (2017). A high resolution atlas of gene expression in the domestic sheep (Ovis aries). *PLoS Genet.* 13:e1006997. doi: 10.1371/journal.pgen.1006997

Consortium International Human Genome Sequencing (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945. doi: 10.1038/nature03001

Cordier, G., Cozon, G., Greenland, T., Rocher, F., Guiguen, F., Guerret, S., et al. (1990). In vivo activation of alveolar macrophages in ovine lentivirus infection. *Clin. Immunol. Immunopathol.* 55, 355–367. doi: 10.1016/0090-1229(90)90124-9

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* 107, 21931–21936. doi: 10.1073/pnas.1016071107

David, S.-A., Piégu, B., Hennequet-Antier, C., Pannetier, M., Aguirre-Lavin, T., Crochet, S., et al. (2017). An assessment of fixed and native chromatin preparation methods to study histone post-translational modifications at a whole genome scale in skeletal muscle tissue. *Biol. Proc. Online* 19:10. doi: 10.1186/s12575-017-0059-0

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

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C., Doyle, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74. doi: 10.1038/nature11247

Escoubet-Lozach, L., Benner, C., Kaikkonen, M. U., Lozach, J., Heinz, S., Spann, N. J., et al. (2011). Mechanisms establishing TLR4-responsive activation states of inflammatory response genes. *PLoS Genet.* 7:e1002401. doi: 10.1371/journal.pgen.1002401

Fang, L., Liu, S., Liu, M., Kang, X., Lin, S., Li, B., et al. (2019). Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biol.* 17:68. doi: 10.1186/s12915-019-0687-8

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nat. Prot.* 7, 1728–1740. doi: 10.1038/nprot.2012.101

Fischle, W., Wang, Y., and Allis, C. D. (2003). Histone and chromatin cross-talk. *Curr. Opin. Cell Biol.* 15, 172–183. doi: 10.1016/s0955-0674(03)00013-9

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., et al. (2019). Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol.* 17:108. doi: 10.1186/s12915-019-0726-5

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Bin Mohamed, Y., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462, 58–64. doi: 10.1038/nature08497

Gendelman, H. E., Narayan, O., Kennedy-Stoskopf, S., Clements, J. E., and Pezeshkpour, G. H. (1984). Slow virus-macrophage interactions. *Characterization of a transformed cell line of sheep alveolar macrophages that express a marker for susceptibility to ovine-caprine lentivirus infections*. *Lab. Invest.* 51, 547–555.

Gendelman, H. E., Narayan, O., Kennedy-Stoskopf, S., Kennedy, P. G., Ghotbi, Z., Clements, J. E., et al. (1986). Tropism of sheep lentiviruses for monocytes: susceptibility to infection and virus gene expression increase during maturation of monocytes to macrophages. *J. Virol.* 58, 67–74. doi: 10.1128/jvi.58.1.67-74.1986

Ghirlando, R., and Felsenfeld, G. (2016). CTCF: making the right connections. *Genes Dev.* 30, 881–891. doi: 10.1101/gad.277863.116

Giuffra, E., and Tuggle, C. K. (2019). Functional annotation of animal genomes (FAANG): current achievements and Roadmap. *Annu. Rev. Anim. Biosci.* 7, 65–88. doi: 10.1146/annurev-animal-020518-114913

Glass, C. K., and Natoli, G. (2016). Molecular control of activation and priming in macrophages. *Nat. Immunol.* 17, 26–33. doi: 10.1038/ni.3306

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., et al. (2015). CRISPR Inversion of ctcf sites alters genome topology and enhancer/promoter function. *Cell* 162, 900–910. doi: 10.1016/j.cell.2015.07.038

Hall, T. J., Vernimmen, D., Browne, J. A., Mullen, M. P., Gordon, S. V., MacHugh, D. E., et al. (2020). Alveolar Macrophage Chromatin Is Modified to Orchestrate Host Response to Mycobacterium bovis Infection. *Front. Genet.* 10:1386. doi: 10.3389/fgene.2019.01386

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, 630–638. doi: 10.1038/ng.857

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39:311. doi: 10.1038/ng1966

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Herrera-Uribe, J., Liu, H., Byrne, K. A., Bond, Z. F., Loving, C. L., and Tuggle, C. K. (2020). Changes in H3K27ac at Gene Regulatory Regions in Porcine Alveolar Macrophages Following LPS or PolyIC Exposure. *Front. Genet.* 11:817. doi: 10.3389/fgene.2020.00817

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. U.S.A.* 106, 9362–9367. doi: 10.1073/pnas.0903103106

Ibeagha-Awemu, E. M., and Zhao, X. (2015). Epigenetic marks: regulators of livestock phenotypes and conceivable sources of missing variation in livestock improvement programs. *Front. Genet.* 6:302. doi: 10.3389/fgene.2015.00302

Jambhekar, A., Dhall, A., and Shi, Y. (2019). Roles and regulation of histone methylation in animal development. *Nat. Rev. Mol. Cell Biol.* 20, 625–641. doi: 10.1038/s41580-019-0151-1

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J. F., Faraut, T., et al. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* 344, 1168–1173. doi: 10.1126/science.1252806

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., et al. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503, 290–294. doi: 10.1038/nature12644

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in Vivo protein-DNA Interactions. *Science* 316, 1497–1502. doi: 10.1126/science.1141319

Kingsley, B. N., Kern, C., Creppe, C., Hales, N. E., Zhou, H., Kalbfleisch, S. T., et al. (2020). Functionally annotating regulatory elements in the equine genome using histone mark ChIP-Seq. *Genes* 11:3. doi: 10.3390/genes11010003

Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., et al. (2014). Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell* 159, 1312–1326. doi: 10.1016/j.cell.2014.11.018

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Mantsoki, A., Devailly, G., and Joshi, A. (2015). CpG island erosion, polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells. *Sci. Rep.* 5:16791. doi: 10.1038/srep16791

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D., et al. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466, 253–257. doi: 10.1038/nature09165

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

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucl. Acids Res.* 47, D419–D426. doi: 10.1093/nar/gky1038

Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., 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

Nakahashi, H., Kieffer Kwon, K.-R., Resch, W., Vian, L., Dose, M., Stavreva, D., et al. (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.* 3, 1678–1689. doi: 10.1016/j.celrep.2013.04.024

Naval-Sanchez, M., Nguyen, Q., McWilliam, S., Porto-Neto, L. R., Tellam, R., Vuocolo, T., et al. (2018). Sheep genome functional annotation reveals proximal regulatory elements contributed to the evolution of modern breeds. *Nat. Commun.* 9:859. doi: 10.1038/s41467-017-02809-1

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., et al. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics* 28, 1919–1920. doi: 10.1093/bioinformatics/bts277

Nguyen, Q. H., Tellam, R. L., Naval-Sanchez, M., Porto-Neto, L. R., Barendse, W., Reverter, A., et al. (2018). Mammalian genomic regulatory regions predicted by utilizing human genomics, transcriptomics, and epigenetics data. *GigaScience* 7, 1-17. doi: 10.1093/gigascience/gix136

Niang, M. (1992). *Interaction of Mycoplasma ovipneumoniae With Sheep Alveolar Macrophages*. Ames: Iowa State University.

Niang, M., Rosenbusch, R. F., Lopez-Virella, J., and Kaeberle, M. L. (1997). Expression of functions by normal sheep alveolar macrophages and their alteration by interaction with Mycoplasma ovipneumoniae. *Vet. Microbiol.* 58, 31–43. doi: 10.1016/s0378-1135(97)00141-7

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

O'Geen, H., Echipare, L., and Farnham, P. J. (2011). Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. *Methods Mol. Biol.* 791, 265–286. doi: 10.1007/978-1-61779-316-5_20

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucl. Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

O'Neill, L. P., and Turner, B. M. (2003). Immunoprecipitation of native chromatin: NChIP. *Methods* 31, 76–82. doi: 10.1016/S1046-2023(03)00090-2

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., et al. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucl. Acids Res.* 44, W160–W165. doi: 10.1093/nar/gkw257

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

Rauschmeier, R., Gustafsson, C., Reinhardt, A., A-Gonzalez, N., Tortola, L., Cansever, D., et al. (2019). Bhlhe40 and Bhlhe41 transcription factors regulate alveolar macrophage self-renewal and identity. *EMBO J.* 38:e101233. doi: 10.15252/embj.2018101233

Rauw, W. (2012). Immune response from a resource allocation perspective. *Front. Genet.* 3:267. doi: 10.3389/fgene.2012.00267

Saeed, S., Quintin, J., Kerstens, H. H. D., Rao, N. A., Aghajanirefah, A., Matarese, F., et al. (2014). Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science* 345:1251086. doi: 10.1126/science.1251086

Salavati, M., Bush, S. J., Palma-Vera, S., McCulloch, M. E. B., Hume, D. A., and Clark, E. L. (2019). Elimination of reference mapping bias reveals robust immune related allele-specific expression in crossbred sheep. *Front. Genet.* 10:863. doi: 10.3389/fgene.2019.00863

Salavati, M., Caulton, A., Clark, R., Gazova, I., Smith, T. P. L., Worley, K. C., et al. (2020). Global analysis of transcription start sites in the new ovine reference genome (Oar rambouillet v1.0*). Front. Genet.* 11:580580. doi: 10.3389/fgene.2020.580580

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6456L–E6465. doi: 10.1073/pnas.1518552112

Schaffner, W. (2015). Enhancers, enhancers – from their discovery to today's universe of transcription enhancers. *Biol. Chem.* 396:311. doi: 10.1515/hsz-2014-0303

Schmidt, S. V., Krebs, W., Ulas, T., Xue, J., Baßler, K., Günther, P., et al. (2016). The transcriptional regulator network of human inflammatory macrophages is defined by open chromatin. *Cell Res.* 26, 151–170. doi: 10.1038/cr.2016.1

Schulte, L. N., Eulalio, A., Mollenkopf, H.-J., Reinhardt, R., and Vogel, J. (2011). Analysis of the host microRNA response to *Salmonella* uncovers the control of major cytokines by the let-7 family. *EMBO J.* 30, 1977–1989. doi: 10.1038/emboj.2011.94

Shannon, J. G., and Heinzen, R. A. (2009). Adaptive immunity to the obligate intracellular pathogen *Coxiella burnetii*. *Immunol. Res.* 43, 138–148. doi: 10.1007/s12026-008-8059-4

Sheehan, M., Markey, B., Cassidy, J., Ball, H. J., Duane, M., and Doherty, M. L. (2005). New transtracheal bronchoalveolar lavage technique for the diagnosis of respiratory disease in sheep. *Vet. Record* 157, 309–313. doi: 10.1136/vr.157.11.309

Sieweke, M. H., and Allen, J. E. (2013). Beyond stem cells: self-renewal of differentiated macrophages. *Science* 342:1242974. doi: 10.1126/science.1242974

Skene, P. J., and Henikoff, S. (2015). A simple method for generating high-resolution maps of genome-wide protein binding. *eLife* 4:e09225. doi: 10.7554/eLife.09225

Soares, L. M., He, P. C., Chun, Y., Suh, H., Kim, T., and Buratowski, S. (2017). Determinants of Histone H3K4 Methylation Patterns. *Mol. Cell* 68, 773.e6–785.e6. doi: 10.1016/j.molcel.2017.10.013

Sundberg, J. P., and Schofield, P. N. (2009). One medicine, one pathology, and the one health concept. *J. Am. Vet. Med. Assoc.* 234, 1530–1531. doi: 10.2460/javma.234.12.1530

Thalheim, T., Herberg, M., Loeffler, M., and Galle, J. (2017). The regulatory capacity of bivalent genes-A theoretical approach. *Int. J. Mol. Sci.* 18:1069. doi: 10.3390/ijms18051069

Tie, F., Banerjee, R., Stratton, C. A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., et al. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development* 136, 3131–3141. doi: 10.1242/dev.037127

Tuggle, C. K., Giuffra, E., White, S. N., Clarke, L., Zhou, H., Ross, P. J., et al. (2016). GO-FAANG meeting: a gathering on functional annotation of animal genomes. *Anim. Genet.* 47, 528–533. doi: 10.1111/age.12466

van der Heijden, C. D. C. C., Noz, M. P., Joosten, L. A. B., Netea, M. G., Riksen, N. P., and Keating, S. T. (2018). Epigenetics and trained immunity. *Antioxidants Redox Signal.* 29, 1023–1040. doi: 10.1089/ars.2017.7310

Vastenhouw, N. L., and Schier, A. F. (2012). Bivalent histone modifications in early embryogenesis. *Curr. Opin. Cell Biol.* 24, 374–386. doi: 10.1016/j.ceb.2012.03.009

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., et al. (2015). Enhancer evolution across 20 mammalian species. *Cell* 160, 554–566. doi: 10.1016/j.cell.2015.01.006

Wagschal, A., Delaval, K., Pannetier, M., Arnaud, P., and Feil, R. (2007). Chromatin immunoprecipitation (ChIP) on unfixed chromatin from cells and tissues to analyze histone modifications. *Cold Spring Harbor Prot.* 2007:db.rot4767. doi: 10.1101/pdb.prot4767

Wang, M., Hancock, T. P., Chamberlain, A. J., Vander Jagt, C. J., Pryce, J. E., Cocks, B. G., et al. (2018). Putative bovine topological association domains and CTCF binding motifs can reduce the search space for causative regulatory variants of complex traits. *BMC Genomics* 19:395. doi: 10.1186/s12864-018-4800-0

Wang, M., Hancock, T. P., MacLeod, I. M., Pryce, J. E., Cocks, B. G., and Hayes, B. J. (2017). Putative enhancer sites in the bovine genome are enriched with variants affecting complex traits. *Genet. Select. Evol.* 49:56. doi: 10.1186/s12711-017-0331-4

Wang, W., Xia, X., Mao, L., and Wang, S. (2019). The CCAAT/Enhancer-binding protein family: its roles in MDSC expansion and function. *Front. Immunol.* 10:1804. doi: 10.3389/fimmu.2019.01804

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* 40, 897–903. doi: 10.1038/ng.154

Won, K.-J., Chepelev, I., Ren, B., and Wang, W. (2008). Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinform.* 9:547. doi: 10.1186/1471-2105-9-547

Wood, S. H., Hindle, M. M., Mizoro, Y., Cheng, Y., Saer, B. R. C., Miedzinska, K., et al. (2020). Circadian clock mechanism driving mammalian photoperiodism. *Nat. Commun.* 11:4291. doi: 10.1038/s41467-020-18061-z

Worley, K. C. (2017). *Baylor College of Medicine Human Genome Sequencing Center (2017). Oar_rambouillet_v1.0.* GCA_002742125.1. https://www.ncbi.nlm.nih.gov/assembly/GCF_002742125.1/

Wreczycka, K., Franke, V., Uyar, B., Wurmus, R., Bulut, S., Tursun, B., et al. (2019). HOT or not: examining the basis of high-occupancy target regions. *Nucl. Acids Res.* 47, 5735–5745. doi: 10.1093/nar/gkz460

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* 9, R137. doi: 10.1186/gb-2008-9-9-r137

Zhao, C., Carrillo, J. A., Tian, F., Zan, L., Updike, S. M., Zhao, K., et al. (2015). Genome-wide H3K4me3 analysis in angus cattle with divergent tenderness. *PLoS One* 10:e0115358. doi: 10.1371/journal.pone.0115358

Zhou, V. W., Goren, A., and Bernstein, B. E. (2010). Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* 12:7. doi: 10.1038/nrg2905

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.

Copyright © 2021 Massa, Mousel, Herndon, Herndon, Murdoch and White. 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: 09 February 2021
doi: 10.3389/fgene.2020.616743

Identification of a Goat Intersexuality-Associated Novel Variant Through Genome-Wide Resequencing and Hi-C

Guang-Xin E[1†], Dong-Ke Zhou[2†], Zhu-Qing Zheng[2†], Bai-Gao Yang[1†], Xiang-Long Li[3], Lan-Hui Li[4], Rong-Yan Zhou[4], Wen-Hui Nai[5], Xun-Ping Jiang[6], Jia-Hua Zhang[1], Qiong-Hua Hong[7], Yue-Hui Ma[8], Ming-Xing Chu[8], Hui-Jiang Gao[8], Yong-Ju Zhao[1], Xing-Hai Duan[1], Yong-Meng He[1], Ri-Su Na[1], Yan-Guo Han[1], Yan Zeng[1], Yu Jiang[2] and Yong-Fu Huang[1]*

[1] Chongqing Key Laboratory of Forage & Herbivore, Chongqing Engineering Research Centre for Herbivores Resource Protection and Utilization, College of Animal Science and Technology, Southwest University, Chongqing, China, [2] Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling, China, [3] College of Animal Science and Technology, Hebei Normal University of Science & Technology, Qinghuangdao, China, [4] College of Animal Science and Technology, Agricultural University of Hebei, Baoding, China, [5] Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China, [6] Key Lab of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, College of Animal Science and Technology, Huazhong Agricultural University, Wuhan, China, [7] Department of Herbivore Science, Yunnan Animal Science and Veterinary Institute, Kunming, China, [8] Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing, China

Background: Polled intersex syndrome (PIS) leads to reproductive disorders in goats and exerts a heavy influence on goat breeding. Since 2001, the core variant of an 11.7 kb deletion at ~129 Mb on chromosome 1 (CHI1) has been widely used as a genetic diagnostic criterion. In 2020, a ~0.48 Mb insertion within the PIS deletion was identified by sequencing in XX intersex goats. However, the suitability of this variation for the diagnosis of intersex goats worldwide and its further molecular genetic mechanism need to be clarified.

Results: The whole-genome selective sweep of intersex goats from China was performed with whole-genome next-generation sequencing technology for large sample populations and a case–control study on interbreeds. A series of candidate genes related to the goat intersexuality phenotype were found. We further confirmed that a ~0.48 Mb duplicated fragment (including *ERG* and *KCNJ15*) downstream of the ~20 Mb PIS region was reversely inserted into the PIS locus in intersex Chinese goats and was consistent with that in European Saanen and Valais black-necked goats. High-throughput chromosome conformation capture (Hi-C) technology was then used to compare the 3D structures of the PIS variant neighborhood in CHI1 between intersex and non-intersex goats. A newly found structure was validated as an intrachromosomal rearrangement. This inserted duplication changed the original spatial structure of goat CHI1 and caused the appearance of several specific loop structures in the adjacent ~20 kb downstream region of *FOXL2*.

Conclusions: Results suggested that the novel complex PIS variant genome was sufficient as a broad-spectrum clinical diagnostic marker of XX intersexuality in goats from

OPEN ACCESS

Edited by:
Elisabetta Giuffra, Institut National de Recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), France

Reviewed by:
Terje Raudsepp, Texas A&M University, United States
Xiangdong Ding, China Agricultural University, China

***Correspondence:**
Yu Jiang
jiangyu96@163.com
Yong-Fu Huang
H67738337@swu.edu.cn

[†] These authors have contributed equally to this work

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

Received: *13 October 2020*
Accepted: *30 December 2020*
Published: *09 February 2021*

Citation:
E G-X, Zhou D-K, Zheng Z-Q, Yang B-G, Li X-L, Li L-H, Zhou R-Y, Nai W-H, Jiang X-P, Zhang J-H, Hong Q-H, Ma Y-H, Chu M-X, Gao H-J, Zhao Y-J, Duan X-H, He Y-M, Na R-S, Han Y-G, Zeng Y, Jiang Y and Huang Y-F (2021) Identification of a Goat Intersexuality-Associated Novel Variant Through Genome-Wide Resequencing and Hi-C. Front. Genet. 11:616743. doi: 10.3389/fgene.2020.616743

Europe and China. A series of private dense loop structures caused by segment insertion into the PIS deletion might affect the expression of *FOXL2* or other neighboring novel candidate genes. However, these structures require further in-depth molecular biological experimental verification. In general, this study provided new insights for future research on the molecular genetic mechanism underlying female-to-male sex reversal in goats.

Keywords: intersexuality, genome-wide selection, Hi-C, copy number variant, translocation

BACKGROUND

In as early as the nineteenth century, people regarded hornlessness as a beneficial and important economic trait and bred specialized hornless goat strains. However, during breeding, the proportion of intersex individuals in the hornless goat population gradually increased. This phenomenon was termed polled intersex syndrome (PIS) (Eaton, 1945). Intersexuality, the phenomenon wherein certain dioecious organisms possess both sexes, has been widely observed in various livestock species (Bosu and Basrur, 1984; Wang and Zhang, 1993), including goats (Ramadan and El Hassan, 1988; Ramadan et al., 1991), within the last century. The proportion of intersex goats within the global population is 4–15% (Zhan et al., 1994; Chen et al., 2010; Song et al., 2015). Reproductive system malformations in PIS goats lead to the loss of reproductive capacity and are thus some of the great challenges encountered in the development of the goat industry.

In 1996, the CHI1q41–q45 genomic regions were confirmed to be linked to hornlessness (Vaiman et al., 1996, 1999). Various molecular methods, such as chromosome walking technology and sequencing, have been used to refine the PIS locus to <100 Kb (Schibler et al., 2000). In 2001, the indicator of PIS in goats was mapped and resolved to an 11.7 kb non-coding deletion in CHI1q43 that was located ∼200 kb upstream of the *FOXL2* gene (Pailhoux et al., 2001). *FOXL2* is an important sex determination gene (Bagheri-Fam et al., 2017; Tao et al., 2018) with a key role in ovarian development (Pannetier et al., 2016; Elzaiat et al., 2017). For example, a previous study on mouse models with ovarian *FOXL2* gene deletion showed that $FOXL2^{+/-}$ mice have a normal phenotype, $FOXL2^{+/+}$ mice have a similar phenotype as patients with human blepharophimosis syndrome, and $FOXL2^{-/-}$ mice exhibit narrow eye slits and premature ovarian failure (Baron et al., 2005). Furthermore, by using gene editing technology, Boulanger et al. (2014) verified that the loss of function and expression silencing of *FOXL2* can cause female-to-male sex reversal in XX goats (Boulanger et al., 2014).

Moreover, the development of PIS diagnostic molecular markers can effectively avoid the misdiagnosis caused by the phenotypic identification of obscure PIS cases. Although a series of PIS diagnostic methods based on PCR amplification has been reported (Yang et al., 2012; Zhang et al., 2019), some studies on the diversification of PIS deletion structure have questioned the accuracy of these methods (Li et al., 2011; Kijas et al., 2013). For example, some intersex Rangeland goats do not exhibit the known homozygous PIS deletion (Kijas et al., 2013). Therefore, whether PIS deletion is specific for the diagnostics of intersexuality in goats remains controversial. Notably, the long-read whole-genome sequencing of two (one Saanen and one Valais Blacknecked black goats) genetically female (XX) intersex goats (Simon et al., 2020) demonstrated that a highly complex structural variant involving a ∼0.48 Mb duplicated segment from ∼21 Mb of chromosome 1 (CHI1) is inversely inserted into the known PIS deletion and that the length of the PIS deletion has also been shortened to 10.159 kb from 11.7 kb (Pailhoux et al., 2001).

In this study, we, for the first time, identified the intersex-related genetic variation structure of the Chinese goat population via high-throughput sequencing technology and analyzed the chromosomal spatial structure of the PIS-related genetic structure through high-throughput chromosome conformation capture (Hi-C) technology to obtain an in-depth understanding of the molecular genetic mechanism of PIS. Our work could also provide a valuable reference for the future development of diagnostic tools with enhanced broad-spectrum recognition capabilities.

METHODS

Genomic Library Construction and Sequencing

All the experimental conditions of this study were approved by the Committee on the Ethics of Animal Experiments of the Southwest University (No. [2007] 3) and the Animal Protection Law of China.

We collected venous blood samples from 55 goats comprising 35 intersex goats (26 XX Tangshan dairy goats and 9 XX Chinese southern native goats) and 20 XX non-intersex Tangshan dairy goats. A total of 2 mL venous blood was collected from each animal (Sampling from Tangshan dairy goat breeding farm, Tangshan, China). The wound was sterilized with 70% medical alcohol. All 55 animals were returned to the pasture to continue living after experimentation. All genomic DNA samples were extracted by using a QIAGEN DNeasy Blood & Tissue kit in accordance with the manufacturer's protocol. Sequencing libraries were constructed with DNA extracts and a NEBNext® Ltra DNA library preparation kit (Illumina®, US). Sequencing

Abbreviations: PIS, Polled intersex syndrome; CNV, Copy number variant; FST, Pairwise fixation index; SNP, Single nucleotide polymorphism; Hi-C, High-through chromosome conformation capture; CHI, *Capra hircus* chromosome; *FOXL2,* Forkhead box L2; *KCNJ15,* Potassium inwardly rectifying channel subfamily J member 15; *PIK3CB,* Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit beta.

was performed on an Illumina HiSeq × Ten platform (pair-end 150 bp). The sequencing data generated in this study were deposited in the NCBI SRA database (SRR10051499-SRR10551533 and SRR10613872-SRR10613891). In addition, we downloaded 166 non-intersex goat genome sequences from the NCBI SRA database. The detailed information of the 221 animals used in this study is shown in **Supplementary Table 1**.

Read Filtering, Read Alignment, and Variant Calling

Raw sequencing reads were trimmed and filtered by using Trimmomatic (version 0.36). We then mapped the clean pair-end reads to a goat (*Capra hircus*) reference genome (ARS1) by using BWA-MEM (version 0.7.13) with default parameters except that "-M" was enabled to flag shortened split hits as secondary data. We used Picard (version 2.1.1, http://broadinstitute.github.io/picard) to remove potential PCR duplicates. Finally, the reads were locally realigned around indels with the IndelRealigner procedure in GATK (version 3.7). We applied GATK to call variants and used the HaplotypeCaller algorithm in Genomic Variant Call Format (GVCF) mode. Variants were called individually for each animal, and one GVCF file that listed genotype likelihoods was generated per animal. Then, the variants were called from the GVCF files through joint genotyping analysis. We removed SNPs that were within the three base pairs of an indel by utilizing bcftools (version 1.8). Biallelic SNPs were retained by applying a hard filter of QD $<$ 2.0, MQ $<$ 40.0, FS $>$ 60.0, SOR $>$ 3.0, MQRankSum $<$ −12.5, or ReadPosRankSum $<$ −8.0. We also used vcftools (version 0.1.14) to remove SNPs with a missing rating of more than 0.1. The copy number variations (CNVs) with a silhouette score of $<$0.65 and a MAF of $<$0.05 were identified by using CNVcaller software (Wang et al., 2017a).

Genome-Wide Selective Sweep Analysis and Gene Annotation

Here, we carried out whole-genome selection signal analysis with two groups: (1) 35 intersex goats (case group, including 26 intersex Tangshan dairy goats [X/X] and 9 intersex Chinese goats [X/X] from southern China) vs. 186 non-intersex individuals (control group) and (2) 26 intersex goats (case group) vs. 20 non-intersex individuals (X/X, control group) from the Tangshan dairy goat population. For the SNP dataset, we calculated the pairwise fixation index (F_{ST}) and π ratio ($\pi_{intersex}/\pi_{non-intersex}$) with 40 kb sliding windows and 10 kb step size. Only windows passing the above two thresholds were retained. Candidate genes were subjected to functional enrichment with an online tool (KEGG, http://www.genome.jp/kegg/pathway.html).

Additionally, we calculated V_{ST} and F_{ST} on the basis of absolute copy number (CN) to identify divergent CNV profiles between XX intersex and normal female goats. V_{ST} was calculated by using following formula: $V_{ST} = \frac{V_{total} - (V_{pop1} \times N_{pop1} + V_{pop2} \times N_{pop2})/N_{total}}{V_{total}}$, where V_{total} is the total variance, V_{pop} is the CN variance for each population, N_{pop} is the sample size for each population, and N_{total} is the total sample size.

Lastly, we calculated linkage disequilibrium by using Arlequin software version 3.5.1.3 (Excoffier and Lischer, 2010) with a permutation test (EM algorithm, permutation number = 100,000).

PCR Amplification to Verify Structural Variant Genotypes

The primers (**Supplementary Table 2**) of breakpoints based on the ∼0.48 Mb fragment (CHI1:150,334,286–150,818,099) that was reversely inserted into the PIS deletion region (∼10.16 kb, CHI1:129,424,780–129,434,939) were designed with the online tool Primer 3.04 software (http://bioinfo.ut.ee/primer3-0.4.0/) to identify the structural characteristics of the identified duplication variant in CHI1 in intersex goats. 2×TransTaq® High Fidelity (HiFi) PCR SuperMix II (TransGen Biotech, China) was used in PCR amplification. The qPCR reaction conditions consisted of an initial denaturation at 94°C for 5 min, followed by 35 cycles of denaturation at 94°C for 30 s, annealing at the locus-specific temperatures presented in **Supplementary Table 2** for 30 s, and extension at 72°C for 120 s. Finally, an elongation step (final extension) was performed at 72°C for 7 min. Two-way Sanger sequencing was performed on an ABI 3730 sequencer platform (Life Technologies, US).

Three-Dimensional Genome Structure Comparison Between Intersex and Non-intersex Goats

Hi-C was performed on two individuals (Dazu black goat, China). One was intersex (hornless, PIS +/+), and the other was non-intersex (horned, female, PIS –/–). Sample processing (treatment of leukocytes from venous blood with the cell cross linker paraformaldehyde) and library construction were performed by using standard methods (four cutter restriction enzyme [MboI], Belton et al., 2012). Briefly as: (1) treat cells with paraformaldehyde (37% formaldehyde) to fix the conformation of DNA; (2) treat cross linked DNA with restriction enzymes (four cutter restriction enzyme, MboI) to produce sticky ends; (3) repair DNA ends with biotin labeling; (4) connect the DNA fragments by DNA ligase; (5) release the cross linked DNA state with 2.5M Glycine; (6) purify the DNA by AMPure XP system (Beckman Coulter, Beverly, USA) and randomly break into 300∼500 bp fragments; (7) construction of small DNA fragment library using NEB Next Ultra DNA Library Prep Kit (NEB, USA). After library construction, Qubit2.0 was used for preliminary quantification. Then, the library was diluted to 1 ng /μL. Agilent 2100 was used to determine whether the insert size of the library met expectations. Q-PCR was used to quantify accurately the effective concentration of the library ($>$2 nM), and sequencing was finally performed with an Illumina HiSeq×Ten PE150™ platform. Sequencing data quality control, reference genome alignment (ARS1), interaction map construction, and loop structure analysis were performed with Juicer software (Durand et al., 2016) with the standard parameters (Mbol restriction enzyme chunk size set at: 80000000

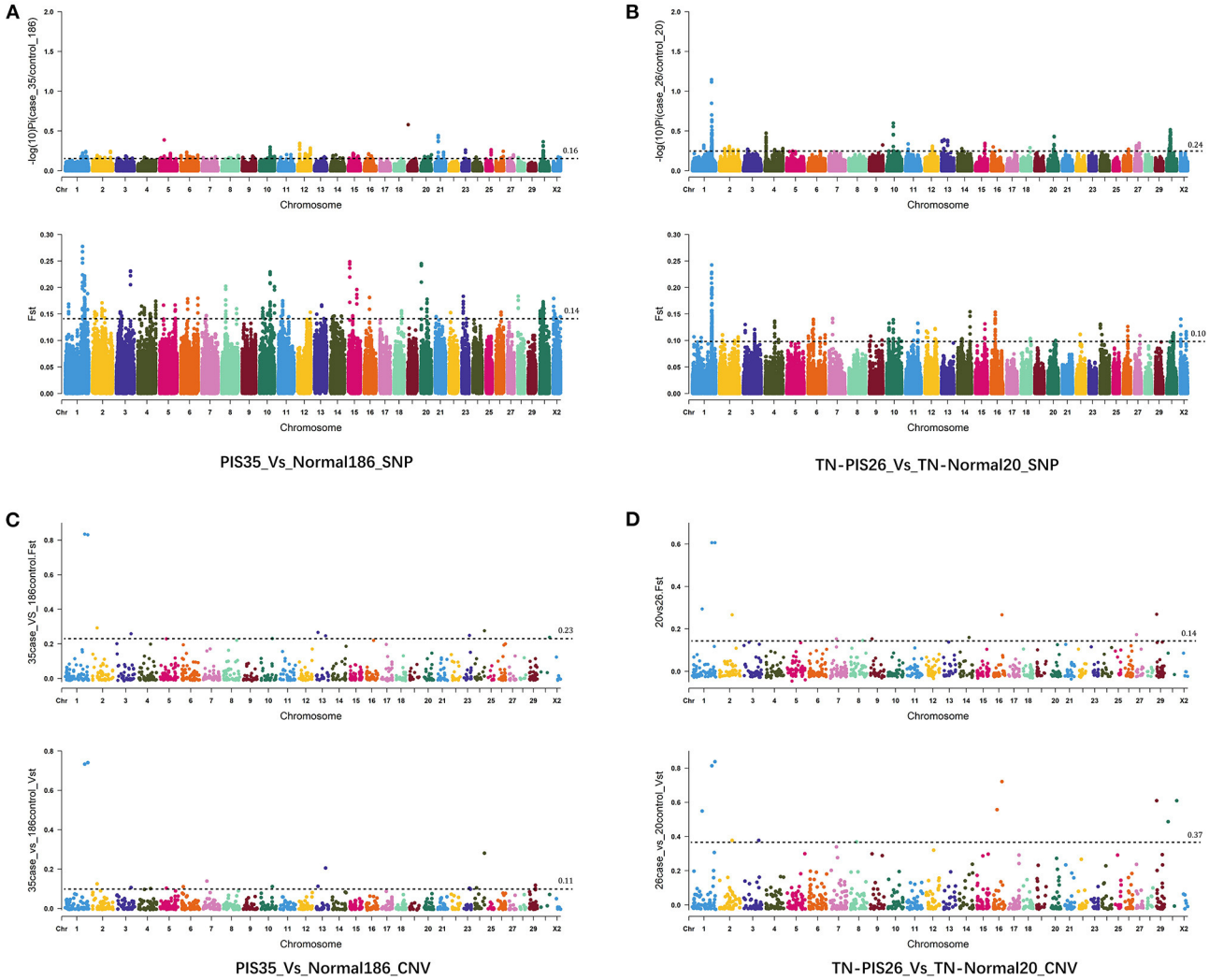


FIGURE 1 | Genome-wide selective sweep of goat intersexuality by using SNPs and CNVs. **(A)** Manhattan plot showing the SNP-based selection signals of 35 intersex goats compared with those of 186 non-intersex goats from a large geographically distributed population. **(B)** Manhattan plot showing the SNP-based selection signal of intersex goats within the Tangshan dairy goat population. **(C)** Manhattan plot showing the CNV-based selection signals of 35 intersex goats compared with those of 186 non-intersex goats from a large geographically distributed population. **(D)** Manhattan plot showing the CNV-based selection signals of intersex goats within the Tangshan dairy goat population.

bp). Image visualization was performed by using the matplotlib package in the Python environment.

RESULTS

A total of 16 462,769 SNPs and 1,058 CNVs were obtained from 221 samples. For the genome-wide selection of SNPs in all individuals (35 intersex vs. 186 non-intersex goats), we screened 258,064 windows and estimated their F_{ST} and π ratios (**Figure 1A**, **Supplementary Table 3**). In total, we identified 40 windows in accordance with the intersection of the top 1% selective regions of both parameters (F_{ST} and π ratio). These regions included 74 coding genes, which encompassed or were located up- and downstream within the 300 kb range of the window. Six of these genes were annotated to seven known signaling pathways (**Supplementary Table 4**), including neuroactive ligand-receptor interaction (*P2RY13* and *P2RY14*), hippo signaling pathway-multiple species (*STK3*), thyroid hormone signaling pathway (*MED12L*), hippo signaling pathway (*STK3*), *MTOR* signaling pathway (*RRAGB*), protein processing in endoplasmic reticulum (*UBQLN2*), and *MAPK* signaling pathway (*STK3*).

Furthermore, we selected 46 Tangshan dairy goats from large samples, set up a scientific case–control analysis test, and identified the selected signal regions of intersexuality within Tangshan dairy goat populations to prevent the genetic background divergence of large sample populations

from interfering with selection signal analysis (**Figure 1B**, **Supplementary Table 5**). The results revealed that 50 windows were generated by the intersection of the top 1% selective regions of the F_{ST} and π ratios and 79 genes, which covered or were located up- and downstream within the 300 kb range of the window. Only four genes were enriched in known pathways (**Supplementary Table 6**). These genes included *RBP2* (vitamin digestion and absorption), *NCK1* (*ErbB* signaling pathway, T cell receptor signaling pathway, and Axon guidance), *IL20RB* (Jak-*STAT* signaling pathway and cytokine–cytokine receptor interaction), and *MRAS* (tight junction, phospholipase D signaling pathway, proteoglycans in cancer, *Rap1* signaling pathway, regulation of actin cytoskeleton, *Ras* and *MAPK* signaling pathways, and HTLV-I infection). Numerous consecutive windows in CHI1 (∼129 to ∼132 Mb) of intersex Tangshan dairy goats showed strong F_{ST} and π ratio signals, and these windows covered the 11.7 kb fragment deletion (PIS) that is widely recognized genomic signature of XX intersex goats (Pailhoux et al., 2001).

The selective CNV-based sweep analysis of intersexuality with a large population and various genetic backgrounds of non-intersex goats (35 vs. 186) revealed that two CNVs had the highest signals with F_{ST} (V1: F_{ST} = 0.834565, CHI1:129,424,780–129,434,939; V2: F_{ST} = 0.830614, CHI1:150,334,286–150,818,099) and V_{ST} (V1–V_{ST} = 0.73290; V2–V_{ST} = 0.74050) (**Figure 1C**, **Supplementary Table 5**). These two variants also carried the most prominent signal in the Tangshan dairy goat population (20 vs. 26; V1: F_{ST} = 0.60641, V_{ST} = 0.81321; V2: F_{ST} = 0.60641, V_{ST} = 0.83742; **Figure 1D**, **Supplementary Table 7**). The V1 variant was contained within the known intersex-related variant region (PIS deletion). Two genes, namely, *MRPS22* (∼120 kb distance) and *COBB2* (∼140 kb distance), were found upstream of V1, whereas the *FOXL2* gene was found 340 kb downstream of V1. Other seven noncoding RNA (*LOC102190268*, *LOC108636915*, *LOC102185085*, *LOC108636375*, *LOC102190822*, *LOC102191084*, and *LOC100861210*) were discovered between *FOXL2* and V1. Furthermore, *KCNJ15* and *ERG* were encompassed by the V2 variant region, and *ETS2* was located 230 kb downstream. No coding gene was found within the 500 kb upstream region of V2.

The length of the PIS deletion (V1) on CHI1 was ∼10.16 kb and was located from 129,424,780 bp to 129,434,939 bp (**Figures 2A,B**) as observed by using the IGV browser (Thorvaldsdóttir et al., 2013). We found that the length of the V2 variant was ∼0.48 Mb and that this variant was distributed on CHI1 at 150,334,286–150,818,099 (**Figure 2C**). The different genotypes of the V1 and V2 variant regions could be clearly identified by comparing the read average coverage of each variant's region with that of the whole genome (**Figures 2D,E**). On the basis of the genotypes of V1 and V2, we found that the homozygous deletion of V1 and the homozygous duplication of a ∼0.48 Mb region of V2 were always simultaneously present in all intersex goats (**Figure 2F**). Linkage disequilibrium analysis revealed significant linkage ($P < 0.0001$) between the V1 and V2 mutations in 221 goats.

In the heterozygous and extra duplication homozygous individuals with the V2 mutation, a considerable number of reads were split-mapped simultaneously to the outer boundary of V1 and the inner boundary of V2 (**Supplementary Figure 1**). We verified the true boundary breakpoints of the two variant regions through PCR amplification and Sanger sequencing (**Figures 2G–J**). Therefore, the precise PIS genome structure was doubly confirmed as an inverted duplication of the ∼0.48 Mb segment that had inserted into the 10.16 kb PIS deletion.

An average of >250 Gb (∼85×) of genome coverage data were obtained from two individuals and used to construct a 3D genome high-resolution interaction map. Firstly, we gathered total 851,483,595 and 808,343,466 reads in case and control individual, respectively. Secondly, according to the mapping results of case and control dataset, there are 300,895,514 (35.34%) and 359,770,248 (44.51%) normal paired reads, 402,929,844 (47.32%) and 347,404,121 (42.98%) chimeric paired reads, 344,755,346 and 117,458,553 PCR duplicates reads, 156,341,608 and 244,912,442 intra-chromosomal interaction reads, 71,982,908 and 173,316,912 short-distance interaction sequence with interaction distance less than 20 kb (<20 kb), 84,358,664 and 149,510,017 long-range interaction sequences with interaction distance larger than 20 kb (>20 kb), respectively.

The heat map of both sets of data with 80 kb resolution revealed a potential intrachromosomal rearrangement site (**Figures 3A,B**), which was initially identified in CHI1 of an intersex individual (PIS–/–). This finding was consistent with the physical location of V2, which was absent from non-intersex goats (PIS–/–). We used a resolution of 10 kb to identify effective breakpoints (**Figures 3C,D**). A small but sharp contact peak suggested that a new intrachromosomal rearrangement occurred in CHI1 of the intersex goats (**Figures 3E,F**). We identified four private consequent loop regions in CHI1 of the intersex goats (**Supplementary Table 8**) and compared these regions with those in non-intersex individuals (**Supplementary Table 9**). These loop regions were densely clustered in the ∼20 kb downstream regions of the *FOXL2* gene, which overlapped with LOC102191651 and LOC108636917 (**Figure 3G**).

DISCUSSION

SNP-based genome-wide selection signal analysis revealed numerous sharp signals in 35 intersex goats and 186 control samples. Within the top 1% selection window, a series of genes were identified and found to be deeply involved in animal reproduction and multiple developmental processes. For example, *STK3* is a key molecule that connects the downstream signaling pathway of estrogen and the Hippo signaling pathway; it also regulates the dynamic development of the uterine epithelium during the estrous cycle through the signal transduction of uterine epithelial cells (Moon et al., 2019). *STK3* was annotated to other signaling pathways, such as the *MAPK* (Bogani et al., 2009; Warr et al., 2016) and Hippo signaling pathways, that also play an important role in gonadal development and sex determination (Frum et al., 2018; Devos et al., 2020). Although *MED12L* has been verified to be associated with fetal mental retardation in human (Nizon et al., 2019), it is also involved in reproductive development (Sayem et al., 2017;

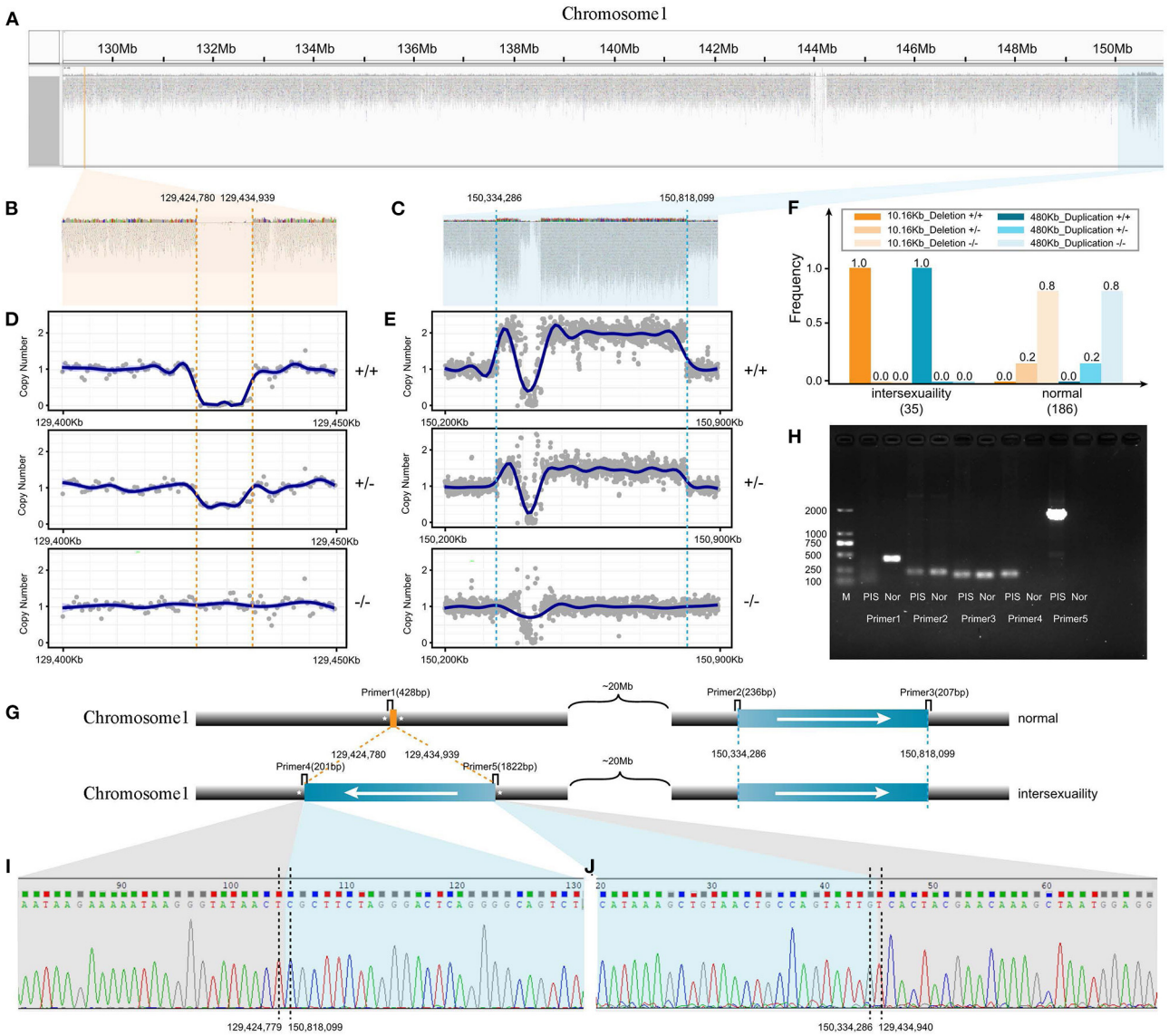


FIGURE 2 | Model detection and verification of novel transposition in intersex goats. **(A)** Two CNVs (V1 and V2) on Chromosome1 observed by IGV software. **(B)** Alignment and coverage of wide-genome short-reads from intersex goats as observed with IGV software. The CNV from 129,424,780 to 129,434,939 bp on Chromosome 1 (V1, **B**) manifested as a deletion. **(C)** Alignment and coverage of wide-genome short-reads from intersex goats as observed by IGV software. The CNV from 150,334,286 to 150,818,099 bp on Chromosome 1 manifested as a duplication (V2). **(D)** Genomic coverage of different genotypes of V1 variant reads (Chromosome 1, 129.40–129.45 Mb). **(E)** Genomic coverage of different genotype of V2 variant reads (Chromosome 1, 150.20–150.90 Mb). **(F)** Two CNV variants (V1 and V2) associated with intersexuality had the same frequency in the population. **(G)** Schematic of the Chromosome 1 PIS transposition model and location map of primers for PCR verification. **(H)** Gel electrophoresis verification of PCR results. **(I,J)** Sanger sequencing results of sequences amplified with primers 4 and 5.

Das and Kumar, 2018; Ulloa-Aguirre et al., 2018). In addition, the *RRAGB* gene is enriched in the *mTOR* signaling pathway, which is widely involved in gonadal development (Bajwa et al., 2017; Correia et al., 2020). Therefore, our findings suggested that numerous molecular mechanisms underlying development and the physiological maintenance of intersexual characteristics await further excavation.

Notably, the different genetic backgrounds of large samples can cause many false-positive genes, and chromosomal regions may thus be incorrectly identified. Given the inconsistent ratios between the numbers of Tangshan dairy goats in the intersex and control groups, a gene with considerable breed specificity caused interference. We performed a strict case–control experiment on the Tangshan dairy goat population to prevent the interference of specific population backgrounds and identified a series of interesting genes. The highest continuous selection signal was observed in CHI1. These signals covered the areas of a previously reported PIS deletion (Pailhoux et al., 2001) and six coding

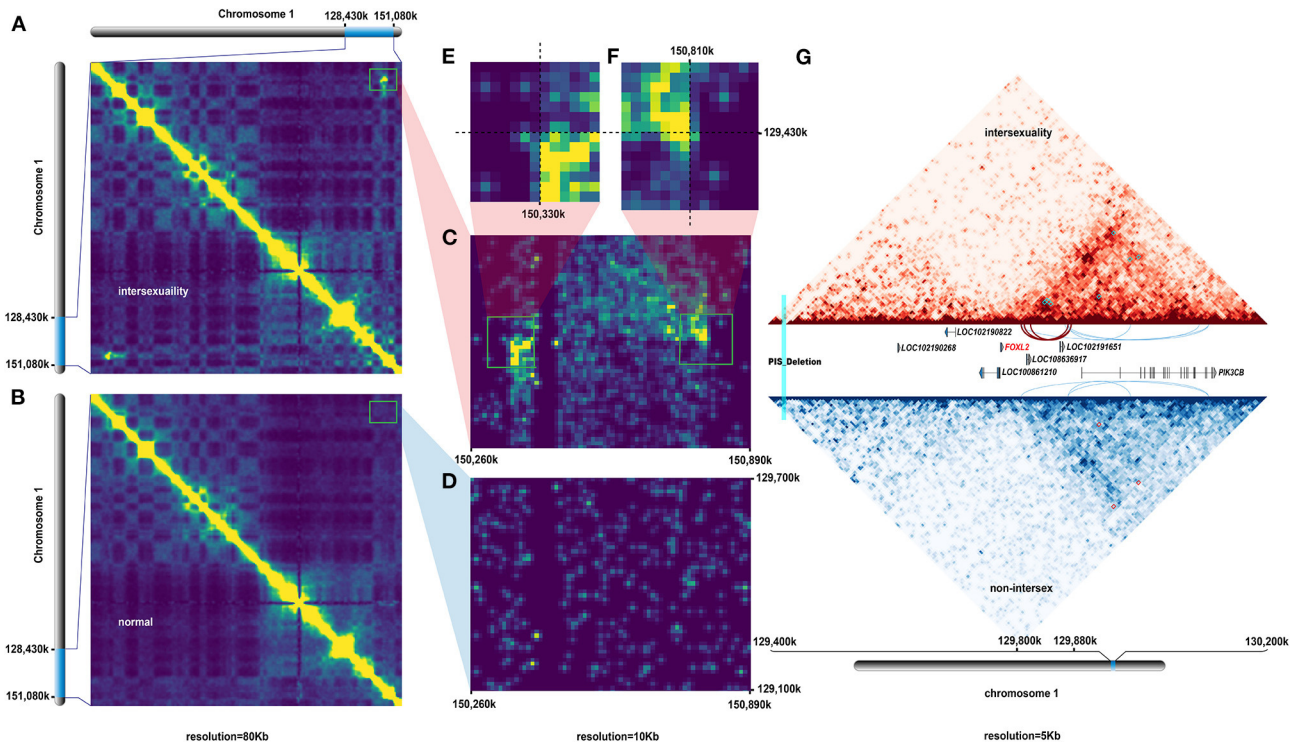


FIGURE 3 | Hi-C analysis results. **(A)** Window interaction matrix on CHI1 from 128.43 to 151.08 Mb in intersexuality individuals with 80 Kb resolution. **(B)** Window interaction matrix on CHI1 from 128.43 to 151.08 Mb in normal control individuals with 80 Kb resolution. **(C–F)** Interaction matrix and comparison of the 129.1–129.7 and 150.3–150.9 Mb regions on CHI1 between intersexuality and normal individuals with 10 Kb resolution. **(G)** Analysis of loop conformation in the 129.4–130.2 Mb region of CHI1 between intersexuality and normal individuals at 5 kb resolution revealed that the intersexuality case had a particular loop conformation in the 129.80–129.88 Mb region at which the sexual development-associated gene *FOXL2* was located 30 Kb upstream.

genes (*MRAS*, *NMNAT3*, *ARMC8*, *DBR1*, *LOC108636376*, and *SCLC35G2*); *SOX14* and *MRPS22* in the upstream region; and *LOCl02190268* and *IL20RB* in the downstream region.

Some genes with cellular biological importance were identified. For example, *NMNAT3* maintains cell differentiation by maintaining mitochondrial content (Son et al., 2016; Yu et al., 2020). *ARMC8* is involved in the adherence of regulatory cells to cells and is associated with cell differentiation in ovarian cancer tumors (Jiang et al., 2015; Gul et al., 2019). *SOX14* is associated with apoptosis in cancer cells in the sexual reproductive system (Stanisavljevic et al., 2017) and is a crucial determinant of allergy development in *Drosophila* (Ritter and Beckstead, 2010). Interestingly, the conserved region of the *MRPS22* gene is a long-range enhancer and regulates the expression of *FOXL2* through an unclear advanced *cis*-regulatory effect of chromatin structure in humans and rats (Crisponi et al., 2004). Thus, our case–control experiment involving a population with a single genetic background enabled us to screen out many false-positive signals and identify a series of credible candidate genes. The results of this experiment provided insight into the molecular genetic mechanism of intersexuality-related physiology.

The recent identification of the gene transcription profiles of intersex and normal goat gonads through the use of RNA-Seq technology suggests that a large number of differentially expressed genes may be involved in the regulation of sex determination and differentiation in intersex goats (Yang et al., 2020). This result reminds us that many potential molecular mechanisms under the goat sexual reversal phenotype remain unclear.

Our CNV-based analysis results showed that equally strong signals were generated in V1 and V2 in the large sample with different genetic backgrounds and the Tangshan dairy goat population in the case–control analysis. These signals were recognized as the 10.16 kb PIS deletion and the ∼0.4838 Mb duplicated segment located ∼20.9 Mb further downstream of the PIS deletion and ∼150 Mb on CHI1. In addition, as expected, the highly complex structure was identified as the additional 0.4838 Mb-sized duplicated segment that was inversely inserted at the breakpoint of the 10.16 kb deletion. Our findings were consistent with the recent research results from a team in Germany that used long-read whole-genome sequencing (Simon et al., 2020). Although we utilized short-read sequencing technology, the large sample size and classic case–control experimental design still achieved the same effect. Our study confirmed that the XX intersex goats from the hornless goat population in China share the same PIS genome variant structure with European goats.

In accordance with previous studies that identified the segment size and polymorphism in PIS deletion (Li et al., 2011).

We believe that some loss in the ASR1 assembly occurred on the last 180 bp section, while was not lost on the previous 11.7 kb PIS deletion sequence (GenBank No. AF404302) investigated by Simon et al. (2020). It was adjacent to the *PISRT1* gene with the closest physical distance. However, previous studies have shown that *PISRT1* does not participate in the expression of *FOXL2* and the determination/differentiation of the gonads. For example, the overexpression of *PISRT1* in PIS−/− fetuses does not affect *FOXL2* expression levels and gonadal development (Boulanger et al., 2008).

The duplicated segment contained the *KCNJ15* and *ERG* genes. The extra copies of these two genes have an essential role in horn and gonadal development. *KCNJ15* is known to participate in insulin secretion (Okamoto et al., 2012), nervous system diseases (Zhou et al., 2018), gastric acid secretion (Yuan et al., 2015), kidney cancer (Liu et al., 2019), and esophageal squamous cell carcinoma (Nakamura et al., 2020). It is also involved in gastric acid secretion and regulation (He et al., 2011), and the relationship between gastric acid secretion and the effects of sex hormones was verified decades ago (Adeniyi, 1991). The high expression of *KCNJ15* in follicle-associated epithelium suggests that *KCNJ15* may be involved in the functional development of the ovary (Kobayashi et al., 2012) and implicates this gene in female gonadal development. Furthermore, a large number of studies have shown that *ERG* is not only an oncogene that is related to a variety of cancers (Wang et al., 2017b; Zhang et al., 2020), it also participates in the embryonic developmental processes, including bone development (Iwamoto et al., 2000), of a variety of organisms (Furlan et al., 2005; Nikolova-Krstevski et al., 2009). This participation indicates that the *ERG* gene may be related to horn and embryonic development.

Furthermore, Hi-C technology was used to study DNA replication, transcription regulation, and DNA damage repair and contact between chromosomal loci (Cremer and Cremer, 2001; de Wit and De Laat, 2012; Maass et al., 2018). Currently, this topic is heavily explored in genomic research, and numerous studies on technical method optimization have been performed (Lin et al., 2018; Yardimci et al., 2019; Janaratne et al., 2020).

Intrachromosomal rearrangement or palindrome duplication is associated with various processes of phenotypic determination and development (Carbonell-Bejerano et al., 2017; Yin et al., 2017; Mendoza et al., 2020). We performed the loop analysis of the 3D genomes to further investigate the special chromosomal spatial structures resulting from the identified intrachromosomal rearrangement. We found several unique loop structures in CHI1 of homozygous PIS intersex goats but not in that of non-intersex individuals. Many intrachromosomal rearrangement structures can alter gene expression levels within and in areas adjacent to a gene region by altering chromosomal structure (Demura et al., 2007; Suzuki et al., 2020).

Substantial evidence indicates that many of the observed loops are related to gene regulation and serve as anchors and promoters (Ahmadiyeh et al., 2010; Hoffman et al., 2013; Rao et al., 2014). The loops that we identified in this study were consistent and clustered near the *FOXL2* gene. Numerous pieces of evidence have shown that silencing *FOXL2* expression directly affects ovarian development and oogenesis in fish (Fan et al., 2019), mice, and humans (Uda et al., 2004; Thanatsis et al., 2019). Specifically, the elimination of *FOXL2* expression is sufficient to induce female-to-male reversal in XX goats (Pannetier et al., 2012; Boulanger et al., 2014). Therefore, although the regulatory relationship between this newly discovered intrachromosomal rearrangement and *FOXL2* expression cannot be evaluated thus far, the change in spatial chromosome 3D structure in the adjacent region of *FOXL2* was evident. Whether these loop structures affect *FOXL2* expression and cause intersexuality by inhibiting *cis*-acting elements or switching *trans*-acting elements should be evaluated through in-depth molecular biology research.

In addition, we found that two genes were located within the loop region: one was trafficking protein particle complex subunit 1 pseudogene (LOC102191651), and the other was an uncharacterized non-coding RNA (LOC108636917). Additional evidence regarding the further functions of these both genes remains lacking. Therefore, we cannot conclude that these loops/two genes participate in gonadal development. However, an interesting gene, *PIK3CB,* that was located further downstream of LOC102191651 and LOC108636917 attracted our attention. Numerous studies have shown that *PIK3CB* plays an important role in the development and physiological function of the ovary (Zheng et al., 2012; Li et al., 2013; Nteeba et al., 2017). However, supernumerary data suggesting that this gene is responsible for the occurrence and maintenance of the intersexual phenotype are unavailable. Therefore, whether the novel loop region containing both genes affects *PIK3CB* expression and whether *PIK3CB* is a new essential factor that is sufficient for causing female-to-male sex reversal in XX goats need to be evaluated.

CONCLUSIONS

We performed the genome-wide selective sweep of intersex goats with wide-genome next-generation sequencing. We doubly verified that the structural variant of caprine PIS structure, a 0.48 Mb duplicated fragment located ∼20 Mb downstream of the PIS region that was reversely inserted into the PIS deletion, was sufficient as a broad-spectrum clinical diagnostic marker of XX intersex goats from Europe and China. The existence of several private dense loop structures in the region adjacent to *FOXL2* of intersex XX goats but not in that of non-intersex individuals suggested that intrachromosomal rearrangement might affect the expression of *FOXL2* or other neighboring novel candidate genes. This effect needs to be further evaluated. This study supported a precise genomic feature of PIS phenotype in intersex goats from Europe and China and provided new insights for future research on the molecular genetic mechanism underlying female-to-male sex reversal in goats.

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 at: https://www.ncbi.nlm.nih.gov/genbank/, SRR10051499-SRR10551533; https://www.ncbi.nlm.nih.gov/genbank/, SRR10613872-SRR10613891.

ETHICS STATEMENT

The animal study was reviewed and approved by Animal Experiments of the Southwest University (No. [2007] 3) Southwest University, Chongqing 400716, China.

AUTHOR CONTRIBUTIONS

G-XE, YJ, Y-FH, Y-JZ, Y-HM, J-HZ, Q-HH, M-XC, and X-LL conceived and designed the experiments. D-KZ, Z-QZ, B-GY, H-JG, Y-MH, and X-HD analyzed the data. G-XE analyzed the data and wrote the paper. YJ, X-PJ, R-SN, Y-GH, YZ, R-YZ, W-HN, and L-HL supported the samples and displayed lab work. G-XE, Y-FH, and YJ provided funding. All authors read and approved the manuscript.

FUNDING

This work was supported by the National Key R&D Program of China (No. 2018YFD0502000), Characteristic Germplasm Resources Population Selection and Innovation on Mutton Sheep and Goats (No. 2015BAD03B05), National Natural Science Foundation of China (No. 31172195), Chongqing Research Program of Basic Research and Frontier Technology (cstc2018jcyjAX0153), and Fundamental Research Funds for the Central Universities (XDJK2018B014). Chongqing Science and Technology Innovation Special Project (No. cstc2019jscx-gksbX0135). We are grateful to thank Prof. Dr. Gang Cao from Huazhong Agricultural University (Wuhan, China) for assisting in part of bioinformatics analysis of Hi-C datasets.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.616743/full#supplementary-material

REFERENCES

Adeniyi, K. O. (1991). Gastric acid secretion and parietal cell mass: effect of sex hormones. *Gastroenterology* 101, 66–69. doi: 10.1016/0016-5085(91)90460-3

Ahmadiyeh, N., Pomerantz, M. M., Grisanzio, C., Herman, P., Jia, L., Almendro, V., et al. (2010). 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9742–9746. doi: 10.1073/pnas.0910668107

Bagheri-Fam, S., Bird, A. D., Zhao, L., Ryan, J. M., Yong, M., Wilhelm, D., et al. (2017). Testis determination requires a specific FGFR2 isoform to repress FOXL2. *Endocrinology* 158, 3832–3843. doi: 10.1210/en.2017-00674

Bajwa, P., Nielsen, S., Lombard, J. M., Rassam, L., Nahar, P., Rueda, B. R., et al. (2017). Overactive mTOR signaling leads to endometrial hyperplasia in aged women and mice. *Oncotarget* 8:7265. doi: 10.18632/oncotarget.13919

Baron, D., Batista, F., Chaffaux, S., Cocquet, J., Cotinot, C., Cribiu, E., et al. (2005). Foxl2 gene and the development of the ovary: a story about goat, mouse, fish and woman. *Reprod. Nutr. Dev.* 45, 377–382. doi: 10.1051/rnd:2005028

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58, 268–276. doi: 10.1016/j.ymeth.2012.05.001

Bogani, D., Siggers, P., Brixey, R., Warr, N., Beddow, S., Edwards, J., et al. (2009). Loss of mitogen-activated protein kinase kinase kinase 4 (MAP3K4) reveals a requirement for MAPK signalling in mouse sex determination. *PLoS Biol* 7:e1000196. doi: 10.1371/journal.pbio.1000196

Bosu, W., and Basrur, P. K. (1984). Morphological and hormonal features of an ovine and a caprine intersex. *Can. J. Comp. Med.* 48:402.

Boulanger, L., Kocer, A., Daniel, N., Pannetier, M., Chesné P., Heyman, Y., et al. (2008). Attempt to rescue sex-reversal by transgenic expression of the PISRT1 gene in XX PIS–/–goats. *Sexual Development* 2, 142–151. doi: 10.1159/000143432

Boulanger, L., Pannetier, M., Gall, L., Allais-Bonnet, A., Elzaiat, M., Le Bourhis, D., et al. (2014). FOXL2 is a female sex-determining gene in the goat. *Curr. Biol.* 24, 404–408. doi: 10.1016/j.cub.2013.12.039

Carbonell-Bejerano, P., Royo, C., Torres-Pérez, R., Grimplet, J., Fernandez, L., Franco-Zorrilla, J. M., et al. (2017). Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in grapevine. *Plant Physiol.* 175, 786–801. doi: 10.1104/pp.17.00715

Chen, Y., Chen, J., and Jiang, H. Z. (2010). Research on the characteristics of intersex goats and it's related research progress. *J. China Anim. Husbandry Vet.* 37, 125–129 (in Chinese).

Correia, B., Sousa, M. I., and Ramalho-Santos, J. (2020). The mTOR pathway in reproduction: from gonadal function to developmental coordination. *Reproduction* 159, R173–R188. doi: 10.1530/REP-19-0057

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

Crisponi, L., Uda, M., Deiana, M., Loi, A., Nagaraja, R., Chiappe, F., et al. (2004). FOXL2 inactivation by a translocation 171 kb away: analysis of 500 kb of chromosome 3 for candidate long-range regulatory sequences. *Genomics* 83, 757–764. doi: 10.1016/j.ygeno.2003.11.010

Das, N., and Kumar, T. R. (2018). Molecular regulation of follicle-stimulating hormone synthesis, secretion and action. *J. Mol. Endocrinol.* 60, R131–R155. doi: 10.1530/JME-17-0308

de Wit, E., and De Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* 26, 11–24. doi: 10.1101/gad.179804.111

Demura, M., Martin, R. M., Shozu, M., Sebastian, S., Takayama, K., Hsu, W.-T., et al. (2007). Regional rearrangements in chromosome 15q21 cause formation of cryptic promoters for the CYP19 (aromatase) gene. *Hum. Mol. Genet.* 16, 2529–2541. doi: 10.1093/hmg/ddm145

Devos, M., Grosbois, J., and Demeestere, I. (2020). Interaction between PI3K/AKT and Hippo pathways during *in vitro* follicular activation and response to fragmentation and chemotherapy exposure using a mouse immature ovary model. *Biol. Reprod.* 102, 717–729. doi: 10.1093/biolre/ioz215

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., et al. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002

Eaton, O.N. (1945). The relation between polled and hermaphroditic characters in dairy goat. *Genetics* 30, 51–61.

Elzaiat, M., Todeschini, A. L., Caburet, S., and Veitia, R. (2017). The genetic make-up of ovarian development and function: the focus on the transcription factor FOXL2. *Clin. Genet.* 91, 173–182. doi: 10.1111/cge.12862

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Fan, Z., Zou, Y., Liang, D., Tan, X., Jiao, S., Wu, Z., et al. (2019). Roles of forkhead box protein L2 (foxl2) during gonad differentiation and maintenance in a fish, the olive flounder (Paralichthys olivaceus). *Reprod. Fertil. Dev.* 31, 1742–1752. doi: 10.1071/RD18233

Frum, T., Murphy, T. M., and Ralston, A. (2018). HIPPO signaling resolves embryonic cell fate conflicts during establishment of pluripotency *in vivo*. *Elife* 7:e42298. doi: 10.7554/eLife.42298.020

Furlan, F., Guasti, L., Avossa, D., Becchetti, A., Cilia, E., Ballerini, L., et al. (2005). Interneurons transiently express the ERG K+ channels during development of mouse spinal networks *in vitro*. *Neuroscience* 135, 1179–1192. doi: 10.1016/j.neuroscience.2005.06.040

Gul, I. S., Hulpiau, P., Sanders, E., Van Roy, F., and van Hengel, J. (2019). Armc8 is an evolutionarily conserved armadillo protein involved in cell–cell adhesion complexes through multiple molecular interactions. *Biosci. Rep.* 39:BSR20180604. doi: 10.1042/BSR20180604

He, W., Liu, W., Chew, C. S., Baker, S. S., Baker, R. D., Forte, J. G., et al. (2011). Acid secretion-associated translocation of KCNJ15 in gastric parietal cells. *Am. J. Physiol. Gastrointest. Liver Physiol.* 301, G591–G600. doi: 10.1152/ajpgi.00460.2010

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., et al. (2013). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* 41, 827–841. doi: 10.1093/nar/gks1284

Iwamoto, M., Higuchi, Y., Koyama, E., Enomoto-Iwamoto, M., Kurisu, K., Yeh, H., et al. (2000). Transcription factor ERG variants and functional diversification of chondrocytes during limb long bone development. *J. Cell Biol.* 150, 27–40. doi: 10.1083/jcb.150.1.27

Janaratne, T., Wang, X. C., Becker, C., Zhao, Y., Leanna, R., and Pritts, W. (2020). Hydrophobic interaction chromatography (HIC) method development and characterization of resolved drug-load variants in site-specifically conjugated pyrrolobenzodiazepine dimer-based antibody drug conjugates (PBD-ADCs). *J. Pharm. Biomed. Anal.* 179:113027. doi: 10.1016/j.jpba.2019.113027

Jiang, G., Yang, D., Wang, L., Zhang, X., Xu, H., Miao, Y., et al. (2015). A novel biomarker ARMc8 promotes the malignant progression of ovarian cancer. *Hum. Pathol.* 46, 1471–1479. doi: 10.1016/j.humpath.2015.06.004

Kijas, J. W., Ortiz, J. S., McCulloch, R., James, A., Brice, B., Swain, B., et al. (2013). Genetic diversity and investigation of polledness in divergent goat populations using 52 088 SNPs. *Anim. Genet.* 44, 325–335. doi: 10.1111/age.12011

Kobayashi, A., Donaldson, D. S., Kanaya, T., Fukuda, S., Baillie, J. K., Freeman, T. C., et al. (2012). Identification of novel genes selectively expressed in the follicle-associated epithelium from the meta-analysis of transcriptomics data from multiple mouse cell and tissue populations. *DNA Res.* 19, 407–422. doi: 10.1093/dnares/dss022

Li, Q., He, H., Zhang, Y.-L., Li, X.-M., Guo, X., Huo, R., et al. (2013). Phosphoinositide 3-kinase p110δ mediates estrogen-and FSH-stimulated ovarian follicle growth. *Mol. Endocrinol.* 27, 1468–1482. doi: 10.1210/me.2013-1082

Li, X., Zhang, J., Zhou, R., Li, L., and Zheng, G. (2011). Special variations within 11.7 kb fragment in goat polled intersex syndrome. *Afr. J. Biotechnol.* 10, 6695–6699. doi: 10.5897/AJB11.001

Lin, D., Hong, P., Zhang, S., Xu, W., Jamal, M., Yan, K., et al. (2018). Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture. *Nat. Genet.* 50, 754–763. doi: 10.1038/s41588-018-0111-2

Liu, Y., Wang, H., Ni, B., Zhang, J., Li, S., Huang, Y., et al. (2019). Loss of KCNJ15 expression promotes malignant phenotypes and correlates with poor prognosis in renal carcinoma. *Cancer Manag. Res.* 11:1211. doi: 10.2147/CMAR.S184368

Maass, P. G., Barutcu, A. R., Weiner, C. L., and Rinn, J. L. (2018). Inter-chromosomal contact properties in live-cell imaging and in Hi-C. *Mol. Cell* 69, 1039–1045. e3. doi: 10.1016/j.molcel.2018.02.007

Mendoza, M. N., Schalnus, S. A., Thomson, B., Bellone, R. R., Juras, R., and Raudsepp, T. (2020). Novel complex unbalanced dicentric X-autosome rearrangement in a thoroughbred mare with a mild effect on the phenotype. *Cytogen. Genome Res.* 160, 597–609. doi: 10.1159/000511236

Moon, S., Lee, O.-H., Lee, S., Lee, J., Park, H., Park, M., et al. (2019). STK3/4 expression is regulated in uterine endometrial cells during the estrous cycle. *Cells* 8:1643. doi: 10.3390/cells8121643

Nakamura, S., Kanda, M., Koike, M., Shimizu, D., Umeda, S., Hattori, N., et al. (2020). KCNJ15 expression and malignant behavior of esophageal squamous cell carcinoma. *Ann. Surg. Oncol.* 27, 2559–2568. doi: 10.1245/s10434-019-08189-8

Nikolova-Krstevski, V., Yuan, L., Le Bras, A., Vijayaraj, P., Kondo, M., Gebauer, I., et al. (2009). ERG is required for the differentiation of embryonic stem cells along the endothelial lineage. *BMC Dev. Biol* 9:72. doi: 10.1186/1471-213X-9-72

Nizon, M., Laugel, V., Flanigan, K. M., Pastore, M., Waldrop, M. A., Rosenfeld, J. A., et al. (2019). Variants in MED12L, encoding a subunit of the mediator kinase module, are responsible for intellectual disability associated with transcriptional defect. *Genet. Med.* 21, 2713–2722. doi: 10.1038/s41436-019-0557-3

Nteeba, J., Ganesan, S., Madden, J. A., Dickson, M. J., and Keating, A. F. (2017). Progressive obesity alters ovarian insulin, phosphatidylinositol-3 kinase, and chemical metabolism signaling pathways and potentiates ovotoxicity induced by phosphoramide mustard in mice. *Biol. Reprod.* 96, 478–490. doi: 10.1095/biolreprod.116.143818

Okamoto, K., Iwasaki, N., Doi, K., Noiri, E., Iwamoto, Y., Uchigata, Y., et al. (2012). Inhibition of glucose-stimulated insulin secretion by KCNJ15, a newly identified susceptibility gene for type 2 diabetes. *Diabetes* 61, 1734–1741. doi: 10.2337/db11-1201

Pailhoux, E., Vigier, B., Chaffaux, S., Servel, N., Taourit, S., Furet, J.-P., et al. (2001). A 11.7-kb deletion triggers intersexuality and polledness in goats. *Nat. Genet.* 29, 453–458. doi: 10.1038/ng769

Pannetier, M., Chassot, A.-A., Chaboissier, M.-C., and Pailhoux, E. (2016). Involvement of FOXL2 and RSPO1 in ovarian determination, development, and maintenance in mammals. *Sex. Dev.* 10, 167–184. doi: 10.1159/000448667

Pannetier, M., Elzaiat, M., Thépot, D., and Pailhoux, E. (2012). Telling the story of XX sex reversal in the goat: highlighting the sex-crossroad in domestic mammals. *Sex. Dev.* 6, 33–45. doi: 10.1159/000334056

Ramadan, R., and El Hassan, A. M. (1988). Intersexuality in goats. *N. Z. Vet. J.* 36, 120–124. doi: 10.1080/00480169.1988.35505

Ramadan, R., Gameel, A., Dafalla, E., and Galil, A. (1991). Bilateral scrotal hysterocele in an intersex goat. *J. Vet. Med. Ser. A* 38, 441–444. doi: 10.1111/j.1439-0442.1991.tb01033.x

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, 1665–1680. doi: 10.1016/j.cell.2014.11.021

Ritter, A. R., and Beckstead, R. B. (2010). Sox14 is required for transcriptional and developmental responses to 20-hydroxyecdysone at the onset of Drosophila metamorphosis. *Dev. Dyn.* 239, 2685–2694. doi: 10.1002/dvdy.22407

Sayem, A. S. M., Giribabu, N., Muniandy, S., and Salleh, N. (2017). Effects of thyroxine on expression of proteins related to thyroid hormone functions (TR-α, TR-β, RXR and ERK1/2) in uterus during peri-implantation period. *Biomed. Pharmacother.* 96, 1016–1021. doi: 10.1016/j.biopha.2017.11.128

Schibler, L., Cribiu, E. P., Oustry-Vaiman, A., Furet, J.-P., and Vaiman, D. (2000). Fine mapping suggests that the goat Polled Intersex Syndrome and the human Blepharophimosis Ptosis Epicanthus Syndrome map to a 100-kb homologous region. *Genome Res.* 10, 311–318. doi: 10.1101/gr.10.3.311

Simon, R., Lischer, H., Pieńkowska-Schelling, A., Keller, I., Häfliger, I. M., Letko, A., et al. (2020). New genomic features of the polled intersex syndrome variant in goats unraveled by long-read whole-genome sequencing. *Anim. Genet.* 51, 439–448. doi: 10.1111/age.12918

Son, M. J., Kwon, Y., Son, T., and Cho, Y. S. (2016). Restoration of mitochondrial NAD+ levels delays stem cell senescence and facilitates reprogramming of aged somatic cells. *Stem Cells* 34, 2840–2851. doi: 10.1002/stem.2460

Song, C. F. L. Y., Zhang, D. X., Jiao, L., Wang, G. M., and Liu, X. J. (2015). Anatomical observation of reproductive organs in intersex goats. *Heilongjiang Anim. Sci. Vet. Med.* 16:81. (in Chinese).

Stanisavljevic, D., Petrovic, I., Vukovic, V., Schwirtlich, M., Gredic, M., Stevanovic, M., et al. (2017). SOX14 activates the p53 signaling pathway and induces apoptosis in a cervical carcinoma cell line. *PLoS ONE* 12:e0184686. doi: 10.1371/journal.pone.0184686

Suzuki, M., Katayama, S., and Yamamoto, M. (2020). Two effects of GATA2 enhancer repositioning by 3q chromosomal rearrangements. *IUBMB Life* 72, 159–169. doi: 10.1002/iub.2191

Tao, W., Chen, J., Tan, D., Yang, J., Sun, L., Wei, J., et al. (2018). Transcriptome display during tilapia sex determination and differentiation as revealed by RNA-Seq analysis. *BMC Genomics* 19:363. doi: 10.1186/s12864-018-4756-0

Thanatsis, N., Kaponis, A., Koika, V., Georgopoulos, N. A., and Decavalas, G. O. (2019). Reduced Foxo3a, FoxL2, and p27 mRNA expression in human

ovarian tissue in premature ovarian insufficiency. *Hormones* 18, 409–415. doi: 10.1007/s42000-019-00134-4

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics* 14, 178–192. doi: 10.1093/bib/bbs017

Uda, M., Ottolenghi, C., Crisponi, L., Garcia, J. E., Deiana, M., Kimber, W., et al. (2004). Foxl2 disruption causes mouse ovarian failure by pervasive blockage of follicle development. *Hum. Mol. Genet.* 13, 1171–1181. doi: 10.1093/hmg/ddh124

Ulloa-Aguirre, A., Zariñán, T., Jardón-Valadez, E., Gutiérrez-Sagal, R., and Dias, J. A. (2018). Structure-function relationships of the follicle-stimulating hormone receptor. *Front. Endocrinol.* 9:707. doi: 10.3389/fendo.2018.00707

Vaiman, D., Koutita, O., Oustry, A., Elsen, J.-M., Manfredi, E., Fellous, M., et al. (1996). Genetic mapping of the autosomal region involved in XX sex-reversal and horn development in goats. *Mamm. Genome* 7, 133–137. doi: 10.1007/s003359900033

Vaiman, D., Schibler, L., Oustry-Vaiman, A., Pailhoux, E., Goldammer, T., Stevanovic, M., et al. (1999). High-resolution human/goat comparative map of the goat polled/intersex syndrome (PIS): the human homologue is contained in a human YAC from HSA3q23. *Genomics* 56, 31–39. doi: 10.1006/geno.1998.5691

Wang, X., Qiao, Y., Asangani, I. A., Ateeq, B., Poliakov, A., Cieślik, M., et al. (2017a). Development of peptidomimetic inhibitors of the ERG gene fusion product in prostate cancer. *Cancer Cell* 31, 532–548.e7. doi: 10.1016/j.ccell.2017.02.017

Wang, X., Zheng, Z., Cai, Y., Chen, T., Li, C., Fu, W., et al. (2017b). CNVcaller: highly efficient and widely applicable software for detecting copy number variations in large populations. *GigaScience* 6:gix115. doi: 10.1093/gigascience/gix115

Wang, Y. Q., and Zhang, X. (1993). Study on cytogenetics of intersex livestock. *J. Zhengzhou Pastoral Sci.* 13, 20–23.

Warr, N., Siggers, P., Carré G.-A., Wells, S., and Greenfield, A. (2016). Genetic analyses reveal functions for MAP2K3 and MAP2K6 in mouse testis determination. *Biol. Reprod.* 94, 1–7. doi: 10.1095/biolreprod.115.138057

Yang, B., Jia, L., Zhao, D., Meng, L., Liu, X., Zhang, Y., et al. (2012). Detection and application of PIS genetic deficiency gene in dairy goat. *Hereditas* 34, 895–900. doi: 10.3724/SP.J.1005.2012.00895

Yang, S., Han, H., Li, J., Zhang, Y., Zhao, J., Wei, H., et al. (2020). Transcriptomic analysis of gene expression in normal goat ovary and intersex goat gonad. *Reprod. Domestic Anim.* doi: 10.1111/rda.13844. [Epub ahead of print].

Yardimci, G. G., Ozadam, H., Sauria, M. E., Ursu, O., Yan, K.-K., Yang, T., et al. (2019). Measuring the reproducibility and quality of Hi-C data. *Genome Biol.* 20, 1–19. doi: 10.1186/s13059-019-1658-7

Yin, Z., Ke, X., Li, Z., Chen, J., Gao, X., and Huang, L. (2017). Unconventional recombination in the mating type locus of heterothallic apple canker pathogen Valsa mali. *G3 Genes Genomes Genet.* 7, 1259–1265. doi: 10.1534/g3.116.037853

Yu, A., Zhou, R., Xia, B., Dang, W., Yang, Z., and Chen, X. (2020). NAMPT maintains mitochondria content via NRF2-PPARα/AMPKα pathway to promote cell survival under oxidative stress. *Cell. Signal* 66:109496. doi: 10.1016/j.cellsig.2019.109496

Yuan, J., Liu, W., Karvar, S., Baker, S. S., He, W., Baker, R. D., et al. (2015). Potassium channel KCNJ15 is required for histamine-stimulated gastric acid secretion. *Am. J. Physiol. Cell Physiol.* 309, C264–C70. doi: 10.1152/ajpcell.00012.2015

Zhan, T. T. Y., Lou, M., Liu, J., Wang, Y., and Zhao, X. (1994). The genetic mechanism of intersexuality in milk goats of Saanen breed of Xinong. *Yi Chuan Xue Bao* 21, 356–361.

Zhang, S., Cao, X., Li, Y., Wang, K., Yuan, M., and Lan, X. (2019). Detection of polled intersex syndrome (PIS) and its effect on phenotypic traits in goats. *Anim. Biotechnol.* 31, 561–565. doi: 10.1080/10495398.2019.1625782

Zhang, Z., Chen, F., Li, S., Guo, H., Xi, H., Deng, J., et al. (2020). ERG the modulates Warburg effect and tumor progression in cervical cancer. *Biochem. Biophys. Res. Commun.* 522, 191–197. doi: 10.1016/j.bbrc.2019.11.079

Zheng, W., Nagaraju, G., Liu, Z., and Liu, K. (2012). Functional roles of the phosphatidylinositol 3-kinases (PI3Ks) signaling in the mammalian ovary. *Mol. Cell. Endocrinol.* 356, 24–30. doi: 10.1016/j.mce.2011.05.027

Zhou, X., Chen, Y., Mok, K. Y., Zhao, Q., Chen, K., Chen, Y., et al. (2018). Identification of genetic risk factors in the Chinese population implicates a role of immune system in Alzheimer's disease pathogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 115, 1697–1706. doi: 10.1073/pnas.1715554115

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.

Copyright © 2021 E, Zhou, Zheng, Yang, Li, Li, Zhou, Nai, Jiang, Zhang, Hong, Ma, Chu, Gao, Zhao, Duan, He, Na, Han, Zeng, Jiang 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

ORIGINAL RESEARCH
published: 08 March 2021
doi: 10.3389/fgene.2021.650305

Generation of a Biobank From Two Adult Thoroughbred Stallions for the Functional Annotation of Animal Genomes Initiative

Callum G. Donnelly[1], Rebecca R. Bellone[1,2], Erin N. Hales[3], Annee Nguyen[1], Scott A. Katzman[4], Ghislaine A. Dujovne[1], Kelly E. Knickelbein[4], Felipe Avila[2], Ted S. Kalbfleisch[5], Elena Giulotto[6], Nicole B. Kingsley[2], Jocelyn Tanaka[2], Elizabeth Esdaile[2], Sichong Peng[1], Anna Dahlgren[1], Anna Fuller[7], Michael J. Mienaltowski[8], Terje Raudsepp[9], Verena K. Affolter[10], Jessica L. Petersen[7] and Carrie J. Finno[1]

[1] Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, Davis, CA, United States, [2] Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California, Davis, Davis, CA, United States, [3] Morris Animal Foundation, Denver, CO, United States, [4] Department of Surgical and Radiological Sciences, School of Veterinary Medicine, University of California, Davis, Davis, CA, United States, [5] Gluck Equine Research Center, University of Kentucky, Lexington, KY, United States, [6] Department of Biology and Biotechnology, University of Pavia, Pavia, Italy, [7] Department of Animal Science, University of Nebraska-Lincoln, Lincoln, NE, United States, [8] Department of Animal Science, College of Agricultural and Environmental Sciences, University of California, Davis, Davis, CA, United States, [9] Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, United States, [10] Department of Pathology, Microbiology, and Immunology, School of Veterinary Medicine, University of California, Davis, Davis, CA, United States

OPEN ACCESS

Edited by:
Hans Cheng,
Agricultural Research Service (USDA), United States

Reviewed by:
Tosso Leeb,
University of Bern, Switzerland
Patrick Venta,
Michigan State University, United States

***Correspondence:**
Carrie J. Finno
cjfinno@ucdavis.edu

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

Received: 06 January 2021
Accepted: 15 February 2021
Published: 08 March 2021

Citation:
Donnelly CG, Bellone RR, Hales EN, Nguyen A, Katzman SA, Dujovne GA, Knickelbein KE, Avila F, Kalbfleisch TS, Giulotto E, Kingsley NB, Tanaka J, Esdaile E, Peng S, Dahlgren A, Fuller A, Mienaltowski MJ, Raudsepp T, Affolter VK, Petersen JL and Finno CJ (2021) Generation of a Biobank From Two Adult Thoroughbred Stallions for the Functional Annotation of Animal Genomes Initiative. Front. Genet. 12:650305. doi: 10.3389/fgene.2021.650305

Following the successful creation of a biobank from two adult Thoroughbred mares, this study aimed to recapitulate sample collection in two adult Thoroughbred stallions as part of the Functional Annotation of the Animal Genome (FAANG) initiative. Both stallions underwent thorough physical, lameness, neurologic, and ophthalmic (including electroretinography) examinations prior to humane euthanasia. Epididymal sperm was recovered from both stallions immediately postmortem and cryopreserved. Aseptically collected full thickness skin biopsies were used to isolate, culture and cryopreserve dermal fibroblasts. Serum, plasma, cerebrospinal fluid, urine, and gastrointestinal content from various locations were collected and cryopreserved. Under guidance of a board-certified veterinary anatomic pathologist, 102 representative tissue samples were collected from both horses. Whole tissue samples were flash-frozen and prioritized tissues had nuclei isolated and cryopreserved. Spatially contemporaneous samples of each tissue were submitted for histologic examination. Antemortem and gross pathologic examination revealed mild abnormalities in both stallions. One stallion (ECA_UCD_AH3) had unilateral thoracic limb lameness and bilateral chorioretinal scars. The second stallion (ECA_UCD_AH4) had subtle symmetrical pelvic limb ataxia, symmetrical prostatomegally, and moderate gastrointestinal nematodiasis. DNA from each was whole-genome sequenced and genotyped using the GGP Equine 70K SNP array. The genomic resources and banked biological samples from these animals

augments the existing resource available to the equine genomics community. Importantly we may now improve the resolution of tissue-specific gene regulation as affected by sex, as well as add sex-specific tissues and gametes.

Keywords: FAANG, horse, male, stallion, biobank

INTRODUCTION

Sex bias in animal model-based biomedical research has been a focus area for the National Institutes for Health (NIH) (Will et al., 2017). While the influence of sex on experimental outcomes is undeniable, the adoption of sex as a consideration for experimental design has been slow (Zucker and Beery, 2010). Genomic regulation is particularly labile to the effect of sex (Lee, 2018). With the intent of cataloging the functional and regulatory elements of economically significant animal species, there is need for the Functional Annotation of Animal Genome (FAANG) project to diligently account for sex differences. This is of importance to the equine industry in which sex of animal is an important component of management and use, such as predominance of females in polo competition and males in show jumping (Fenner et al., 2019). The increased resolution of tissue specific genomic, transcriptional, and epigenomic data will be of significant benefit to the progress of equine genomics. In particular, the ability to more precisely characterize regulation of genomic regions associated with traits pertinent to equine health and performance are made increasingly more possible (Kingsley et al., 2019). The objective of the current study was to augment the existing equine biobank with male samples. We report here the addition of two male Thoroughbred horses to the biobank of two female Thoroughbred horses initially established in 2016.

MATERIALS AND METHODS

Animals

Two Thoroughbred stallions (aged 3 and 4 years; ECA_UCD_AH3 and ECA_UCD_AH4, respectively) were donated for this project. ECA_UCD_AH3 had been in racing training and sustained a career ending musculoskeletal injury prior to donation. ECA_UCD_AH4 had not engaged in racing training, and is a son of the reference genome donor Twilight. Approval for all protocols was granted by the UC Davis Institutional Animal Care and Use Committee (Protocol #21033).

Clinical Assessments

Complete physical examination (including complete blood count and serum biochemistry) was performed and interpreted by two board-certified internists (CJF and CGD) for both stallions. Complete neurologic and lameness examinations were performed by relevant recognized experts (CJF and SAK, respectively) for both stallions within 24 h prior to euthanasia. Complete ophthalmic examination of the anterior and posterior segment (performed by KEK) as well as photopic and scotopic electroretinography (ERG) was performed on both stallions within 48 h prior to euthanasia.

Clinical Sample Collection

Serum and plasma samples were collected immediately prior to euthanasia *via* an indwelling intravenous cannula. Heparin and EDTA whole blood was centrifuged at 2,000× g for 10 min at 4°C. Plasma was harvested, flash frozen in liquid nitrogen and stored at −80°C. Whole blood in plain tubes was allowed to clot for 30 min at room temperature, with serum collected in the same fashion as plasma.

Ejaculated sperm and seminal plasma were collected from ECA_UCD_AH3 at the time of ophthalmic examination. Ejaculated sperm was unable to be collected from ECA_UCD_AH4. Seminal plasma was retrieved by un-cushioned centrifugation at 2,000× g for 15 min. The supernatant was separated and centrifuged again before being passed through a 20 μm filter. The supernatant and sperm pellet were flash frozen in liquid nitrogen and stored at −80°C.

Immediately following euthanasia, the left testicle, epididymis, and distal ductus deferens were removed from each horse. The epididymes were removed and incised sharply transversely at the cauda epididymis. Commercial semen extender was flushed retrograde *via* the remaining ductus deferens. Recovered spermatozoa were frozen in 0.5 mL polyethylene straws by controlled rate freezing protocol and stored in liquid nitrogen, at 323 and 455.5 million sperm/mL for ECA_UCD_AH3 and ECA_UCD_AH4, respectively.

Synovial fluid was collected aseptically (by centesis) from the left middle carpal, radio-carpal joint and the left medial compartment of the femoro-tibial joint from ECA_UCD_AH3 and the right middle carpal, radio-carpal joint and left medial compartment of the femoro-tibial joint from ECA_UCD_AH4 immediately following euthanasia. A veterinary clinical pathologist cytologically evaluated an aliquot of each sample within 2 h of collection. The remaining sample was centrifuged at 2,000× g for 10 min at 4°C, and the supernatant was flash frozen in liquid nitrogen and stored at −80°C.

Cerebrospinal fluid was collected aseptically *via* atlanto-occipital centesis. A veterinary clinical pathologist cytologically evaluated an aliquot of each sample within 2 h of collection. The remaining sample was then processed in the same manner as synovial fluid.

Peripheral Blood Mononuclear Cell Collection

Whole blood was collected in EDTA tubes 24 h prior to humane euthanasia. Using density gradient centrifugation, peripheral blood mononuclear cells (PBMCs) were harvested as previously described (Hida et al., 2002). Briefly, whole blood was overlayed on a double gradient of 1,077 and 1,119 histopaque media (Sigma-Aldrich). Then, samples were centrifuged at 700 g for

30 min with the enriched PBMC layer carefully removed, flash frozen in liquid nitrogen and stored at −80°C.

Tissue Specific Sampling

Full thickness skin biopsies were aseptically collected from the area overlying the left gluteal muscles for dermal fibroblast isolation and cultured as previously described (Raimondi et al., 2011). Briefly, biopsies were washed in ice cold PBS with the addition of penicillin and streptomycin. Fragments of the dermis (2–3 mm^2) were placed into a 24-well tissue-culture-treated plate and covered in complete media (Dulbecco's Minimum Essential Medium, 20% fetal bovine serum, 2× non-essential amino acids, 2 mM L-glutamine, 2× penicillin/streptomycin, 2 μg/ml amphotericin B, and 1 μg/ml fluconazole). At confluence, cells were trypsinized, counted, and seeded in a 12-well plate. Cells were passaged and frozen at passage three and four in a DMSO-based cryoprotectant media and stored in liquid nitrogen (Raimondi et al., 2011). DNA isolated from whole blood and from cultured fibroblast cells of each horse were genotyped and compared for 15 standard genetic markers (fourteen microsatellites and one sex link marker amelogenin) routinely tested for use in identify and parentage testing services as the UC Davis Veterinary Genetics Laboratory.

A total of 102 issue samples were taken from all body systems (**Supplementary Table 1**). Sample stations were arranged in teams, with a veterinarian overseeing each team to ensure appropriate tissue identification. Additionally, all tissues were examined by a veterinary anatomic pathologist (VKA) for gross abnormalities prior to collection (**Supplementary Table 2**). Tissue samples were collected as previously described (Burns et al., 2018). Briefly, samples from the most representative portion of each tissue were collected and preserved in 10% buffered formalin for histopathology. Samples for nuclei preparation and tissue banking were taken from sites immediately adjacent (proximal and distal) to the representative histopathology sections and flash frozen. Nuclei isolation was performed as similarly to that described for the mare biobank, with the exceptions of testis replacing ovary and C6 spinal cord instead of T1 (Burns et al., 2018).

Genotyping and Whole-Genome Sequencing

Genomic DNA was isolated from whole blood using a previously validated technique (Burns et al., 2018). Whole-genome sequencing using an Illumina NovaSeq (150 bp paired-end reads) platform was performed at Admera Health, LLC (South Plainfield, NJ), targeting 35X coverage for each horse. Library preparation was performed using the KAPA library quantification kit for Illumina (Roche Holding AG, Basel, Switzerland). SNP genotyping was performed using the GGP Equine 70K SNP bead chip array (Neogen GeneSeek, Lincoln, NE). Genotype data were merged with those collected from the prior sampling effort of Thoroughbred mares (Burns et al., 2018). Marker positions were translated from EquCab2 to EquCab3 using the NCBI Genome Remapping Service (https://www.ncbi.nlm.nih.gov/genome/tools/remap), removing markers that were not mapped to any of the 31 autosomes. Using SNP data, runs of homozygosity (ROH), a measure of diversity, was quantified in DetectRuns (https://github.com/bioinformatics-ptp/detectRUNS/tree/master/detectRUNS) using the method of Marras et al. (2015). Analysis parameters required a minimum of 15 SNPs, maxOpp Run = 1, max MIssRun = 1, maxGap = 10,000,000, minLengthBps = 100,000. Whole-genome sequence data of the stallions and mares were processed and variants called according to the pipeline outlined in Sieck et al. (2020), with the exception of mapping to the EquCab3 reference genome (Burns et al., 2018; Sieck et al., 2020). Bi-allelic autosomal SNPs were extracted from the resulting vcf for ROH analysis in the same manner as used for the SNP data.

Additionally, in an effort to provide extensive phenotyping and genotyping for future studies on pigmentation biology, both stallions were genotyped for coat color loci using the commercially available horse full coat color and white pattern panel (https://vgl.ucdavis.edu/panel/full-coat-color-pattern-panel).

Pedigree Analysis

Individual inbreeding coefficients (F) were calculated from five-generation pedigrees of both stallions and the previously sampled mares (Burns et al., 2018) using Pedigraph (Garbe and Da, 2008).

Karyotyping

Sodium heparin stabilized whole blood samples were collected 7 days prior to euthanasia for karyotype characterization. Karyotypes were generated for each horse using the previously validated Pokeweed-stimulated lymphocyte culture (Raudsepp et al., 2010). Thirty metaphase cells from ECA_UCD_AH3 and ECA_UCD_AH4 were captured for analysis. Genomic DNA was isolated from lymphocyte cultures for PCR detection of the sex determining region Y *(SRY)* and androgen receptor *(AR)* genes, as preciously described (Raudsepp et al., 2004).

RESULTS

Clinical Assessments

ECA_UCD_AH3 had several mild abnormalities apparent on clinical examination, including a grade III/V left thoracic limb lameness [AAEP scale (Baxter, 2020)] and chorioretinal scarring of both eyes. Scoptopic ERG data for both horses is provided in **Supplementary Table 3**. No other abnormalities were detected antemortem for ECA_UCD_AH3. ECA_UCD_AH4 did not have clinically detectable lameness, however, a mild grade I/V [Modified Mayhew scale (Furr and Reed, 2015)] symmetrical pelvic limb ataxia was detected with a presumptive neuroanatomical localization to the cervical spinal cord. No other abnormalities were detected for ECA_UCD_AH4 on clinical examination. No abnormalities were detected on complete blood count or serum biochemistry for either stallion.

Clinical Sample Collection

Serum, plasma (collected in EDTA and heparin tubes), buffy coat and urine were collected from both stallions. Seminal plasma and ejaculated spermatozoa were obtained for ECA_UCD_AH3 only. Epididymal recovered spermatozoa from both stallions

were successfully isolated and cryopreserved with adequate post thaw kinematic parameters (**Supplementary Table 4**). Synovial and cerebrospinal fluid samples were cytologically normal (**Supplementary Tables 5**, **6**). PBMCs were successfully collected from both stallions.

Genotyping

Achieved coverage of the whole-genome sequence data was 38.7 and 37.5X for ECA_UCD_AH3 and ECA_UCD_AH4, respectively. Relative to the reference genome, between 5.4 (ECA_UCD_AH3) and 3.8 (ECA_UCD_AH4) million variants were observed. Data are publically available through the FAANG consortium. Combining the 70K genotype data of the stallions to that from the two mares and removing variants present on sex chromosome data and on unmapped contigs, 59,823 variants remained. The count of ROH per horse was similar (498–503) for all four horses, except ECA_UCD_AH4 (280). However, ECA_UCD_AH4 had a greater number of long (>8 Mb) ROH than all other horses (**Supplementary Table 7**). Considering SNP data from WGS, between 1,822 (ECA_UCD AH1) and 2,867 (ECA_UCD AH4) ROH were observed. Five-generation pedigree-based inbreeding estimates ranged from 0 (ECA_UCD AH1) to 0.344 (ECA_UCD AH4).

Genetic profiles from cultured fibroblast cells matched to that of the expected horse (15/15 markers matched for each horse, presented in **Supplementary Table 8**).

ECA_UCD_AH3 is genetically defined as a bay horse (E/e A/a) with one copy of dominant white 20 (N/W20), while ECA_UCD_AH4 was gray with a bay base coat (E/e A/A) before graying out and was homozygous for dominant white 20 (W20/W20) (all coat color genotypes are presented in **Supplementary Table 9**).

Karyotype

Both stallions had a normal 64 XY male karyotype with no apparent structural or numerical chromosomal abnormalities. Both stallions were positive for the *SRY* and *AR* genes.

Tissue Specific Sampling

Sample collection commenced 30 min following euthanasia and was concluded within 3 h for all tissue samples. Samples from 102 tissues were collected and flash frozen (**Supplementary Table 1**). Sixteen of these tissues also had nuclei isolated and cryopreserved (**Supplementary Table 1**).

Pathology

Gross examination of ECA_UCD_AH3 was without abnormality, except for a 15 × 2 × 1.3 cm firm area at the caudal margin of the left caudal lung lobe. This area was therefore avoided for sampling. Histologic examination of the area identified locally extensive fibrosis, Type II pneumocyte proliferation and alveolar histiocytosis, which is consistent with a previous pulmonary inflammatory insult. Mild lymphocytic inflammation was noted at several sections of the gastrointestinal tract, but were considered incidental. Gross examination of ECA_UCD_AH4 revealed marked gastrointestinal nematodiasis throughout the large and small intestinal segments. Gastrointestinal segments had marked lymphocytic and eosinophilic infiltrates consistent with gastrointestinal parasitism. ECA_UCD_AH4 had marked prostatomegally grossly, but was without abnormality on histopathology. There were no gross or histologic findings of the central nervous system of ECA_UCD_AH4 that could explain the mild sensory ataxia. There were no other significant histologic abnormalities. Summary of all pathologic findings are summarized in **Supplementary Table 2**.

DISCUSSION

We report the successful completion of the male equine biobank for the FAANG consortium. This project has added 102 tissues from two stallions in addition to seven body fluids, one cell line and spermatozoa to the biobank for use by the equine research community. This study recapitulated the guidelines of the original equine tissue collection, demonstrating that this method is repeatable and appropriate for this type of endeavor.

Stringent phenotype information both ante- and postmortem are a critical feature of this biobank and lends strength to the interpretation of results of subsequent studies using these tissues. Similarly, it also establishes the boundaries for the interpretation of data derived from this biobank. This is especially true for the gastrointestinal tissues that had lymphocytic and/or eosinophilic infiltrates. Previous tissue collection from both mares also had a similar infiltrate pattern in the gastrointestinal sections to the stallions described here (Burns et al., 2018). While this makes the direct comparison between male and female samples more consistent, it underscores the necessity to perform detailed phenotypic characterization of the tissues. The available phenotype information was expanded in this data set with the addition of ERG measurements. Further, this additional information enhances the ability to select tissue with the most normal function or, at minimum, document pathology that may affect future assays. In this case, the left retina from ECA_UCD_AH3 with an abnormal waveform was not used for the tissue bank, with the normal right retina retained.

The available whole-genome sequencing and SNP genotyping data from these horses to accompany the biobank and pending tissue-specific regulatory information will assist in future genomic investigations. In this sampling effort, the availability of a stallion closely related to the mare from which the reference genome is based (Twilight) is expected to provide high accuracy mapping and allow for more in depth investigations of achieved vs. predicted inbreeding. The intentional inbreeding that resulted in ECA_UCD_AH4 (by breeding record) was corroborated by the frequency and relatively large ROH as identified in the analysis of the SNP genotype data. This also highlights the limitation of this biobank, as it is restricted to a single breed. Future investigation may need to include more diverse biologic replicates to account for between breed variations in gene regulation (Li et al., 2012).

While the equine FAANG project has not progressed to the level of the ENCODE project in humans, important data have already been generated from the equine biobank

(ENCODE Project Consortium, 2012). Histone modifications as analyzed by chromatin immunoprecipitation sequencing have been established for eight tissues from the original two mares (Kingsley et al., 2019). The addition of repressive and enhancing epigenetic marks to the catalog of equine genome regulation is of great benefit to the equine research community. Sex-biased gene regulation is strongly intertwined with epigenetic regulation, especially as it relates to histone modifications (Tsai et al., 2009). This reinforces the need for the equine FAANG project to expand the existing biobank with male samples. This undertaking was no small feat and required the sacrifice of two additional animals. However, the benefit to future equine genomic studies is immense.

Although this study concludes the tissue collection arm of the FAANG project, the work to annotate the equine genome continues in earnest. This addition to the biobank expands the list of tissues available to the equine research community. Groups with specific interest in certain tissues are encouraged to contact the corresponding author for availability, as the "adopt-a-tissue" program successfully used with the original collection is also available for these male samples.

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 at: https://www.ebi.ac.uk/ena, PRJEB42407.

ETHICS STATEMENT

The animal study was reviewed and approved by UC Davis Institutional Animal Care and Use Committee.

AUTHOR CONTRIBUTIONS

Clinical examinations were performed by CF, SK, KK, and CD. Semen evaluation and cryopreservation performed by GD. Karyotyping performed by TR. Tissue collection performed by EH, AN, KK, NK, JT, EE, SP, AD, AF, JP, and CD. All histopathoogy reviewed by VA. Nuclei preparation performed by RB and FA. Cell culture and cryopreservation performed by MM. Genetic analysis performed by JP and RB. All authors contributed to manuscript preparation.

FUNDING

Funding was provided by the Grayson Jockey Club Foundation, USDA NRSP-8 and the UC Davis Center for Equine Health. Support for CF was provided by the National Institutes of Health (NIH) (L40 TR001136). None of the funding agencies had any role in the design of the study, analysis, interpretation of the data, or writing of the manuscript. Salary support for SP was provided by the Ann T. Bowling Fellowship at the UC Davis Veterinary Genetics Laboratory.

ACKNOWLEDGMENTS

The authors would like to thank all of the technicians at the Center for Equine Health and Anatomic Pathology.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.650305/full#supplementary-material

REFERENCES

Baxter, G. M. (ed.). (2020). *Adams and Stashak's Lameness in Horses*. Hoboken, NJ: John Wiley & Sons.

Burns, E. N., Bordbari, M. H., Mienaltowski, M. J., Affolter, V. K., Barro, M. V., Gianino, F., et al. (2018). Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. *Anim. Genet.* 49, 564–570. doi: 10.1111/age.12717

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

Fenner, K., Caspar, G., Hyde, M., Henshall, C., Dhand, N., Probyn-Rapsey, F., et al. (2019). It's all about the sex, or is it? Humans, horses and temperament. *PLoS ONE* 14:e0216699. doi: 10.1371/journal.pone.0216699

Furr, M., and Reed, S. (eds.). (2015). *Equine Neurology*. Hoboken, NJ: John Wiley & Sons.

Garbe, J. R., and Da, Y. (2008). *User Manual Version 2.4*. Minneapolis, MN: Department of Animal Science, University of Minnesota.

Hida, N., Maeda, Y., Katagiri, K., Takasu, H., Harada, M., and Itoh, K. (2002). A simple culture protocol to detect peptide-specific cytotoxic T lymphocyte precursors in the circulation. *Cancer Immunol. Immunother.* 51, 219–228. doi: 10.1007/s00262-002-0273-7

Kingsley, N. B., Kern, C., Creppe, C., Hales, E. N., Zhou, H., Kalbfleisch, T. S., et al. (2019). Functionally annotating regulatory elements in the equine genome using histone mark ChIP-Seq. *Genes* 11:3. doi: 10.3390/genes11010003

Lee, S. K. (2018). Sex as an important biological variable in biomedical research. *BMB Rep.* 51, 167–173. doi: 10.5483/BMBRep.2018.51.4.034

Li, M., Wu, H., Luo, Z., Xia, Y., Guan, J., Wang, T., et al. (2012). An atlas of DNA methylomes in porcine adipose and muscle tissues. *Nat. Commun.* 3, 1–11. doi: 10.1038/ncomms1854

Marras, G., Gaspa, G., Sorbolini, S., Dimauro, C., Ajmone-Marsan, P., Valentini, A., et al. (2015). Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. *Anim. Genet.* 46, 110–121. doi: 10.1111/age.12259

Raimondi, E., Piras, F. M., Nergadze, S. G., Di Meo, G. P., Ruiz-Herrera, A., Ponsa, M., et al. (2011). Polymorphic organization of constitutive heterochromatin in Equus asinus (2n = 62) chromosome 1. *Hereditas* 148, 110–113. doi: 10.1111/j.1601-5223.2011.02218.x

Raudsepp, T., Durkin, K., Lear, T. L., Das, P. J., Avila, F., Kachroo, P., et al. (2010). Molecular heterogeneity of XY sex reversal in horses. *Anim. Genet.* 41(Suppl. 2), 41–52. doi: 10.1111/j.1365-2052.2010.02101.x

Raudsepp, T., Lee, E. J., Kata, S. R., Brinkmeyer, C., Mickelson, J. R., Skow, L. C., et al. (2004). Exceptional conservation of horse-human gene order on X chromosome revealed by highresolution radiation hybrid mapping. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2386–2391. doi: 10.1073/pnas.0308513100

Sieck, R. L., Fuller, A. M., Bedwell, P. S., Ward, J. A., Sanders, S. K., Xiang, S. H., et al. (2020). Mandibulofacial dysostosis attributed to a recessive mutation of CYP26C1 in hereford cattle. *Genes* 11:1246. doi: 10.3390/genes11111246

Tsai, H. W., Grant, P. A., and Rissman, E. F. (2009). Sex differences in histone modifications in the neonatal mouse brain. *Epigenetics* 4, 47–53. doi: 10.4161/epi.4.1.7288

Will, T. R., Proaño, S. B., Thomas, A. M., Kunz, L. M., Thompson, K. C., Ginnari, L. A., et al. (2017). Problems and progress regarding sex bias

and omission in neuroscience research. *eNeuro* 4:ENEURO.0278-17.2017. doi: 10.1523/ENEURO.0278-17.2017

Zucker, I., and Beery, A. (2010). Males still dominate animal studies. *Nature* 465:690. doi: 10.1038/465690a

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.

Copyright © 2021 Donnelly, Bellone, Hales, Nguyen, Katzman, Dujovne, Knickelbein, Avila, Kalbfleisch, Giulotto, Kingsley, Tanaka, Esdaile, Peng, Dahlgren, Fuller, Mienaltowski, Raudsepp, Affolter, Petersen and Finno. 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

DATA REPORT
published: 26 March 2021
doi: 10.3389/fgene.2021.649959

"Adopt-a-Tissue" Initiative Advances Efforts to Identify Tissue-Specific Histone Marks in the Mare

*N. B. Kingsley[1,2], Natasha A. Hamilton[3], Gabriella Lindgren[4,5], Ludovic Orlando[6], Ernie Bailey[7], Samantha Brooks[8], Molly McCue[9], T. S. Kalbfleisch[7], James N. MacLeod[7], Jessica L. Petersen[10], Carrie J. Finno[2] and Rebecca R. Bellone[1,2]**

[1] Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California, Davis, Davis, CA, United States, [2] Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, Davis, CA, United States, [3] Faculty of Science, School of Life and Environmental Science, University of Sydney, Camperdown, NSW, Australia, [4] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden, [5] Livestock Genetics, Department of Biosystems, KU Leuven, Leuven, Belgium, [6] Centre d'Anthropobiologie et Génomique de Toulouse (CAGT), Faculté de Médecine Purpan, Université Toulouse III-Paul Sabatier, Toulouse, France, [7] Maxwell H. Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Lexington, KY, United States, [8] Department of Animal Sciences, University of Florida, Gainesville, FL, United States, [9] Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, St. Paul, MN, United States, [10] Department of Animal Science, University of Nebraska-Lincoln, Lincoln, NE, United States*

Keywords: genome, annotation, epigenetics, horse, chromatin, consortium, collaboration, regulation

OPEN ACCESS

Edited by:
Hans Cheng, United States Department of Agriculture, United States

Reviewed by:
Cong-jun Li, Agricultural Research Service (USDA), United States
Herve Acloque, Institut National de Recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), France

****Correspondence:***
Rebecca R. Bellone
rbellone@ucdavis.edu

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

Received: *06 January 2021*
Accepted: *01 March 2021*
Published: *26 March 2021*

Citation:
Kingsley NB, Hamilton NA, Lindgren G, Orlando L, Bailey E, Brooks S, McCue M, Kalbfleisch TS, MacLeod JN, Petersen JL, Finno CJ and Bellone RR (2021) "Adopt-a-Tissue" Initiative Advances Efforts to Identify Tissue-Specific Histone Marks in the Mare. Front. Genet. 12:649959. doi: 10.3389/fgene.2021.649959

INTRODUCTION

The equine genetics and genomics research community has a long history of synergistic collaborations for developing tools and resources to advance equine biology. Starting in 1995 with the first International Equine Gene Mapping Workshop supported by the Dorothy Russell Havemeyer Foundation Inc. (Bailey, 2010), researchers collaborated to build comprehensive equine linkage maps (Guérin et al., 1999, 2003; Penedo et al., 2005; Swinburne et al., 2006), radiation hybrid and comparative maps (Caetano et al., 1999; Chowdhary et al., 2002), physical marker and BAC contig maps (Raudsepp et al., 2004, 2008; Leeb et al., 2006), reference genomes for the horse (Wade et al., 2009; Kalbfleisch et al., 2018), and genotyping arrays to economically map and study traits of interest for horse owners and breeders (McCue et al., 2012; McCoy and McCue, 2014; Schaefer et al., 2017). Continuing the legacy of community-based advancements, a new collective effort began in 2015 to functionally annotate DNA elements in the horse as part of the international Functional Annotation of ANimal Genomes (FAANG) Consortium (Andersson et al., 2015; Tuggle et al., 2016; Burns et al., 2018).

Reminiscent of the ENCODE project in humans and mice (Dunham et al., 2012), the ultimate goal of the FAANG consortium is to annotate the major functional elements in the genomes of domesticated animal species (Andersson et al., 2015). In particular, four histone modifications were chosen by the consortium to characterize the genomic locations of enhancers (H3K4me1), promoters and transcription start sites (H3K4me3), open chromatin with active regulatory elements (H3K27ac), and facultative heterochromatin with inaccessible or repressed regulatory elements (H3K27me3) (Andersson et al., 2015; Giuffra and Tuggle, 2019). The initial equine FAANG efforts identified putative regulatory regions in eight prioritized tissues of interest (TOI) by performing Chromatin Immuno-Precipitation Sequencing (ChIP-Seq) for the four target histone marks (Kingsley et al., 2020). In that investigation, more than one million putative regulatory sites were characterized across the equine genome. With more than 80 tissues, cell lines, and body fluids stored in the equine biobank (Burns et al., 2018), further opportunities to expand the scope of the annotation work exist. To leverage the benefits of the biobank, a collaborative sponsorship

program titled "Adopt-a-Tissue" was created to enable researchers from across the globe to select and support annotation of a tissue by the equine FAANG group. Through this effort, four additional "Adopted" tissues— spleen, metacarpal 3 (MC3), sesamoid, and full thickness skin— were assayed by histone mark ChIP-Seq to expand the tissue-specific annotation resources for the entire equine research community.

METHODS

All ChIP-Seq assays were performed by Diagenode ChIP-Seq Profiling Service (Diagenode, Cat# G02010000, Liège, Belgium). Summarized experimental procedures are available in more detail at the FAANG FTP portal hosted by EBI (ftp://ftp.faang.ebi.ac.uk/faang/ftp/protocols/assays/ and ftp://ftp.faang.ebi.ac.uk/faang/ftp/protocols/experiments/). Spleen samples were processed following the assay procedures outlined in UCD_SOP_ChIP-Seq_for_Histone_Marks_20191101.pdf.
Skin and both bone tissues were processed following the experimental protocols outlined in UCD_SOP_ChIP-seq_for_Histone_Marks_Skin_20201218.pdf and UCD_SOP_ChIP-seq_for_Histone_Marks_Bone_20201218.pdf, respectively. "Adopted" tissues, as summarized in **Supplementary Table 1**, were collected from two Thoroughbred mares (denoted as ECA_UCD_AH1 for SAMEA104728862 and ECA_UCD_AH2 for SAMEA104728877) as part of the FAANG equine biobank (Burns et al., 2018) following protocols approved by the University of California, Davis Institutional Animal Care and Use Committee (Protocol #19037).

Chromatin was isolated from the two bone tissues using the TrueMicro ChIP-Seq kit (Diagenode Cat# C01010140) and from spleen and skin using the iDeal ChIP-Seq kit for Histones (Diagenode Cat# C01010059). Starting amounts for each replicate varied by tissue with ∼100 mg for spleen, 375–770 mg for MC3, 445–650 mg for sesamoid, and ∼125 mg for skin. After homogenization, fixed samples were sheared with the Bioruptor® Pico (Diagenode Cat# B01060001) for 12 (spleen), 10–12 (MC3 and sesamoid), and 8 (skin) cycles of 30 s on and 30 s off. The amount of chromatin yield and thus chromatin per IP varied by tissue. Spleen and skin had the greatest amounts (1.5 μg and 600 ng, respectively) per IP and MC3 and sesamoid had the least (350 ng each). The following antibody concentrations were used for MC3, sesamoid, and skin: 0.5 μg for H3K4me1, 0.5 μg for H3K4me3, 1 μg for H3K27ac, and 1 μg for H3K27me3. To account for the greater amount of chromatin from spleen, twice the amount of antibody was used for each mark compared to the other three tissues. For all tissues, 10% of the total chromatin from each replicate was saved for the input.

Libraries were prepared with the IP-Star® Compact Automated System (Diagenode Cat# B03000002) using the MicroPlex Library Preparation Kit v2 (Diagenode Cat# C05010013). Spleen, MC3, and sesamoid were sequenced as 50 base pair single-end (SE) reads on the HiSeq 4000 platform (Illumina, San Diego, CA, USA). For these tissues, the broad mark (H3K27me3) was sequenced to a minimum of 50M raw reads while the remaining marks (H3Kme1, H3K4me3, and H3K27ac) and the input were sequenced to a minimum depth of 30 M raw reads. Due to advancements in sequencing technology, skin tissue was sequenced as 50 base pair paired-end (PE) reads on the NovaSeq 6000 (Illumina, San Diego, CA, USA). For skin, the broad mark (H3K27me3) was sequenced to a minimum of 100 M raw fragments while the remaining marks (H3Kme1, H3K4me3, and H3K27ac) and the input were sequenced to a minimum depth of 40 M raw fragments.

Methods for analyzing SE reads followed the procedures described previously (Kingsley et al., 2020) and modifications were made to the SE analysis methods to accommodate PE data generated from skin. After trimming with Trim-Galore version 0.4.0 (Martin, 2011; Andrews et al., 2012), reads were aligned to EquCab3.0 (Kalbfleisch et al., 2018) with BWA-MEM version 0.7.9a (Li and Durbin, 2009). Alignments in BAM format were filtered using SAMtools version 1.9 (Li et al., 2009). Reads were removed if they did not map, had secondary alignments (including split hits), failed platform/vendor quality tests, were identified as optical duplicates, or had an alignment quality score <30. PE reads were also removed if the mates did not map. PCR duplicates were marked with PicardTools version 2.7.1 (Picard toolkit, 2019) and removed with SAMtools. For peak-calling, MACS2 version 2.1.1.20160309 (Zhang et al., 2008) was used to call peaks for all marks with PE data denoted by a PE flag (-f BAMPE). SICERpy version 0.1.1 was also used to call peaks for H3K27me3 as it specializes in broad peak calling (SICERpy, SICERpy, GitHub Repository; Zang et al., 2009). To use SICERpy with the PE data, the second read in each pair was removed and data were processed as SE based on recommendations from the software developers. Peak-calls were combined by identifying overlapping regions of enrichment in both biological replicates where at least one replicate was significantly enriched for a given mark. Heatmaps and quality metrics were generated using deepTools 2.4.2 (Ramírez et al., 2016), SPP 1.13 (Kharchenko et al., 2008), and custom scripts. Detailed bioinformatic workflows are available at ftp://ftp.faang.ebi.ac.uk/faang/ftp/protocols/analysis/.

QUALITY ASSESSMENT

Library Complexity

Data were assessed for library complexity with metrics established by ENCODE and endorsed by FAANG, including nonredundant fraction (NRF), PCR bottleneck coefficient 1 (PBC1), and PCR bottleneck coefficient 2 (PBC2) (Landt et al., 2012; Kingsley et al., 2020). All of the libraries prepared surpassed the quality threshold for the PBC2 metric (PBC2 > 1), however, several marks and tissues fell below the quality threshold for NRF and PBC1 (**Table 1**). For example, three of the four marks for spleen passed all library complexity measures while the H3K27me3 data from both biological replicates failed NRF and PBC1. Additionally, both replicates for sesamoid and MC3 passed all three metrics for H3K4me1 and H3K27me3 but fell below threshold for H3K4me3 and H3K27ac. All skin libraries passed NRF and PBC1 thresholds with three exceptions:

TABLE 1 | Quality metrics and peak-calling summary for each biological replicate.

Mark	Tissue	Replicate	NRF	PBC1	PBC2	NSC	RSC	JSD	FRiP	Peak calls
		Threshold:	(>0.5)	(>0.5)	(>1)	(>1.05)	(>0.8)	(>0.05)	(>0.01)	
K4me1	Spleen	AH1	0.65	0.64	2.82	1.03	0.98	0.39	0.20	84,146
K4me1	Spleen	AH2	0.74	0.74	3.80	1.04	0.96	0.39	0.25	113,447
K4me3	Spleen	AH1	0.62	0.64	2.98	2.67	1.44	0.62	0.57	28,735
K4me3	Spleen	AH2	0.57	0.60	2.67	2.53	1.34	0.62	0.57	31,198
K27ac	Spleen	AH1	0.65	0.66	2.96	1.34	1.54	0.46	0.29	50,977
K27ac	Spleen	AH2	0.70	0.70	3.40	1.31	1.46	0.45	0.29	60,281
K27me3	OriginalSpleen	AH1	0.43	0.42	1.79	1.01	0.64	0.02	0.05	1,136
K27me3	RepeatSpleen	AH1	0.13	0.28	2.94	1.03	0.33	0.20	0.02	164
K27me3	MergedSpleen	AH1	0.32	0.40	1.86	1.01	0.59	0.50	0.05	6,297
K27me3	OriginalSpleen	AH2	0.44	0.43	1.82	1.01	0.67	0.05	0.08	6,647
K27me3	RepeatSpleen	AH2	0.66	0.67	3.02	1.02	0.74	0.06	0.10	32,492
K27me3	MergedSpleen	AH2	0.53	0.55	2.40	1.01	0.83	0.06	0.11	37,629
K4me1	Sesamoid	AH1	0.63	0.63	2.64	1.02	0.63	0.31	0.08	43,397
K4me1	Sesamoid	AH2	0.84	0.85	6.46	1.01	0.49	0.21	0.00	4
K4me3	Sesamoid	AH1	0.37	0.39	1.79	2.33	1.26	0.57	0.48	19,617
K4me3	Sesamoid	AH2	0.39	0.40	1.78	1.40	1.20	0.41	0.21	16,524
K27ac	Sesamoid	AH1	0.42	0.42	1.81	1.16	1.19	0.40	0.19	34,223
K27ac	Sesamoid	AH2	0.34	0.34	1.67	1.02	0.60	0.31	0.02	5,013
K27me3	Sesamoid	AH1	0.68	0.68	3.11	1.01	0.48	0.25	0.06	1,840
K27me3	Sesamoid	AH2	0.75	0.75	3.97	1.01	0.56	0.18	0.00	0
K4me1	MC3	AH1	0.74	0.74	3.76	1.02	0.68	0.12	0.09	56,238
K4me1	MC3	AH2	0.77	0.77	4.28	1.02	0.63	0.13	0.07	47,452
K4me3	MC3	AH1	0.14	0.27	2.91	2.73	1.22	0.48	0.50	19,209
K4me3	MC3	AH2	0.32	0.34	1.71	2.56	1.20	0.48	0.50	21,339
K27ac	MC3	AH1	0.27	0.29	1.65	1.25	1.19	0.30	0.23	36,022
K27ac	MC3	AH2	0.09	0.26	4.56	1.38	0.95	0.32	0.19	16,638
K27me3	MC3	AH1	0.57	0.58	2.36	1.01	0.55	0.25	0.10	17,001
K27me3	MC3	AH2	0.65	0.65	2.88	1.01	0.51	0.22	0.08	13,790
K4me1	Skin	AH1	0.43	0.47	2.14	1.15	2.85	0.29	0.34	115,470
K4me1	Skin	AH2	0.53	0.55	2.39	1.13	3.06	0.25	0.29	109,322
K4me3	Skin	AH1	0.41	0.46	2.12	3.14	1.27	0.60	0.69	24,442
K4me3	Skin	AH2	0.32	0.40	2.12	3.19	1.30	0.60	0.68	23,584
K27ac	Skin	AH1	0.50	0.53	2.30	1.51	1.45	0.41	0.47	58,278
K27ac	Skin	AH2	0.58	0.59	2.57	1.47	1.46	0.40	0.47	57,737
K27me3	Skin	AH1	0.50	0.53	2.36	1.06	3.20	0.14	0.24	95,788
K27me3	Skin	AH2	0.50	0.54	2.40	1.06	4.09	0.11	0.21	77,151

The summary includes six quality metrics—NRF, Non-Redundant Fraction; PBC1 and PBC2, PCR Bottleneck Coefficient 1 and 2; NSC, Normalized Strand Cross-Correlation Coefficient; RSC, Relative Strand Cross-Correlation Coefficient; FRiP, Fraction of Reads in Peaks– and thresholds originally established by ENCODE, and the Jensen Shannon Distance (JSD). Samples include all of the original spleen IP, the repeated spleen IP for H3K27me3, and the merged (original + repeated) spleen IPs for H3K27me3. Peaks used to determine FRiP and peak numbers for H3K27me3 were called with SICER. All other peaks were generated with MACS2. Biological replicates are denoted as AH1 for SAMEA104728862 and AH2 for SAMEA104728877. Red highlighting indicates values below the quality thresholds.

both replicates for H3K4me3 and ECA_UCD_AH2 replicate for H3K4me1.

In addition to quality metrics, sequencing data were evaluated at several processing stages of the analysis including alignment and PCR deduplication. All datasets generated high mapping quality scores (>35) and exceeded the minimum sequencing targets as described in the methods (**Supplementary Table 2**). Skin and spleen tissues retained a high number of reads for H3K4me1, H3K4me3, and H3K27ac after alignment, filtering, and deduplication (>20 M reads per replicate). Although all three activating marks were sequenced to the same target for both bone tissues, H3K4me1 retained more than 20 M reads per replicate while H3K4me3 and H3K27ac fell below 20 M processed reads per replicate with the majority of reads removed by deduplication. More than 40 M reads remained for each H3K27me3 replicate after processing with the exception of ECA_UCD_AH2 for sesamoid.

IP Enrichment

Data were also evaluated for IP enrichment using a variety of metrics to determine signal quality. Using normalized strand cross correlation (NSC) and relative strand cross correlation (RSC) assessments established by ENCODE (Landt et al., 2012), all marks for skin tissue exceeded the minimum quality threshold (**Table 1**). Additionally, the biological replicates for H3K4me3 and H3K27ac from spleen and MC3, as well as the H3K4me3 replicates for sesamoid, passed both cross-correlation measures. Similar to the library complexity metrics, several tissues fell below the quality thresholds (NCS $>$ 1.05 and RSC $>$ 0.8) including H3K4me1 from sesamoid and MC3; H3K27ac from ECA_UCD_AH2 sesamoid; and H3K27me3 from spleen, sesamoid, and MC3. Alignments were also assessed using the Jensen Shannon distance (JSD) to compare the distribution of reads with that of the background (input). Using JSD, H3K27me3 from both spleen replicates had values below 0.05, which is indicative of insufficient IP enrichment.

The final measure of IP enrichment evaluated the fraction of reads in peaks (FRiP) by comparing the peak calls with the read distribution for each sample. All tissues produced a high proportion of aligned reads within peaks for H3K4me3, ranging from 0.21 for sesamoid to 0.69 for skin. Similarly, MC3, skin, and spleen generated high FRiP scores for H3K27ac (0.47–0.19), and peaks from skin and spleen also scored well for H3K4me1 (0.47–0.29). Although lower than the values from skin and spleen, FRiP scores from MC3 indicated sufficient enrichment was obtained for H3K4me1 (0.07–0.09). For sesamoid tissue, the ECA_UCD_AH2 replicate generated peaks with comparable enrichment for H3K4me1, H3K27ac, and H3K27me3, while the ECA_UCD_AH2 replicate scored below threshold for both H3K4me1 and H3K27me3 (0.0005 and 0.0043, respectively). Further, H3K27me3 peaks from skin generated a substantially higher fraction of reads compared with MC3 and spleen (0.21–0.24 vs. 0.05–0.10), although all three of these tissues obtained sufficient enrichment based on this assessment.

Replicate Comparison

In addition to quality assessments for the read alignments, peaks called from the biological replicates were compared. For most of the marks, the percentage of genome covered by peaks was consistent with previously reported values for the TOI (**Table 1**). For sesamoid tissue, at least one replicate for H3K4me1, H3K27ac, and H3K27me3 generated fewer peak calls than expected based on results from the other replicate and the MC3 replicates. Additionally, the initial data for H3K27me3 from both spleen replicates yielded fewer peaks in accordance with the low complexity and enrichment scores for those libraries. The Jaccard similarity coefficient identified the highest correlation between the biological replicates for H3K4me3 across all "adopted" tissues, ranging from 0.65 to 0.84 (**Table 2**), and data from skin also showed high correlation for all marks (0.44–0.84). Replicates for spleen and MC3 had moderate levels of similarity for H3K4me1 and H3K27ac (0.32–0.58), while the biological replicates for H3K4me1 and H3K27me3 from sesamoid had no identity detected, consistent with the low-scoring quality assessments.

Additional Data Collection

Due to insufficient enrichment and replicate identity, IP and sequencing were repeated for H3K27me3 from both spleen replicates. Unfortunately, the repeated ECA_UCD_AH1 data had low library complexity and IP enrichment (**Table 1** and **Supplementary Table 2**). To achieve sufficient data for accurate peak calling from spleen tissue, the first round of IP and sequencing from ECA_UCD_AH1 for H3K27me3 and both rounds from ECA_UCD_AH2 were used for combined peak calling. Reads from the two input files for ECA_UCD_AH2 were also merged. The number of combined peaks increased from 4,955 covering 1.98% from the first round of sequencing to 5,267 covering 2.18% of the genome when data were merged (**Table 2**). Similar issues with enrichment prevented sufficient signal for peak calling in sesamoid for three of the four marks, and therefore, a second round of IP and quality evaluation of ECA_UCD_AH2 sesamoid is underway for H3K4me1, H3K27ac, and H3K27me3.

DATA METRICS

After combining replicates, the number of retained peaks for each mark from the SE data ranged from 4,933 to 73,528 for spleen and from 5,628 to 46,511 for MC3 (**Table 2**). For both tissues, H3K4me1— the mark indicative of enhancers— was found to have the highest number of peaks while the repressive mark was found to have the lowest. This pattern is also consistent with the TOI data (Kingsley et al., 2020). For PE skin data, the number of combined peaks varied from 24,353 to 92,971 regions, and H3K4me3, which denotes promoters, was the mark with the lowest number of peaks. Additionally, the amount of the genome covered by H3K27me3 peaks was substantially higher for skin compared to the other equine FAANG tissues analyzed to date (6.28 vs. 2.94%), while the number of reads retained for H3K27me3 from the PE data after filtering (42.8%) was comparable to the average retained for all of the equine H3K27me3 SE data (41.3%, PRJEB42315 and PRJEB35307).

Evaluating general enrichment patterns revealed that the "adopted" tissues detected mark distributions for the activating marks that were consistent with those identified previously for the TOI (**Supplementary Figures 1–3**). Data for H3K27me3 from skin, however, generated strong enrichment around the TSS and upstream of an average gene, while still maintaining a similar level of relative enrichment for H3K27me3 distributed throughout the rest of the gene body and downstream as seen for other tissues (**Supplementary Figure 4**). Evaluation of the spleen datasets detected the strongest H3K27me3 enrichment when combining the original ECA_UCD_AH1 dataset and the merged ECA_UCD_AH2 dataset (denoted as "spleen" on **Supplementary Figure 4**). While enrichment distributions for sesamoid tissue detected consistent patterns for H3K4me1, H3K27ac, and H3K27me3, the relative level of enrichment is lower than expected based on the other tissues. In addition to genome-wide evaluations, the replicate-combined peak calls were also manually evaluated across a small number of well-characterized regions. Consistent

TABLE 2 | Summary of the combined peak calls and replicate comparison.

Mark	Tissue	Combined peak number	% Covered	Jaccard similarity coefficient
H3K4me1	Spleen	73,528	2.98	0.44
H3K4me3	Spleen	28,661	1.56	0.80
H3K27ac	Spleen	51,427	1.82	0.58
H3K27me3 MACS2	Spleen	7,349	0.09	0.01
H3K27me3 SICER	Spleen	449	0.22	0.03
H3K4me1	MC3	46,511	1.16	0.32
H3K4me3	MC3	20,556	1.10	0.75
H3K27ac	MC3	31,547	1.08	0.38
H3K27me3 MACS2	MC3	15,304	0.40	0.28
H3K27me3 SICER	MC3	5,628	2.57	0.28
H3K4me1	Sesamoid	750	0.01	0.00
H3K4me3	Sesamoid	17,361	1.07	0.65
H3K27ac	Sesamoid	13,160	0.67	0.08
H3K27me3 MACS2	Sesamoid	390	0.01	0.00
H3K27me3 SICER	Sesamoid	703	0.26	0.00
H3K4me1	Skin	92,971	4.56	0.50
H3K4me3	Skin	24,353	1.60	0.84
H3K27ac	Skin	54,946	3.38	0.67
H3K27me3 MACS2	Skin	51,480	6.02	0.44
H3K27me3 SICER	Skin	11,764	6.28	0.44

The summary includes the combined number of peaks and the percentage of the genome covered by those peaks. The Jaccard Similarity Coefficient compares the two biological replicates with 1 being perfectly concordant and 0 being entirely discordant. Peaks for H3K4me1, H3K4me3, and H3K27ac were called with MACS2.

with expectations, activating marks were detected at the TSS and upstream of ubiquitously expressed genes such as *ACTB* for all tissues (**Supplementary Figures 5A,B**). Additionally, all "adopted" tissues lacked peaks indicative of active transcription for a liver-specific gene known as *CYP2E1* (**Supplementary Figures 5C,D**).

DISCUSSION

The ENCODE project profoundly impacted scientific understanding of genome function in humans by enabling researchers to explore previously impossible challenges, such as charting genomic landscape shifts during development and uncovering enhancer networks associated with disease (Nord et al., 2013; Rhie et al., 2016). The advancements made by ENCODE paved a path for the FAANG consortium to characterize genomic function in numerous agricultural species (Andersson et al., 2015; Tuggle et al., 2016; Giuffra and Tuggle, 2019), which will expand research opportunities across diverse genera. As a part of the larger consortium, the equine FAANG group established a community-based initiative to "adopt" additional tissues for annotation. As a result of that expansive collaborative effort, characterization of putative regulatory regions was performed in spleen, sesamoid, MC3, and skin. The four additional tissues are of major importance for equine health and traits of economic impact. Specifically, research on catastrophic fracture involving sesamoid and MC3 can benefit from bone-specific annotations as recent advances in treatment have focused on transgenically modified stem cell therapeutics (Ball et al., 2019). Similarly, many diseases and traits under artificial selection in horses, such as melanoma, insect bite hypersensitivity, and coat colors including Appaloosa spotting among others, involve skin tissue (Rieder et al., 2000, 2001; Bellone et al., 2008, 2013; Rosengren Pielberg et al., 2008; Curik et al., 2013; Lanz et al., 2017). Several of these characterized phenotypes have been associated with mutations affecting gene expression (Rieder et al., 2000; Rosengren Pielberg et al., 2008; Bellone et al., 2013), making regulatory regions identified from whole skin a valuable resource for equine researchers. The "Adopt-a-Tissue" effort fits into a broader legacy of collaborative resource development that has historically led to rapid advancements for equine genomics and will continue to push equine science toward new frontiers. In concordance with past community efforts, the high quality data generated from the "Adopted" tissues are publicly available to benefit all investigators and lead to further progress in equine research.

Using quality metrics first standardized by ENCODE (Dunham et al., 2012), we identified low IP enrichment for the broad mark in spleen, sesamoid, and MC3 tissues. Unlike the SE datasets, the skin replicates sequenced with PE reads generated a higher enrichment signal for H3K27me3 as determined by quality metrics and enrichment topology plots. In particular, enrichment near the TSS was more strongly

detected for skin than for any of the TOI or the other "adopted" tissues, suggesting that PE reads may better evaluate the broad repressive mark than SE datasets. With only one tissue evaluated as PE, we cannot exclude the possibility that this enrichment pattern may be skin-specific rather than evidence of a better method for detecting H3K27me3. Although enrichment difficulties have been previously recognized for the broad domains like those of H3K27me3 (Landt et al., 2012; Carelli et al., 2017), investigation of specific ChIP methods for broad histone marks appear to be rare. O'Geen et al. (2011) used both short and long sonication periods to account for the different rates of shearing efficiency for compact versus. open chromatin. They found that the larger DNA fragments after sonication were more enriched for broad repressive histone marks while smaller fragments were more likely to contain active chromatin modifications (O'Geen et al., 2011). Their work suggests that shorter sonication times and stringent size selection may bias ChIP samples toward higher enrichment of regions containing narrow marks at the expense of more condensed areas with broad marks, yet current ChIP-Seq standards do not encourage separate protocols for the different mark topologies (Landt et al., 2012; ENCODE Guidelines for Experiments Generating ChIP-seq Data, 2017). Instead, advances in ChIP-Seq methods have focused on analysis and software development to accommodate the different enrichment levels expected from broad and narrow domains assayed with the same protocol (Zhang et al., 2008; Zang et al., 2009). Future investigations involving H3K27me3 and other broad histone modifications may benefit from developing bench protocols, including sequencing parameters, that are specific for broad marks.

To account for insufficient H3K27me3 signal from spleen tissue, IP and sequencing were repeated for both biological replicates. By combining the reads from both sets of data for ECA_UCD_AH2, we were able to obtain sufficient enrichment for peak identification. These data support that combining results from different IPs performed on the same tissue sample can be a useful approach to obtain the enrichment needed for annotation purposes. Study of the best means for combining information from biological and technical replicates for differential enrichment analyses suggests that combining ChIP datasets without accounting for enrichment levels may lead to more false negatives (Bao et al., 2013). Although our data may not have captured all possible peaks, combining data enabled detection of more H3K27me3 peak calls with higher consistency than possible with the first dataset alone. Therefore, the current peak calls can serve as the starting point for spleen-specific annotations, which can be improved upon with characterization of heterochromatin regions from additional equine spleen samples.

The low quality metrics for three of the four marks from ECA_UCD_AH1 sesamoid tissue indicated there was low IP enrichment. To the best of the authors' knowledge, the MC3 and sesamoid data generated here represent the first histone mark peak calls from healthy, whole bone tissue. The overall lower quality metrics for bone tissues support the difficulty of working with these tissues, however, one of the two replicates for sesamoid showed sufficient quality for all four marks, suggesting the issue may be sample specific. To determine if any issues arose during chromatin extraction or IP, further evaluation of H3K4me1, H3K27ac, and H3K27me3 marks in sesamoid tissue from ECA_UCD_AH1 is warranted. Additional data generated from ECA_UCD_AH1 sesamoid tissue will be added to PRJEB42315 when available.

Previous equine annotations were developed based on homology and transcriptomics, leaving much of the genome, especially noncoding regions, uncharacterized (Hestand et al., 2015; Aken et al., 2016; Mansour et al., 2017). While valuable, annotation of regulatory regions based solely on homology with other species is not expected to be sufficient given the evolutionary role of these elements within and among species (Schmidt et al., 2010; McLean et al., 2011; Shibata et al., 2012; Lowdon et al., 2016). With the first publication of the equine FAANG data from eight prioritized tissues (Kingsley et al., 2020) and the four "adopted" tissues presented in this manuscript, researchers can begin to interrogate the role of regulatory regions in equine traits, such as the recent investigation of a novel 16 KB deletion associated with an ocular disorder known as distichiasis (Hisey et al., 2020). Future annotations for the horse will include maps of regulatory states characteristic of healthy tissue, making it a vital resource to compare against disease states. The histone ChIP-Seq data from the horse have already been integrated into a useable annotation resource by a new project known as FAANGMine (FAANGMine, FAANGMine). Similar to FlyMine (Lyne et al., 2007), the project aims to combine the results from all of the genomic assays used by the FAANG consortium into a single resource for easier use. Thanks to these integration effort, additional equine FAANG datasets including the "adopted" tissue peak calls will open up opportunities for variant investigations in previously uncharacterized noncoding regions and expand research opportunities in equine omics.

DATA AVAILABILITY STATEMENT

Data were submitted to the European Nucleotide Archive following the best practices established by the FAANG Metadata and Data Sharing Committee and the FAANG Data Coordination Centre (Harrison et al., 2018). All of the new data referenced in the article were submitted under project ID PRJEB42315. The following files types were submitted for all high quality data: raw fastq for each mark and input from both replicates (38 files), processed BAM files for each mark and input from both replicates (34 files), bed files with peak calls per replicate including both SICERpy and MACS2 calls for H3K27me3 (32 files), and bed files with combined peak calls including those from both SICERpy and MACS2 for H3K27me3 (16 files). All files and metadata can be accessed from the FAANG Data Portal (https://data.faang.org/home). Previously published FAANG data used in the comparisons are also available from the FAANG data portal under project ID PRJEB35307 (Kingsley et al., 2020).

ETHICS STATEMENT

The animal study was reviewed and approved by University of California, Davis Institutional Animal Care and Use Committee.

AUTHOR CONTRIBUTIONS

CF, RB, JP, TK, JM, NH, GL, LO, SB, MM, and EB: conceptualization. NK, RB, CF, and JP: methodology. NK: formal analysis, investigation, data curation, and visualization. CF, RB, and JP: resources. NK and RB: writing—original draft preparation. CF, JP, JM, TK, NH, GL, LO, SB, MM, and EB: writing—review and editing. RB: supervision. CF, RB, and JP: project administration. CF, RB, JP, TK, JM, NH, GL, LO, EB, SB, and MM: funding acquisition. All authors have read and agreed to the published version of the manuscript.

FUNDING

Funding for experimental materials, and other resources was provided by the Grayson Jockey Club Foundation, United States Department of Agriculture (USDA) NRSP-8 equine species coordinator funds, a Priority Partnership Collaboration Award from the University of Sydney and University of California, Davis, and the Center for Equine Health (CEH) at UC Davis with funds provided by the State of California pari-mutuel fund and contributions by private donors. Support for NK was provided by the Grayson Jockey Club Foundation, USDA (2018-06530), Morris Animal Foundation (D16-EQ-028), and a CEH fellowship. Support for CF was provided by the National Institutes of Health (L40 TR001136). Publication fees supplied by UC Davis Open Access Publication Fund.

ACKNOWLEDGMENTS

The authors would like to acknowledge Drs. Catherine Creppe, Paola Genevini, and the other members of the Diagenode Epigenomics Services Team for their high-quality histone mark profiling service. The authors would also like to thank Drs. Colin Kern, Huaijun Zhou, Pablo Ross, Ying Wang, and the other members of the UC Davis FAANG group for providing technical expertise and feedback. Additionally, the authors would like to recognize the contributions from horse owners that made this work possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.649959/full#supplementary-material

REFERENCES

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., et al. (2016). The Ensembl gene annotation system. *Database.* 2016, 1–19. doi: 10.1093/database/baw093

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* 16, 4–9. doi: 10.1186/s13059-015-0622-4

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., and Wingett, S. (2012). *Trim Galore*. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed May 6, 2020).

Bailey, E. (2010). Horse genomics and the Dorothy Russell Havemeyer Foundation. *Anim. Genet.* 41:1. doi: 10.1111/j.1365-2052.2010.02136.x

Ball, A. N., Phillips, J. N., McIlwraith, C. W., Kawcak, C. E., Samulski, R. J., and Goodrich, L. R. (2019). Genetic modification of scAAV-equine-BMP-2 transduced bone-marrow-derived mesenchymal stem cells before and after cryopreservation: An "off-the-shelf" option for fracture repair. *J. Orthop. Res.* 37, 1310–1317. doi: 10.1002/jor.24209

Bao, Y., Vinciotti, V., Wit, E., and 't Hoen, P. A. C. (2013). Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data. *BMC Bioinformatics* 14:169. doi: 10.1186/1471-2105-14-169

Bellone, R. R., Brooks, S. A., Sandmeyer, L., Murphy, B. A., Forsyth, G., Archer, S., et al. (2008). Differential gene expression of TRPM1, the potential cause of congenital stationary night blindness and coat spotting patterns (LP) in the appaloosa horse (*Equus caballus*). *Genetics* 179, 1861–1870. doi: 10.1534/genetics.108.088807

Bellone, R. R., Holl, H., Setaluri, V., Devi, S., Maddodi, N., Archer, S., et al. (2013). Evidence for a retroviral insertion in TRPM1 as the cause of congenital stationary night blindness and leopard complex spotting in the horse. *PLoS ONE* 8, 1–14. doi: 10.1371/journal.pone.0078280

Burns, E. N., Bordbari, M. H., Mienaltowski, M. J., Affolter, V. K., Barro, M. V., Gianino, F., et al. (2018). Generation of an equine biobank to be used for functional annotation of animal genomes project. *Anim. Genet.* 49, 564–570. doi: 10.1111/age.12717

Caetano, A. R., Shiue, Y. L., Lyons, L. A., O'Brien, S. J., Laughlin, T. F., Bowling, A. T., et al. (1999). A comparative gene map of the horse (*Equus caballus*). *Genome Res.* 9, 1239–1249. doi: 10.1101/gr.9.12.1239

Carelli, F. N., Sharma, G., and Ahringer, J. (2017). Broad chromatin domains: an important facet of genome regulation. *BioEssays* 39, 1–7. doi: 10.1002/bies.201700124

Chowdhary, B. P., Raudsepp, T., Honeycutt, D., Owens, E. K., Piumi, F., Guérin, G., et al. (2002). Construction of a 5000rad whole-genome radiation hybrid panel in the horse and generation of a comprehensive and comparative map for ECA11. *Mamm. Genome* 13, 89–94. doi: 10.1007/s00335-001-2089-8

Curik, I., Druml, T., Seltenhammer, M., Sundström, E., Pielberg, G. R., Andersson, L., et al. (2013). Complex inheritance of melanoma and pigmentation of coat and skin in grey horses. *PLoS Genet.* 9:e1003248. doi: 10.1371/journal.pgen.1003248

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

ENCODE Guidelines for Experiments Generating ChIP-seq Data (2017). Available online at: https://www.encodeproject.org/about/experiment-guidelines/ (accessed June 30, 2020).

FAANGMine. Available online at: http://faangmine.org (accessed July 16, 2020).

Giuffra, E., and Tuggle, C. K. (2019). Functional annotation of animal genomes (FAANG): current achievements and roadmap. *Annu. Rev. Anim. Biosci.* 7, 65–88. doi: 10.1146/annurev-animal-020518-114913

Guérin, G., Bailey, E., Bernoco, D., Anderson, I., Antczak, D. F., Bell, K., et al. (1999). Report of the international equine gene mapping workshop:

male linkage map. *Anim. Genet.* 30, 341–354. doi: 10.1046/j.1365-2052.1999.00510.x

Guérin, G., Bailey, E., Bernoco, D., Anderson, I., Antczak, D. F., Bell, K., et al. (2003). The second generation of the International Equine Gene Mapping Workshop half-sibling linkage map. *Anim. Genet.* 34, 161–168. doi: 10.1046/j.1365-2052.2003.00973.x

Harrison, P. W., Fan, J., Richardson, D., Clarke, L., Zerbino, D., Cochrane, G., et al. (2018). FAANG, establishing metadata standards, validation and best practices for the farmed and companion animal community. *Anim. Genet.* 49, 520–526. doi: 10.1111/age.12736

Hestand, M. S., Kalbfleisch, T. S., Coleman, S. J., Zeng, Z., Liu, J., Orlando, L., et al. (2015). Annotation of the protein coding regions of the equine genome. *PLoS ONE* 10, 1–13. doi: 10.1371/journal.pone.0124375

Hisey, E. A., Hermans, H., Lounsberry, Z. T., Avila, F., Grahn, R. A., Knickelbein, K. E., et al. (2020). Whole genome sequencing identified a 16 kilobase deletion on ECA13 associated with distichiasis in Friesian horses. *BMC Genomics* 21, 1–13. doi: 10.1186/s12864-020-07265-8

Kalbfleisch, T. S., Rice, E. S., DePriest, M. S., Walenz, B. P., Hestand, M. S., Vermeesch, J. R., et al. (2018). Improved reference genome for the domestic horse increases assembly contiguity and composition. *Commun. Biol.* 1, 1–8. doi: 10.1038/s42003-018-0199-z

Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* 26, 1351–1359. doi: 10.1038/nbt.1508

Kingsley, N. B., Kern, C., Creppe, C., Hales, E. N., Zhou, H., Kalbfleisch, T. S., et al. (2020). Functionally annotating regulatory elements in the equine genome using Histone Mark ChIP-Seq. *Genes* 11:3. doi: 10.3390/genes11010003

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22, 1813–1831. doi: 10.1101/gr.136184.111

Lanz, S., Brunner, A., Graubner, C., Marti, E., and Gerber, V. (2017). Insect bite hypersensitivity in horses is associated with airway hyperreactivity. *J. Vet. Intern. Med.* 31, 1877–1883. doi: 10.1111/jvim.14817

Leeb, T., Vogl, C., Zhu, B., de Jong, P. J., Binns, M. M., Chowdhary, B. P., et al. (2006). A human-horse comparative map based on equine BAC end sequences. *Genomics* 87, 772–776. doi: 10.1016/j.ygeno.2006.03.002

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lowdon, R. F., Jang, H. S., and Wang, T. (2016). Evolution of epigenetic regulation in vertebrate genomes. *Trends Genet.* 32, 269–283. doi: 10.1016/j.tig.2016.03.001

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., et al. (2007). FlyMine: an integrated database for drosophila and anopheles genomics. *Genome Biol.* 8:R129. doi: 10.1186/gb-2007-8-7-r129

Mansour, T. A., Scott, E. Y., Finno, C. J., Bellone, R. R., Mienaltowski, M. J., Penedo, M. C., et al. (2017). Tissue resolved, gene structure refined equine transcriptome. *BMC Genomics* 18, 1–12. doi: 10.1186/s12864-016-3451-2

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

McCoy, A. M., and McCue, M. E. (2014). Validation of imputation between equine genotyping arrays. *Anim. Genet.* 45:153. doi: 10.1111/age.12093

McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., et al. (2012). A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet.* 8:e1002451. doi: 10.1371/journal.pgen.1002451

McLean, C. Y., Reno, P. L., Pollen, A. A., Bassan, A. I., Capellini, T. D., Guenther, C., et al. (2011). Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* 471, 216–219. doi: 10.1038/nature09774

Nord, A. S., Blow, M. J., Attanasio, C., Akiyama, J. A., Holt, A., Hosseini, R., et al. (2013). Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* 155, 1521–1531. doi: 10.1016/j.cell.2013.11.033

O'Geen, H., Echipare, L., and Farnham, P. J. (2011). Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods Mol. Biol.* 791, 265–286. doi: 10.1007/978-1-61779-316-5_20

Penedo, M. C. T., Millon, L. V., Bernoco, D., Bailey, E., Binns, M., Cholewinski, G., et al. (2005). International equine gene mapping workshop report: a comprehensive linkage map constructed with data from new markers and by merging four mapping resources. *Cytogenet. Genome Res.* 111, 5–15. doi: 10.1159/000085664

Picard toolkit (2019). *Broad Institute, GitHub Repos.* Available online at: http://broadinstitute.github.io/picard/ (accessed July 8, 2020).

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., et al. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44, W160–W165. doi: 10.1093/nar/gkw257

Raudsepp, T., Gustafson-Seabury, A., Durkin, K., Wagner, M. L., Goh, G., Seabury, C. M., et al. (2008). A 4,103 marker integrated physical and comparative map of the horse genome. *Cytogenet. Genome Res.* 122, 28–36. doi: 10.1159/000151313

Raudsepp, T., Santani, A., Wallner, B., Kata, S. R., Ren, C., Zhang, H., et al. (2004). A detailed physical map of the horse Y chromosome. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9321–9326. doi: 10.1073/pnas.0403011101

Rhie, S. K., Guo, Y., Tak, Y. G., Yao, L., Shen, H., Coetzee, G. A., et al. (2016). Identification of activated enhancers and linked transcription factors in breast, prostate, and kidney tumors by tracing enhancer networks using epigenetic traits. *Epigenet. Chromatin* 9, 1–17. doi: 10.1186/s13072-016-0102-4

Rieder, S., Stricker, C., Joerg, H., Dummer, R., and Stranzinger, G. (2000). A comparative genetic approach for the investigation of ageing grey horse melanoma. *J. Anim. Breed. Genet.* 117, 73–82. doi: 10.1111/j.1439-0388.2000x.00245.x

Rieder, S., Taourit, S., Mariat, D., Langlois, B., and Guérin, G. (2001). Mutations in the agouti (ASIP), the extension (MCIR), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (*Equus caballus*). *Mamm. Genome* 12, 450–455. doi: 10.1007/s003350020017

Rosengren Pielberg, G., Golovko, A., Sundström, E., Curik, I., Lennartsson, J., Seltenhammer, M. H., et al. (2008). A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. *Nat. Genet.* 40, 1004–1009. doi: 10.1038/ng.185

Schaefer, R. J., Schubert, M., Bailey, E., Bannasch, D. L., Barrey, E., Bar-Gal, G. K., et al. (2017). Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. *BMC Genomics* 18, 1–18. doi: 10.1186/s12864-017-3943-8

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328, 1036–1040. doi: 10.1126/science.1186176

Shibata, Y., Sheffield, N. C., Fedrigo, O., Babbitt, C. C., Wortham, M., Tewari, A. K., et al. (2012). Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genet.* 8:e1002789. doi: 10.1371/journal.pgen.1002789

SICERpy. *GitHub Repository*. Available online at: https://github.com/dariober/SICERpy (accessed July 8, 2020).

Swinburne, J. E., Boursnell, M., Hill, G., Pettitt, L., Allen, T., Chowdhary, B., et al. (2006). Single linkage group per chromosome genetic linkage map for the horse, based on two three-generation, full-sibling, crossbred horse reference families. *Genomics* 87, 1–29. doi: 10.1016/j.ygeno.2005.09.001

Tuggle, C. K., Giuffra, E., White, S. N., Clarke, L., Zhou, H., Ross, P. J., et al. (2016). GO-FAANG meeting: a gathering on functional

annotation of animal genomes. *Anim. Genet.* 47, 528–533. doi: 10.1111/age.12466

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science.* 326, 865–867. doi: 10.1126/science.1178158

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 25, 1952–1958. doi: 10.1093/bioinformatics/btp340

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

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.

Copyright © 2021 Kingsley, Hamilton, Lindgren, Orlando, Bailey, Brooks, McCue, Kalbfleisch, MacLeod, Petersen, Finno and Bellone. 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: 27 April 2021
doi: 10.3389/fgene.2021.656334

A *de novo* Full-Length mRNA Transcriptome Generated From Hybrid-Corrected PacBio Long-Reads Improves the Transcript Annotation and Identifies Thousands of Novel Splice Variants in Atlantic Salmon

*Sigmund Ramberg[1], Bjørn Høyheim[2], Tone-Kari Knutsdatter Østbye[3] and Rune Andreassen[1]**

[1] Department of Life Sciences and Health, Faculty of Health Sciences, OsloMet – Oslo Metropolitan University, Oslo, Norway, [2] Department of Preclinical Sciences and Pathology, Faculty of Veterinary Medicine, Norwegian University of Life Sciences, Ås, Norway, [3] Nofima (Norwegian Institute of Food, Fisheries and Aquaculture Research), Ås, Norway

OPEN ACCESS

Edited by:
Amanda Jane Chamberlain, Agriculture Victoria, Australia

Reviewed by:
Emily Louise Clark, University of Edinburgh, United Kingdom
Shannon Marie Clarke, AgResearch Ltd., New Zealand

***Correspondence:**
Rune Andreassen
rune.andreassen@oslomet.no

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

Received: *20 January 2021*
Accepted: *01 April 2021*
Published: *27 April 2021*

Citation:
Ramberg S, Høyheim B, Østbye TK and Andreassen R (2021) A de novo Full-Length mRNA Transcriptome Generated From Hybrid-Corrected PacBio Long-Reads Improves the Transcript Annotation and Identifies Thousands of Novel Splice Variants in Atlantic Salmon. Front. Genet. 12:656334. doi: 10.3389/fgene.2021.656334

Atlantic salmon (*Salmo salar*) is a major species produced in world aquaculture and an important vertebrate model organism for studying the process of rediploidization following whole genome duplication events (Ss4R, 80 mya). The current *Salmo salar* transcriptome is largely generated from genome sequence based *in silico* predictions supported by ESTs and short-read sequencing data. However, recent progress in long-read sequencing technologies now allows for full-length transcript sequencing from single RNA-molecules. This study provides a *de novo* full-length mRNA transcriptome from liver, head-kidney and gill materials. A pipeline was developed based on Iso-seq sequencing of long-reads on the PacBio platform (HQ reads) followed by error-correction of the HQ reads by short-reads from the Illumina platform. The pipeline successfully processed more than 1.5 million long-reads and more than 900 million short-reads into error-corrected HQ reads. A surprisingly high percentage (32%) represented expressed interspersed repeats, while the remaining were processed into 71 461 full-length mRNAs from 23 071 loci. Each transcript was supported by several single-molecule long-read sequences and at least three short-reads, assuring a high sequence accuracy. On average, each gene was represented by three isoforms. Comparisons to the current Atlantic salmon transcripts in the RefSeq database showed that the long-read transcriptome validated 25% of all known transcripts, while the remaining full-length transcripts were novel isoforms, but few were transcripts from novel genes. A comparison to the current genome assembly indicates that the long-read transcriptome may aid in improving transcript annotation as well as provide long-read linkage information useful for improving the genome assembly. More than 80% of transcripts were assigned GO terms and thousands of transcripts were from genes or splice-variants expressed in an organ-specific manner demonstrating that hybrid error-corrected long-read transcriptomes may be applied to study genes and splice-variants

expressed in certain organs or conditions (e.g., challenge materials). In conclusion, this is the single largest contribution of full-length mRNAs in Atlantic salmon. The results will be of great value to salmon genomics research, and the pipeline outlined may be applied to generate additional *de novo* transcriptomes in Atlantic Salmon or applied for similar projects in other species.

Keywords: Atlantic salmon, transcriptome, full-length mRNA, hybrid error correction, PacBio Iso-seq, Illumina sequencing

INTRODUCTION

Atlantic Salmon (*Salmo Salar*) is a species with significant value both economically and scientifically. It is an important aquaculture species, and there is also substantial commercial harvesting of wild salmon (FAO, 2018). Both of these activities benefit greatly from increased knowledge of salmon genetics, aiding in breeding to improve yield, quality and welfare for farmed salmon, and in monitoring the health of wild populations (Yanez et al., 2014; Abdelrahman et al., 2017; Houston and Macqueen, 2019).

Salmonids have undergone a relatively recent whole genome duplication (WGD) event, (the salmonid-specific fourth vertebrate whole genome duplication, Ss4R) approximately 80 million years ago (Allendorf and Thorgaard, 1984; Macqueen and Johnston, 2014). They are now undergoing rediploidization, which makes Atlantic salmon a model species useful for studying post WGD phenomena like rediploidization and conservation of partial tetrasomy (Lien et al., 2016; Campbell et al., 2019). Smoltification, the process by which Atlantic salmon and other anadromous salmonids adapt from life in freshwater to saltwater (Hoar, 1988), represents another scientifically interesting and unique developmental transformation that would be of interest to study by omics-technologies. This transition is also a management challenge in aquaculture due to the high mortality rate associated with the post saltwater transfer period (Hjeltnes et al., 2019). Infectious diseases caused by various pathogens are also a major challenge for the aquaculture industry and continues to lead to large economic losses and reduced fish health (Hjeltnes et al., 2019). High-quality transcriptomic resources are extremely valuable when studying the underlying molecular processes governing such developmental transformations, molecular details of infectious diseases as well as in the study of the post WGD phenomena. They are also very important resources for the continuing knowledge-based aquaculture management to improve fish welfare and ensure growth of the aquaculture industry (Abdelrahman et al., 2017).

A chromosome level assembly of the *Salmo Salar* genome has been publicly available since 2015 thanks to the efforts of the International Cooperation to Sequence the Atlantic Salmon Genome (ICSASG) (Lien et al., 2016), but the transcript-level resources are limited. There has been some work on generating full-length mRNA transcripts for *Salmo Salar* (Andreassen et al., 2009; Leong et al., 2010). The vast majority of the protein coding transcripts in the NCBI RefSeq database were, however, annotated by use of *in silico* predictions from the genome sequence supported and corrected by ESTs originating from sequencing of cDNA libraries and high-throughput sequencing (HTS) of transcriptomes on platforms producing short-read data (at time of writing, 498 177 ESTs and 4 475 852 530 short-reads) (Hagen-Larsen et al., 2005; Adzhubei et al., 2007; Koop et al., 2008; NCBI, 2015). While these methods are useful for identifying the presence of particular gene products, they are much less useful for characterizing transcript isoforms (Conesa et al., 2016). Thus, the identification of splice variants and possible misplacement of short transcript sequences on the genome due to the existence of highly similar ohnologous genes (resulting from the salmonid specific WGD) are challenges not easily solved in Salmonids if relying on short-read transcript sequencing alone (Leong et al., 2010; Liu et al., 2012). A full-length protein coding transcriptome from a species (the CDS as well as the 5′- and 3′UTRs) and its repertoire of splice variants is an essential resource to reliably annotate protein coding transcripts and understand how such structural variants impact disease and economical important traits in farmed animals (Abdelrahman et al., 2017; Giuffra et al., 2019).

Long-read sequencing based on the PacBio Iso-Seq SMRT sequencing technology produce full-length transcript sequences (e.g., full-length mRNAs) by sequencing single molecules (Rhoads and Au, 2015). This method solves the problems associated with assembly of short-read HTS data by producing reads that span the entirety of a protein coding transcript, including the CDS and the 5′- and the 3′UTRs of mRNAs. This method, thus, allows one to accurately identify different splice variants and, in salmonids, would allow one to distinguish between transcripts with highly similar ohnologous coding sequences as the complete read (including the less conserved 3′UTR) is generated from a single molecule. Being able to unambiguously characterize the 3′UTRs of specific transcript variants would also benefit the study of regulatory elements targeting these regions, such as microRNAs (Woldemariam et al., 2019, 2020; Shwe et al., 2020). The vast majority of the 3′UTRs in the present Atlantic reference transcriptome (RefSeq) (NCBI, 2015) are, however, predicted from the Atlantic salmon genome sequence with short-read support (RefSeq XM entries).

The development of the PacBio Iso-Seq SMRT sequencing technology has allowed for high-throughput long-read sequencing suitable for sequencing the complete transcriptome of a species or generate tissue specific transcriptomes to study tissue specific gene expression (Wang et al., 2016). The higher error-rates associated with long-read sequencing

may be counteracted by generating a consensus out of multiple reads from a single molecule (High-Quality reads) applying the Iso-Seq method (Gordon et al., 2015; Rhoads and Au, 2015). The error rate can be further improved using graph-based hybrid error correction methods (Au et al., 2012; Salmela and Rivals, 2014; Sahraeian et al., 2017). This approach utilizes the long-reads as basis for short-read alignment that include sequence error correction. Thus, the long-reads provide the structural information of all isoforms, while the long-read isoforms are error-corrected by a much higher read number of short-reads with a superior read accuracy. This allows for the generation of a *de novo* full-length transcriptome with a quality comparable to any reference resource without the use of the current genome assembly as an error-correcting source (Feng et al., 2019). This is particularly important in non-model species where genome assemblies have, in general, considerable potential for quality improvement.

The aim of this study has been to provide the first high-quality full-length protein coding transcriptome resource for Atlantic salmon. We have a particular interest in the study of expression changes and regulation of gene expression during smoltification and sea-water transfer as well as expression changes and gene regulation in response to infectious diseases (Woldemariam et al., 2019, 2020; Shwe et al., 2020). Two of the samples included in this study were therefore selected from a challenge study to reveal the full-length sequences of mRNAs expressed in head-kidney when infected with salmonid alpha virus (SAV) (McLoughlin and Graham, 2007; Andreassen et al., 2017; Bernhardt et al., 2021). Samples from head-kidney were chosen, as this is one of the main immune organ in fish and frequently used in fish immunological studies of gene expression (Bjørgen and Koppang, 2021). The samples from the three different main stages of smoltification; pre-smolt, smoltified fish and post sea-water transfer were chosen from gills, liver and head-kidney. These are all important organs in this developmental transition, and the samples were from our recent and ongoing study of smoltification (Shwe et al., 2020). The development and evaluation of a long-read based transcriptome pipeline was another aim. We have used a combination of existing tools for sequence analysis, curation and annotation of PacBio Iso-Seq transcript data applying both sequel I and sequel II platforms. The quality of long-reads from the PacBio platform were then further improved by use of additional transcript data generated on the Illumina short-read platform. Applying hybrid error correction algorithms that were complemented with in-house developed scripts the sequence accuracy was increased. Finally, after producing a high-quality transcriptome data set that consisted of full-length mRNAs with complete CDSs, the transcripts were functionally annotated. The pipeline processing is independent of the Atlantic salmon genome sequence and other transcript sources like RefSeq for error correction. This allowed transcripts previously predicted from the genome sequence and ESTs to be experimentally validated by a long-read *de novo* transcriptome, and new splice variants and paralogs to be reliably characterized.

MATERIALS AND METHODS

Fish Sample Materials

Table 1 gives an overview of the samples sequenced including information about the experimental condition and their organ type. The table also gives the unique labels used for each sample in the following analysis. Two of the head-kidney samples included in the study (SAV_Control and SAV_challenge, **Table 1**) were from one healthy control fish and one fish challenged with Salmonid Alphavirus, respectively. The challenge trial was carried out at the Industrial and Aquatic Laboratory (ILAB, Bergen High Technology Centre, Bergen, Norway) in February/March of 2018 (Bernhardt et al., 2021). Post-smolt fish from the breed SF Optimal (Stofnfiskur Iceland) were challenged by cohabitation in saltwater with salmon shedders (carrier fish) injected with salmonid alphavirus subtype 3 (SAV3) from Norway (Taksdal et al., 2015). All the fish used in the challenge trial were unvaccinated allowing the study of the immune response following viral infection with SAV. All fish tested negative for SAV3, Infectious salmon anemia virus (ISAV), Infectious pancreatic necrosis virus (IPNV), Piscine myocarditis virus (PMCV), Piscine orthoreovirus (PRV), and Salmon gill poxvirus (SGPV) prior to the challenge trial confirming that the control fish were healthy fish not infected by any of the fish virus commonly seen in aquaculture industry. The average weight of the fish was 110.9 g, and in the experimental period dissolved oxygen was between 79–97%, water temperature was between 11.5–12.4°C, and salinity was between 34.1–34.5‰ across the tanks. Samples (challenged fish and control) were collected at day 37 in the SAV challenge trial and fixed in RNA later (Life Technologies, Carlsbad, CA, United States) immediately after collection. A successful SAV3 infection was confirmed by detection of the viral sequence in the challenge sample. The experimental study was approved by the National Research Authority in Norway (NARA). All salmon used for sampling in the experiment were euthanized according to standard protocols approved by the Norwegian Food Safety Authorities prior to sampling. For simplicity, these two samples are referred to as the SAV samples.

The nine remaining samples included were collected from fish used in a study of miRNA gene expression changes during smoltification and early saltwater period (Shwe et al., 2020). Samples were taken from the head-kidney, gills and liver prior to smoltification (day 0; HKU1, GiU1, LiU1), at the end of the smoltification (day 81; HKU4, GiU4, LiU4), and 4 weeks after sea water transfer (day 111; HKU7, GiU7, LiU7). The liver and gill samples from day 0 were from one fish while the head-kidney sample were from another. The three samples from day 81 were from same fish and this was also the case for the three samples from day 111. Details about smoltification conditions and the sample collection is given in Shwe et al. (2020). Shortly, fish were anesthetized by an overdose of MS-222 (tricaine 123 methanesulfonate, 0.1 g/L) prior to sampling and killed by a blow to the head. The tissue samples were immediately collected, frozen in liquid hydrogen and stored at −80°C. All fish handling procedures complied with the guidelines of the EU-legislation (2010/63/EU), as well as with

TABLE 1 | Distribution of tissue types and experimental conditions of samples sequenced with sample labels as used in the final dataset.

Tissue type	Pre-Smoltification[1]	Post-Smoltification[1]	Post-Seawater Transfer[1]	Control Adult[2]	SAV-challenged Adult[2]
Gill	GiU1	GiU4	GiU7	–	–
Liver	LiU1	LiU4	LiU7	–	–
Head-Kidney	HKU1	HKU4	HKU7	SAV_Control	SAV_Challenge

The smoltification samples[1] are from a study of Shwe et al. (2020) while the challenge samples[2] are from Bernhardt et al. (2021).

the Norwegian legislation. The experiment was considered as a non-regulated procedure according to the National Legislation on Animal Research since the fish had not been exposed to any pain or distress. Thus, this experiment did not require application for approval from the Norwegian Food Safety Authority. For simplicity, these samples are referred to as the smoltification samples.

PacBio Library Preparation and Iso-Seq Sequencing, Illumina Library Preparation and RNA Sequencing

The nine smoltification samples were all processed at the Earlham Institute (Norwich, England). RNA extraction was performed using the Qiagen RNeasy Mini Kit (Qiagen, Hilden, Germany), with On-column DNase Digestion using the RNAse-Free DNase Set, according to the manufacturer's protocol. The total RNA extracts were used for both PacBio long-read sequencing and Illumina short paired-end sequencing. The PacBio non-size selected Iso-seq library preparation was used for the for long-read sequencing. The Express Template Prep 2 protocol requiring RIN-values > 8 on the RNA-samples was applied, and each of the nine samples were individually processed. The resulting cDNA-adapter complexes from the nine samples were sequenced on one PacBio Sequel II 8M SMRT Cell each. For short-read sequencing on the Illumina platform, the automated NEBNext Ultra II Directional RNA-Seq library kit with Poly-A selection (New England Biolabs, Inc., Ipswich, MA, United States) was used for library preparation, and the paired-end sequencing (150 bp) was performed using one Illumina NovaSeq 6000 SP flow cell for all nine samples multiplexed together.

The RNA extraction, library preparation and sequencing of the two SAV samples was carried out by Genewiz Germany GmbH (Leipzig, Germany). RNA was extracted with the RNeasy Plus Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol and the samples used for sequencing had a RIN > 8. For long-read PacBio Iso-seq sequencing, the cDNA synthesis was performed using the SMRTer PCR cDNA synthesis kit (Clontech Laboratories, Inc., Mountain View, CA, United States) without size-selection, and the cDNA-adapter complex was generated using the SMRTbell template prep kit V1.0. (Pacific Biosciences of California, Inc., Menlo Park, CA, United States). Each sample was sequenced using the PacBio Sequel I. Each sample was sequenced on two 1M v2 SMRT cells to compensate for the lower read count on the Sequel I compared to the Sequel II. The short-read sequencing applied the NEBnext Ultra RNA library preparation kit (New England Biolabs, Inc., Ipswich, MA, United States) in accordance with manufacturers protocol. The paired-end (150 bp) sequencing was carried out on the Illumina HiSeq 4000 platform.

Pipeline for Generation of a Non-redundant *de novo* Transcriptome Resource With Error-Corrected High-Quality Long-Reads Free of Interspersed Repeats

The Iso-Seq raw long-reads from the PacBio Iso-seq sequencing were processed through the IsoSeq3 pipeline (PacBio, 2020) as illustrated in **Figure 1**. SMRT link version 8.0 was used for the smoltification samples, while 6.0 was used for the SAV samples, as they were sequenced and processed before 8.0 was released. Data from each sample was processed independently. Only High Quality (HQ) reads, meaning they were supported by at least two FLNCs and with a predicted sequence accuracy $\geq$ 99% (>Q20) (fasta output, **Figure 1**) were used in our downstream analysis.

Cutadapt 1.18 (Martin, 2011) was applied for adapter removal of Illumina reads and quality was checked with FASTQC (Andrews, 2010; **Figure 2**, Cutadapt). The high quality paired-end reads were applied for hybrid error correction of the Iso-Seq generated HQ reads using version 0.9 of the LoRDEC algorithm [Long-read De Bruijn Graph (DBG) Error Correction] (Salmela and Rivals, 2014) with the Illumina reads originating from the same sample, using k-mer size 21 and solidity threshold 3 (**Figure 2**, LoRDEC).

The error-corrected HQ reads (EC-HQ reads) were then filtered using an inhouse python script which removed any EC-HQ read that was less than 99% covered by the De Brujin graph generated from the Illumina reads (Inhouse python script, **Figure 2**). Internal sequence gaps not supported by the graph were not allowed, while one percent or less of the terminal ends were allowed to be without graph support with solidity threshold 3 (a strict filtering with 100% coverage would have demanded at least three Illumina reads to start or end at the very 5′ or 3′ terminal bases in the HQ reads). The eleven sample files with the filtered EC-HQ reads were concatenated, and each EC-HQ read was given a new unique ID number. At the same time, the sample origin and number of supporting FLNC sequences from IsoSeq3 for each new ID was noted in a separate file to make their sample origin and FLNC support traceable (**Figure 2**, Inhouse Python script). In practice, the error correction ensured that any base pair in the long-reads were supported by at least three Illumina reads (average phred quality 36) that consecutively covered the long HQ reads. In cases where there were differences in single base positions or small numerical differences in homopolymer stretches between the HQ read and the supporting sections of the

FIGURE 1 | The PacBio Isoseq3 pipeline for processing SMRT-sequencing data. Each Zero-Mode Wave (ZMW) provides information from a single DNA polymerase, which sequences each cDNA-SMRTBell adapter repeatedly. Consensus: the CCS program generates a consensus sequence for each read that contains a complete repeated insert-adapter complex. Demulitplex: lima filters away sequences with unwanted primer combinations, trims away the adapter sequences, and orients the reads in the 5′→3′orientation. Refine: the refine program filters away concatemers, and sequences without polyA tails of at least 20 bp. Finally, it trims the polyA tails from the remaining sequences. Cluster: Isoseq cluster performs conservative clustering of sequences and uses partial order alignment to generate a consensus sequence for each cluster. The output is classified as High Quality or Low Quality based on the predicted accuracy. The final outputs are high quality and low-quality sequences in fastq format. This figure is used with permission from Pacific Biosciences.

De Brujin graph, the HQ reads were in effect corrected by the Illumina sequences.

All concatenated EC-HQ reads were searched against version 3.0 of the Dfam database (Hubley et al., 2016) using version 4.0.9 of Repeatmasker (Smit et al., 2013) as part of the OmicsBox software package, using the RMBlast search engine with default speed and sensitivity settings (Smith-Waterman score threshold of 225). Any EC-HQ read that matched any of the interspersed repeats present in bony fishes (*Actinopterygii*) were removed using an inhouse python script (Repeatmasker, **Figure 2**).

A grouping of the EC-HQ reads likely to originate from same genome locus was carried out applying cdna_Cupcake

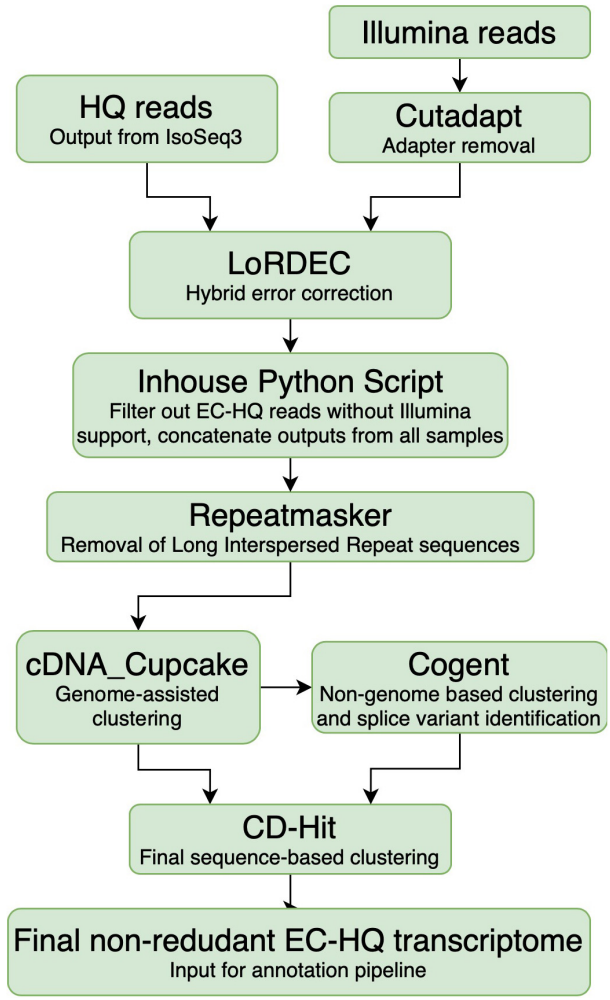


FIGURE 2 | Overview of the analysis pipeline from processing of sequences up to a non-rendundant Error Corrected High Quality transcriptome. The PacBio SMRT High Quality reads were the input from the PacBio platform. The Illumina reads were first trimmed using cutadapt to remove the adapter sequences. Subsequently, they were used to generate a De Bruijn graph for LoRDEC to error-correct of the High Quality reads on a sample-by-sample basis. Inhouse python script: The error-corrected reads were filtered based on degree of Illumina support and coverage of the High Quality reads. Repeatmasker was used to identify and remove reads containing known Long Interspersed Repeats. Sequences that could be mapped accurately to the *Salmo salar* or *Salmo trutta* genome were clustered using cdna_Cupcake, while the remaining sequences were instead clustered using Cogent. All the reads were additionally clustered using CD-Hit prior to annotation. The final output was a non-redundant Error Corrected High Quality transcriptome.

version 12.1.0 (Tseng, 2020a) with minimap2 version 2.17 (Li and Birol, 2018). The EC-HQ reads were aligned against the ICSASG_v2 assembly of the Atlantic Salmon genome (RefSeq accession no GCF_000233375.1) with default parameters, apart from allowing up to 10 000 bp overhangs in the 5′ end and 1000 bp overhangs in the 3′ end. This allowed any EC-HQ reads to be grouped by genome co-ordinates and assigned into groups of transcripts originating from same locus. Furthermore, reads with the same splice-pattern, representing the same isoform, were clustered together keeping the longest one as the representative isoform sequence of the cluster. The groupings of clusters and genome mapping co-ordinates were retained in the fasta headers as information that was utilized by downstream applications (SQANTI3). The relaxed 5′ and 3′ overhang cutoffs were used to allow shorter EC-HQ reads that represented fragments of other full-length EC-HQ read transcripts to be clustered together with the full-length transcripts representing this isoform, rather than being erroneously identified as separate isoform variants.

The EC-HQ reads which did not map well to the Atlantic Salmon genome were mapped against the RefSeq version of the fSalTru1.1 assembly of the *Salmo trutta* genome (RefSeq accession no GCF_901001165.1) applying cdna_Cupcake and same parameters. The remaining sequences that did not map to either the *Salmo salar* or the *Salmo trutta* genomes were mapped against the SAV genome (GenBank accession no. KC122923) as one of the individual samples was SAV infected, and the matching reads were discarded.

The remaining EC-HQ reads that did not map to either of the Salmonid genomes were clustered using Cogent 6.1.0 using default parameters (Tseng, 2020b) (Cogent, **Figure 2**). Cogent generates a pseudo-genome by attempting to recreate a genome sequence that could give rise to all observed transcripts in the dataset, and then clusters and groups the EC-HQ reads into families using cdna_Cupcake. In this manner, the EC-HQ reads that did not map well to either of the two genome sequences were grouped into transcripts likely to be structural variants from same gene.

All the EC-HQ reads, both the ones that were mapped to either of the genome assembly sequences as well as those grouped by Cogent were finally clustered with CD-Hit version 4.8.1 (Li and Godzik, 2006; Fu et al., 2012) (CD-Hit, **Figure 2**). This final clustering aligned any shorter EC-HQ read to a longer identical isoform if present in the dataset. The settings used in CD-hit alignment were:

1. Sequence identity threshold 0.99.
2. Local sequence alignment.
3. Cluster reads to most similar longer EC-HQ read if there were more than one fitting the alignment criteria.
4. The short EC-HQ read must align with 99% of its length to the longer (1% "uncovered" bases). If the shorter EC-HQ read were larger than 3000 bp the 1% limit was replaced by 30 bp or less uncovered bases.
5. The one long EC-HQ read that other shorter reads were aligned to was allowed to have any amount of overhang.

This final clustering assured that identical structural variants were aligned into a single representative EC-HQ read. In most cases the ones aligned to a longer EC-HQ reads would be 5′ incomplete EC-HQ reads or EC-HQ reads with an incomplete 3′UTR due to mispriming in the cDNA synthesis. This assured that every different structural isoform was represented by a single full-length EC-HQ read. An inhouse python script was used to identify FLNC support and contributing samples for each non-redundant EC-HQ read following the final clustering. This non-redundant EC-HQ read transcriptome was further analyzed by classification of structural variants (Materials and Methods,

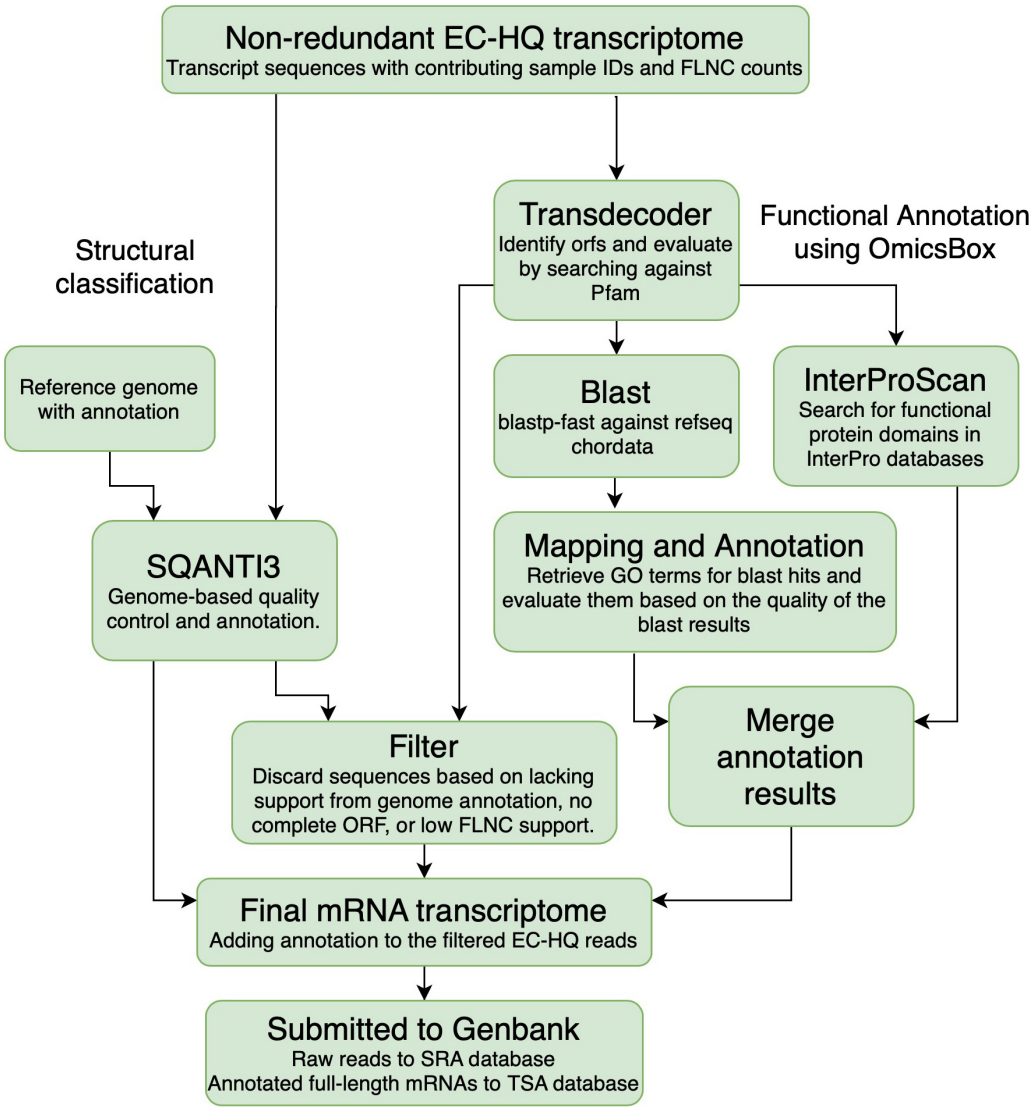


FIGURE 3 | Overview of the annotation process. Transcripts that were clustered using the *Salmo salar* or *Salmo trutta* genomes were characterized against the genome annotation for the corresponding species using SQANTI3. All the sequences were also used for open reading frame prediction using Transdecoder. The sequences predicted to contain a complete coding sequence were Blasted against the RefSeq protein database and searched through the Interpro database to retrieve gene names and functional annotation. The reads were filtered based on the SQANTI-classification, open reading frame prediction and support in the PacBio sequencing data. Information from the structural classification path and the functional annotation path was added to the final filtered mRNA transcriptome.

"Classification of Structural Variants") and functional annotation (Materials and Methods, "Functional Annotation").

Classification of Structural Variants

The EC-HQ reads that had been clustered by cDNA_cupcake using the *Salmo salar* or the *Salmo trutta* genome were classified and compared against the existing annotation for the respective genomes. This analysis was carried out with SQANTI3 (v 1.0.0) with default parameters (Tardaguila et al., 2018) (SQANTI3, **Figure 3**). SQANTI3 compares each EC-HQ read to genome annotation information on the locus where it maps. Based on how the reads match to the genome annotation the comparisons can give the following main category classification of each EC-HQ read:

Full Splice Match (FSM) which is an identical match to a transcript isoform present in the genome annotation with all the same splice junctions and exons.

Incomplete Splice Match (ISM) representing an incomplete but otherwise identical match to a known isoform. All the present splice junctions match, but there are exons missing in either or both ends of the EC-HQ transcript.

Novel In Catalog (NIC), a novel isoform with another combination of exons than those isoforms annotated in the genome, but with a combination of known splice-junctions from previously annotated isoforms.

Novel Not in Catalog (NNC), novel isoform containing at least one splice junction not present in the annotation. Consequently, these have at least one novel exon.

Intergenic, meaning a novel transcript that maps to a locus with no previously annotated genes in the current version of the Atlantic salmon genome.

Genic intron, meaning the transcript maps entirely within the intron of an annotated gene in the Atlantic salmon genome.

Genic genomic, which means the transcript overlaps annotated introns and exons.

Antisense, which means there is no annotated gene at the locus on the strand where the transcript matches, but there is one on the reverse strand.

Fusion, meaning the read spans across two different annotated loci in the current genome annotation.

Sequences were not mapped against the *Salmo trutta* genome if they were successfully mapped against the *Salmo salar* genome. A structural classification indicating the EC-HQ was of any other category than FSM in *Salmo salar* does therefore not exclude that it would have mapped different (e.g., FSM) in *Salmo trutta.*

SQANTI also provides additional useful information like number of exons, splice junction signals (canonical or not), CDS length, polyA-signals upstream on 3′end and genomic percentage of A's downstream of a 3′ termination site (used to judge whether there was cDNA synthesis mispriming).

Functional Annotation

All EC-HQ reads in the final non-redundant transcriptome dataset (**Figure 2**) were subjected to functional annotation using the OmicsBox software suite (**Figure 3**; BioBam, 2019). Coding sequences were predicted with TransDecoder v5.5.0. (Haas et al., 2013; Haas and Papanicolaou, 2015) using homology search against Pfam 32 to confirm ORFs with a minimum length of 150 bp (termed complete CDS in the following manuscript if both the start and stop codon were present in-frame in the CDS). The search for ORFs was set to be strand specific, and the single best hit was kept for further analysis.

All complete CDS sequences were fed into the functional annotation workflow of OmicsBox (equivalent to the older Blast2GO software) (Gotz et al., 2008) with the following modified blast parameters: plastp-fast, species filter 89593 Craniata < chordates >, HSP-Hit coverage 70%. Default e-value cutoff for the blastp step was 1.0E-3. The HSP-Hit coverage criteria ensured that any hits were similar across the majority of the sequence, rather than just containing a highly similar partial sequence. Blast results were fed into the GO mapping and GO annotation modules. The complete CDS were also searched for functional motifs using InterProscan. The results were merged into a final functional annotation file. This process allowed us to identify mRNA transcript isoforms with complete coding regions, identify genecodes where these could not be provided by SQANTI, and provide functional descriptions of the proteins in the GO framework as well as enzyme codes.

Sequences that were mapped onto the *Salmo salar* mitochondrial sequence by SQANTI were also fed to TransDecoder using the Vertebrate Mitochodrial genetic code to search for ORFs. Any complete CDS from these sequences were also functionally annotated as described above.

Additionally, for the transcripts in the final full-length mRNA transcriptome (see section "Final mRNA Filtering Based on Supporting Evidence, and TSA Submission"), we identified subgroups of full-length mRNAs expressed in at least three samples from a specific organ type, but not in any of the other organ types in our materials. For these, we generated multi-level GO charts, showing the most specific GO-terms appearing in the dataset over the default abundance cutoffs suggested by OmicsBox, in a non-redundant way. These cutoffs were respectively 144 transcripts for gills, 98 in head-kidney, and 116 in liver.

Final mRNA Filtering Based on Supporting Evidence, and TSA Submission

An inhouse python script (Filter, **Figure 3**) was used to identify EC-HQ reads that represented mRNA transcripts. The inhouse script filtered all EC-HQ reads, and only EC-HQ reads predicted to contain a complete CDS by TransDecoder were kept. Furthermore, they should be classified by SQANTI as full splice match, novel in catalog, or novel not in catalog with canonical splice junctions to be included. If classified differently by SQANTI or only grouped by Cogent, a minimum support by at least 5 FLNC reads was used as threshold for including such structural isoforms in our final full-length mRNA transcriptome. A somewhat stricter FLNC-support criteria, thus, was used for isoforms with these structural classifications, as they did not have the same level of support in the existing genome annotations as the FSM, NIC, and NNCs described above (which only required the minimum 2 FLNC support needed to be classified as HQs). The script also collected all the structural and functional annotation information for the filtered transcripts into a tsv file (**Supplementary File 1**).

The PacBio and Illumina raw sequencing data was submitted to NCBI's SRA database, and the final transcript sequences in our full-length mRNA transcriptome were submitted to the Transcriptome Shotgun Assembly under the accession GIYK00000000. The version described in this paper is the first version, GIYK01000000.

Transcriptome Comparisons Applying BLAST Analysis and SQANTI3 Annotations

Blast analysis was applied to find the degree of sequence similarity between the final *de novo* mRNA transcriptome and the *Salmo salar* mRNAs in the RefSeq database.

The complete set of RefSeq *Salmo salar* mRNA transcripts was identified and downloaded as a full record fasta files using a filtered nucleotide search through the NCBI RefSeq database. These were searched against our full-length mRNA transcriptome using blast 2.9.0 + with e-value cutoff 1e-15, and outputfmt "6 std qcovhsp slen." The same search settings were used in the reverse comparison with our transcriptome as query against the RefSeq sequences. An inhouse python script was used to

filter the blast results. The filter classified transcripts into three categories. Matches between query and subject meeting the following criteria were categorized as identical isoforms: E-value less than 10^{-50}, percentage identity $\geq$ 99%, and either query coverage per high-scoring segment pair >99% or alignment length * 100/subject length > 99. This ensured that any match meeting the e-value and identity thresholds had a greater than 99% coverage of the query by the subject sequence, or a greater than 99% coverage by the subject of the query sequence. These thresholds ensured that matches categorized as identical isoforms were consecutively matching sequences originating from the same isoforms, but allowed either of them to differ in UTR length compared to the other. Matches within this category were further grouped depending on whether the RefSeq RNAs had the longer UTR (query coverage per high-scoring segment pair < 99%) or whether the matching mRNA in our full-length mRNA transcriptome had the longer UTR (alignment length * 100/subject length < 99). The second category, named significant hits, were all matches not meeting the identical isoform criteria, but with an *E*-value of less than 10^{-15}. The remaining query sequences returning E-values more than 10^{-15} were categorized as non-matching RefSeq mRNAs (or non-matching full-length mRNAs in the reverse search).

The overlap between mRNAs in our dataset and the genome-annotation based mRNAs (genome reference sequence GCF_000233375.1) would be the number of sequences classified as FSM by SQANTI3. These were retrieved from the final.tsv file (**Supplementary File 1**). The number of FSM's, the total listed number of mRNAs given in the genome annotation report (NCBI, 2015) and the total number of full-length transcripts in the final mRNA dataset were used to generate the Venn diagram in section "Clustering and Grouping of Unique EC-HQ Reads Revealed That 22% of These Could Not Be Mapped to the Current Atlantic Salmon Genome Sequence."

RefSeq mRNAs with sequences that are mismatches if compared with the correlated mRNAs (exons) in the Atlantic salmon genome assembly was identified by adding the term 'AND "assembly gap:" [All Fields]' to the search in the RefSeq database. These represented RefSeq RNAs that are not supported by the current genome sequence. Furthermore, the number of such RefSeq mRNAs that are supported by our full-length mRNAs was retrieved by searching their accession numbers among those categorized in the identical isoform category described above.

RESULTS

Hybrid Error-Correction Increased the Sequence Accuracy and Allowed for Removal of Sequencing Artifacts

The results from the Pacbio sequencing and Illumina sequencing are summarized in **Supplementary File 2**. As expected one Sequel II cell generated about 4–7 times more HQ reads than two Sequel I cells, but the percentage of HQ reads generated from the CCS reads were similar. Also, the size distribution of reads from the two platforms showed very similar distribution (**Supplementary File 3**) indicating that the reads generated by the two platforms were of equal quality. Following HQ filtering there were a total of 2 080 166 HQ reads distributed across the 11 samples (**Supplementary File 2**). The number of Illumina reads in all samples applied for error correction was more than 900 million with an average phred quality of 36. This resulted in a total of 1 596 834 EC-HQ reads.

The coverage distribution by Illumina reads on the HQ reads is illustrated in **Figure 4**, while the exact number of HQ reads covered by a certain percentage of Illumina reads is given in **Supplementary File 4**. The figure shows that the majority of HQ reads were preserved (and error-corrected) by this filtering process. More than 81% of the HQ reads had a coverage of 99% or more, illustrating that most HQ reads were error-corrected across their entire sequence. The main reason for being removed was not a general poor coverage. Instead, the sequences that were removed (23,3%) were heavily weighted toward high coverage HQ reads (95–99% coverage), but with internal small gaps in their Illumina read coverage. At 99% coverage there were, e.g., approximately 100 000 HQ reads with internal gaps not covered by the De Brujin Graph (DBG). This could indicate that these HQ reads were artifacts produced in the PacBio pipeline representing contaminating genomic sequence or fusion products (different transcripts fused together and SMRT-sequenced). This is the most likely explanation since only smaller parts of the EC-HQ sequences were not covered by the independent and much "deeper" transcriptome sequences from the Illumina platform. While the degree of single bp correction could not be directly measured we did compare the CDS lengths before and after error correction. This comparison revealed that 118 199 (7%) of the reads increased ORF size after correction, 1 455 732 (91%) had the same ORF length, and 22 903 (2%) had a shorter ORF. A substantial portion of the reads were, thus, corrected and 75% of these increased the length of their CDS. The number of short-reads applied for error correction was more than 927 million (**Supplementary File 1**). This is equal to 20% of all short-reads used to annotate exome sequences in the current genome assembly.

One Third of the Transcripts Were Interspersed Repeats

The EC-HQ read dataset were analyzed by Repeatmasker to identify transcripts originating from interspersed repeats (Repeatmasker, **Figure 2**). After filtering out any EC-HQ read matching interspersed repeats the dataset was reduced to a total of 1 090 532 EC-HQ reads. This showed that about one third (32%) of all transcripts in Atlantic salmon are simply expressed interspersed repeats.

Clustering and Grouping of Unique EC-HQ Reads Revealed That 22% of These Could Not Be Mapped to the Current Atlantic Salmon Genome Sequence

The mapping of the EC-HQ reads to the *Salmo salar* genome by cDNA_cupcake (**Figure 2**) showed an 89% success rate with

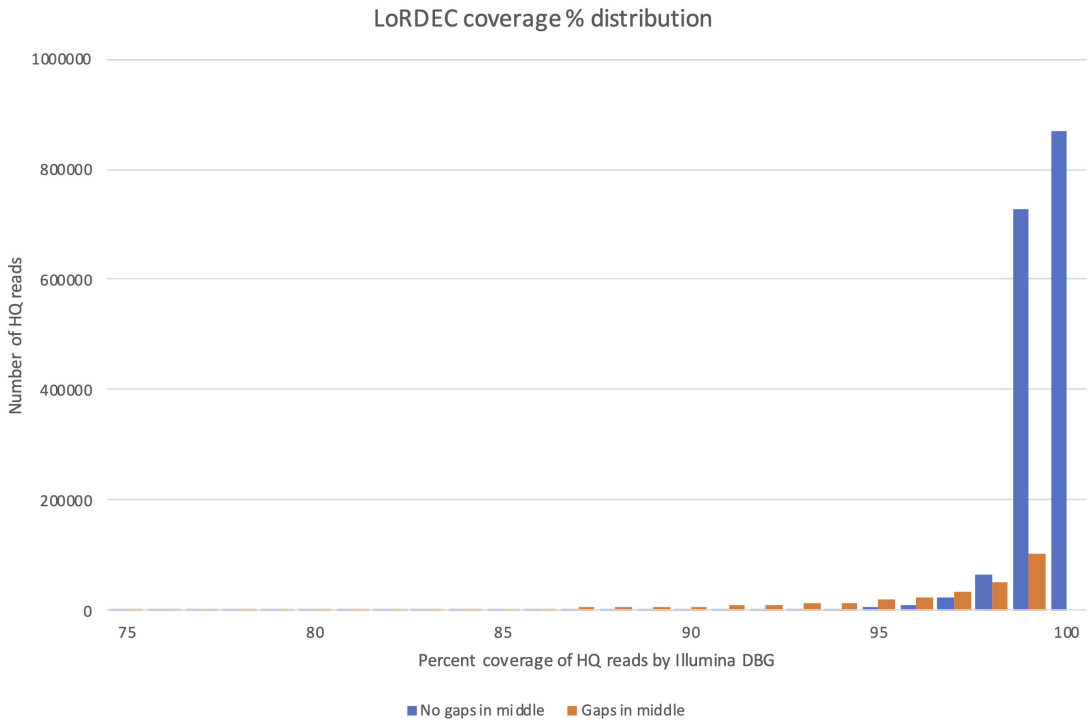


FIGURE 4 | Distribution of coverage by LoRDEC for sequences with (orange bars) and without (orange bars) internal gaps in the coverage interval 75–100%.

972 904 of the EC-HQ reads mapped. These were reduced to 87 315 unique reads following CD-Hit clustering. Approximately half (59 913 of 117 628) of the remaining EC-HQ reads could, however, be cDNA_cupcake clustered to the *Salmo trutta* genome. These were reduced down to 8721 unique reads by CD-Hit. The remaining 57 715 EC-HQ reads were clustered using Cogent, which were reduced to 16 367 unique reads by CD-Hit. cDNA_Cupcake and Cogent also assigned a locus number to each sequence. Thus, if they mapped to the same location on the genomes or the Cogent pseudogenome in an overlapping manner, they were grouped together as sequences that were likely isoforms of the same transcript. Altogether, the cDNA_cupcake, Cogent and final CD-Hit clustering reduced the complete dataset to 112 404 unique EC-HQ transcripts (Final non-redundant EC-HQ transcriptome, **Figure 2**). In summary, 78% of the unique EC-HQ transcripts were mapped to the *Salmo salar* genome while approximately 8% was mapped to the *Salmo trutta*. The remaining 14% unique EC-HQ reads (clustered by Cogent) could not be mapped to either of the two salmonid genomes. It is very unlikely that eight percent of the transcripts would map well in the Salmo trutta genome if the reason for mismatch against the Salmo salar genome was low quality or error in the EC-HQ reads. On the contrary, this indicated that there are missing or misassembled sequences in the current Atlantic salmon genome sequence that prohibited a successful mapping of a surprisingly large proportion (22%) of the of the unique Atlantic salmon EC-HQ reads. Following from this, errors in the genome sequence would be the likely explanation for why 14% of transcripts grouped by Cogent could not be mapped to the *Salmo salar* genome sequence (or the *Salmo trutta* genome). The fact that our final full-length transcriptome matches the current RefSeq mRNAs much better than the transcripts (annotated exons) in the genome sequence (identical isoforms vs. FSMs, section "Full-length Transcriptome Comparison With the Transcript Annotation of the Genome Applying SQANTI3") also points to incorrect annotation of splice products. The *de novo* full-length transcriptome from this study may therefore aid to improve the current transcript annotation in the genome sequence. In light of this, the 22% of transcripts not currently mapped to the Atlantic salmon genome sequence represents a very useful source for long range linkage information that may be used to improve the genome sequence assembly.

A *de novo* Transcriptome That Consisted of 71 461 Full-Length mRNAs From 23 071 Loci

The 87 315 non-redundant EC-HQ reads that had been mapped to the *Salmo salar* genome and the 8721 that had been mapped to the *Salmo trutta* genome were structurally annotated using SQANTI3 (**Figure 3**). The remaining sequences that had been clustered by Cogent could not be meaningfully annotated in this manner, but rather relied on the functional annotations (OmicsBox, **Figure 3**) for RNA category classification. Complete distribution of structural classifications for all non-redundant EC-HQ reads is shown in **Table 2**.

All 112 404 unique sequences, regardless of clustering method, were also used for ORF-prediction with Transdecoder, and

further functional annotation by OmicsBox if they were predicted to have a complete CDS. Following the last filtering step (Filter, **Figure 3**), our final mRNA dataset consisted of 71 461 non-redundant EC-HQ reads that, by use of SQANTI3 and our OmicsBox criteria, were annotated as protein coding transcripts with a complete CDS. These transcripts were predicted to stem from a total of 23 071 loci, or likely loci in the case of Cogent transcripts, with an average of three transcripts per locus. The list of EC-HQ reads, along with SQANTI analysis outputs, sample origin, FLNC support, gene descriptions, GO codes, enzyme codes and TSA contig IDs are provided in **Supplementary File 1**. Despite being classified as anti-sense, genic, intergenic or fusion by SQANTI, these categories contained 2591 transcripts classified by TransDecoder as having a complete CDS and with support as full-length protein coding mRNAs by Pfam. An additional 4803 transcripts similarly classified as full-length protein coding mRNAs by TransDecoder did not map to any of the Salmonid genomes (clustered by Cogent). Sixty-seven percent of the transcripts not mapped or mapped with a non-protein coding category by SQANTI3 were also supported as protein-coding transcripts by the following GO-annotation (section "More Than 80% of the Transcripts Were Assigned GO Terms and Subsets Revealed Organ Specific Expression Patterns") In summary, a total of 71 461 of the unique transcripts (63%) were classified as mRNAs while the remaining transcripts (37%) likely represented some other kind of non-coding RNAs. The length distribution of the mRNAs is given in **Figure 5**. The mRNA transcripts ranged from 319 bases to 13 331 bases in length, with a median length of 1402 bp and a mean length of 3209 bp.

The EC-HQ reads in the final mRNA dataset categorized as Full Splice Match (FSM) (final in **Table 2**) represented an identical match to a known isoform in the current genome annotations (*Salmo salar* or *Salmo trutta*) in terms of splice pattern and sequence identity (SQANTI3 default 95% cutoff). The majority of such matching transcripts in the current Atlantic salmon annotation have been generated from predictions based on the genome sequence with variable support from short HTS reads or ESTs (96% of *Salmo salar* RefSeq mRNAs are XM entries). The FSMs in our dataset, therefore, represents experimental validation of 17 787 transcript isoforms by single molecule sequenced full-length mRNAs. Surprisingly, there were 960 EC-HQ reads mapped as FSMs in the Salmo trutta genome. Obviously, these are also true Atlantic salmon full-length mRNAs, but inconsistencies in the current genome assembly of *Salmo salar* prevented these transcripts from mapping to the *Salmo salar* genome sequence.

The transcripts mapped as either Novel in Catalog (NIC), or Novel Not in Catalog (NNC) with exons defined by canonical splice junctions, represents novel isoforms not currently annotated in the genome sequences. The NICs have combinations of known splice sites which makes them isoforms with new combinations of annotated exons. There were 17 039 such novel transcripts in the final dataset, approximately the same number as all FSMs. There were an even larger number of NNCs, a total of 25 581 transcripts, illustrating the capability of long-read based methods to identify new isoforms not possible to predict reliably using short-reads and the genome sequence alone. Again, a substantial proportion (8,5%) of the transcripts in these categories could only map to, and be classified by SQANTI3, using the *Salmo trutta* genome.

The remaining SQANTI3 categories (ISM, antisense, genic, intergenic, and fusion) comprised the smaller part of the final mRNA transcriptome (a total of 5291 transcripts). Despite their structural classification by SQANTI3 bringing into question if these were true mRNAs, they were all supported as full-length mRNAs by OmicsBox. Also, as a threshold of at least 5 supporting FLNCs was used to filter out possible artifacts from these categories, they are likely to represent true Atlantic salmon protein-coding transcripts. Although the ISM reads were supported by the genome annotation as matching a known transcript but missing exonic sequence in the 5′ or 3′ end (or both), they all had a complete CDS. We did employ the more conservative FLNC-support criteria (at least 5 FLNCs) for ISMs, and given this threshold, it is less likely that they are incomplete, but rather represent full-length mRNAs. A small number of ISM sequences (3.7%) had 80% or higher A content in the 20 bases immediately downstream of where they mapped in the genome sequence. This could have allowed mispriming during cDNA synthesis leading to incorrect 3′UTR lengths in these transcripts. The fraction of transcripts grouped by Cogent with at least 5

TABLE 2 | Distribution of SQANTI-classifications and Cogent grouping for the non-redundant Error Corrected High Quality transcriptome (**Figure 2**) and filtered mRNA transcriptome (**Figure 3**).

Structural Category	*Salmo salar* Pre-filter	*Salmo salar* Final	*Salmo trutta* Pre-Filter	*Salmo trutta* Final	Cogent Pre-Filter	Cogent Final
FSM	19837	17787	1073	960	–	–
ISM	6080	2569	455	131	–	–
NIC	18338	16218	946	821	–	–
NNC	37411	22756	5443	2825	–	–
Antisense	787	256	22	8	–	–
Genic	942	385	24	5	–	–
Intergenic	2331	828	263	80	–	–
Fusion	1589	807	494	222	–	–
Cogent	–	–	–	–	16367	4803

*Sequences that could not be mapped to the Salmo salar or Salmo trutta genomes could not be classified using SQANTI. These are listed in the row Cogent. Pre-filter refers to all non-redundant EC-HQ transcripts (output from pipeline **Figure 2**). Final refers to the final filtered mRNA dataset (output pipeline **Figure 3**).*

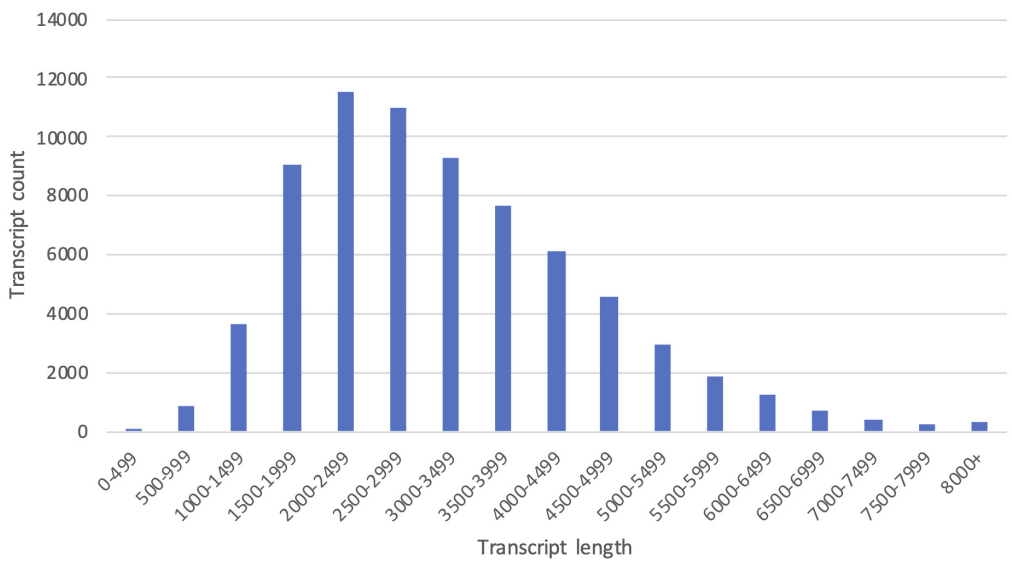


FIGURE 5 | The final full-length mRNA transcriptome distributed by transcript length. Each column shows the number of transcripts falling within the given 500 bp length interval.

supporting FLNCs is approximately as large (4803 transcripts) as those mapped to *Salmo trutta*. Again, this illustrates the ability of long-read based transcriptome sequencing to identify transcripts not detected by short-read supported genome predictions.

An overview of the number of FLNCs supporting each EC-HQ mRNA in the final full-length mRNA transcriptome is given in **Figure 6**. The figure illustrates that more than 74% had a support of more than five FLNCs independent of category. This means that approximately 70% of the FSM, NIC, and NNC categories were supported by five or more FLNCs even if the inclusion criteria for these categories was two or more FLNCs.

A python script was used to estimate the number of instances were transcripts annotated as encoding the same gene by SQANTI or OmicsBox had been assigned to different loci by cdna_Cupcake, or different gene families by Cogent. In all, 12% of genes had at least two EC-HQ mRNA transcripts that were mapped to different loci. This indicates that at least 12% of the expressed genes were likely represented by multiple paralogs in our dataset.

The transcripts have been deposited at DDBJ/EMBL/GenBank as a Transcriptome Shotgun Assembly project under the accession GIYK00000000. The version described in this paper is the first version, GIYK01000000. The TSA Contig ID for each sequence is listed in **Supplementary File 1**.

Comparisons Revealed That the *de novo* Transcriptome Better Supported the RefSeq mRNA Transcripts Than the Genome Assembly Sequence

The current transcript information on Atlantic salmon mRNA sequences is provided from two NCBI sources. One is the transcript information (transcript and isoform variants defined by their exons) given in the annotation to the current *Salmo salar* reference genome sequence. The other source is the current collection of *Salmo salar* mRNA transcripts in the RefSeq database. Although these would be expected to correspond well, they differ in the sequence information given for thousands of transcripts. There are 4475 RefSeq mRNAs that are annotated as either having a gap or additional sequence that is not present in the current genome reference sequence. The long-read based *de novo* mRNA transcriptome from the present study could possibly aid to resolve what are the correct transcript sequences. The single-molecule based method applied here is also expected to be the superior one for isoform identification. This potential for increasing the quality of annotated transcript isoforms with our dataset was investigated by comparing our long-read transcriptome to each of the available sources (the RefSeq mRNA sequences and the annotation of transcripts in the RefSeq genome sequence).

Full-Length Transcriptome and RefSeq Transcriptome Comparisons Applying BLASTN

The *Salmo salar* mRNAs in RefSeq consist mostly of transcripts that are predicted by use of the genome sequence but supported and error-corrected based on EST and short-read sequences as part of the NCBI Eukaryotic Genome Annotation pipeline[1]. All 97 604 *Salmo salar* mRNA sequences in the RefSeq database were blasted against our dataset, classifying hits into three categories (**Figure 7**). Identical isoforms [99% identity and coverage of the shorter matching sequence by the longer matching sequence (See section "Transcriptome Comparisons

[1]https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/

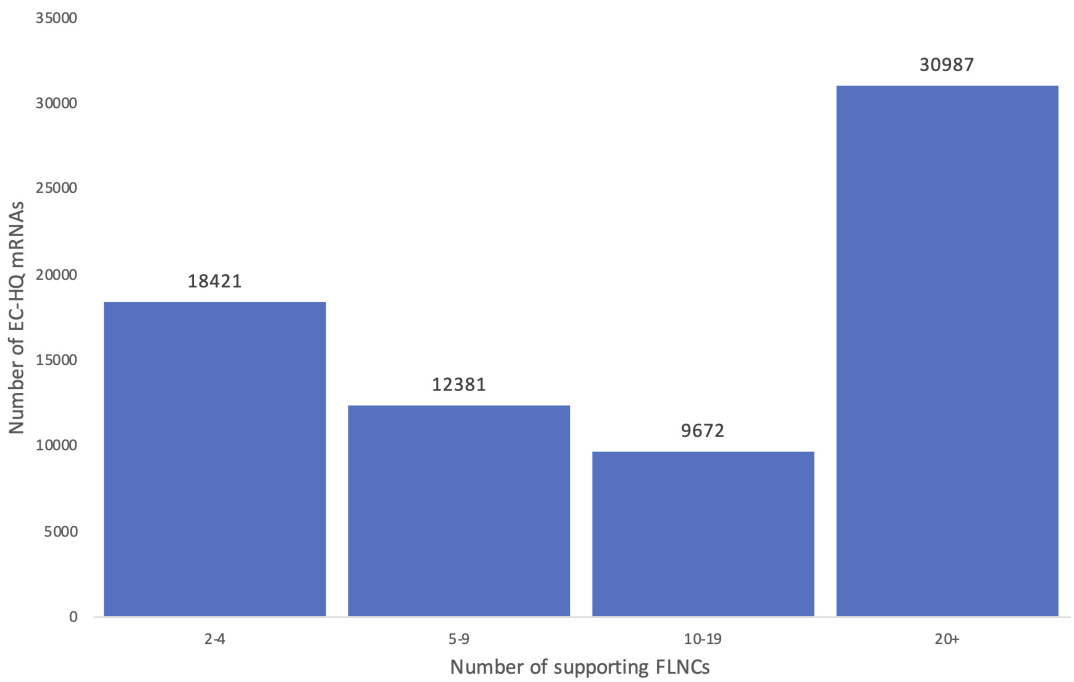


FIGURE 6 | Distribution of Full-Length Non-Concatemer-support in the final mRNA dataset. Each column shows the number of transcripts in the final mRNA transcriptome with a number of Full-Length Non-Concatemer reads supporting the long-reads.

Applying BLAST Analysis and SQANTI3 Annotations" for more detail)]. The second category was significant hits, (all transcripts with e-values smaller than a 1e-15, but not meeting the very strict identical isoform criteria) and the third category was termed non-matching RefSeq mRNAs (all transcripts with e-values larger than e-15 or no hits). Filtering according to these criteria, 24 415 (25%) of the RefSeq transcripts were in the identical isoforms category, providing experimental validation for a quarter of the isoforms in the RefSeq database by full-length mRNAs from our long-read transcriptome. Furthermore, there were around twice as many mRNAs with significant hits (49 785) against our transcriptome, indicating that an additional half of all RefSeq transcripts were present as splice variants or paralogs in our dataset. Given that parts of the RefSeq mRNA sequences are predicted from the genome sequence (96% are XM entries), there is also a possibility that some of the matches in the category significant hits are, in fact, identical isoforms, but not meeting the very strict criteria we applied for this category due to sequence errors. The 23 404 mRNAs in the category of non-matching RefSeq-mRNAs likely represent transcripts from genes not expressed in the organs included in this study. The mRNAs in the category identical isoforms showed a distribution of length differences with the longer transcript being from the RefSeq dataset in 42% of cases, while the two matching transcripts deviated by less than one hundredth of their length in 28% of cases, and in the remaining 30% the longer transcripts were from our *de novo* transcriptome. These size differences were in most cases small and affected only the UTRs, not the CDSs.

We also reversed the comparison, with a new blastn search where our dataset was the query sequences against the *Salmo salar* RefSeq mRNAs (**Figure 8**). This revealed how well our transcriptome is represented in the current RefSeq transcriptome. Notably, the number of sequences with at least one blast hit meeting the identical isoforms criteria was lower when using the full-length transcriptome as the query than when using the RefSeq mRNAs as the query (20 582, **Figure 8** vs. 24 415, **Figure 7**). This indicates that some of our transcripts were classified as identical isoforms to multiple sequences currently listed separately in RefSeq (**Figure 7**). Some possible explanations for this finding are that some sequences in RefSeq are redundant, and/or that some of our transcripts have incomplete UTRs, making them unable to distinguish between some RefSeq entries. The comparison also showed that only 566 of our sequences did not have a significant blast hit (<1e-15). The result from the reversed blastn analysis, thus, indicated that our transcriptome, having 99% support in RefSeq, consisted nearly exclusively of transcript variants of known genes rather than there being transcripts from novel genes (**Figure 8**). Taken together, the two blast analyses showed that our full-length *de novo* transcriptome validated 25% the of currently known Atlantic salmon transcripts in RefSeq, provided a large number of new isoforms significantly matching 50% of the known transcripts in RefSeq, but did not discover many transcripts from novel genes (1%, **Figure 8**).

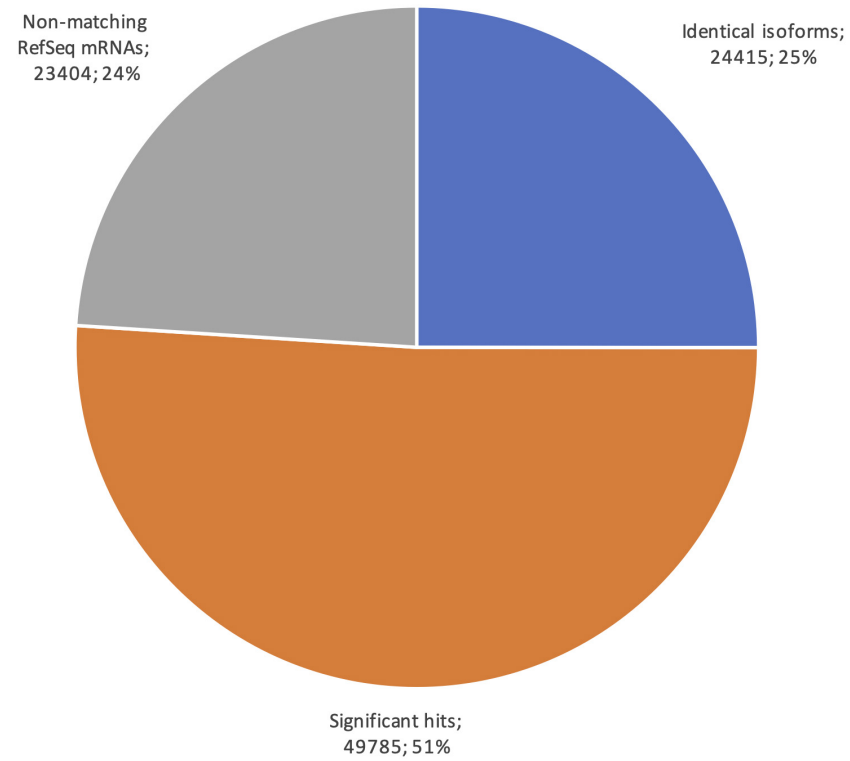


FIGURE 7 | Pie chart showing the distribution of blast results when searching all *Salmo salar* RefSeq mRNAs against the final full-length mRNA dataset. Blue are identical isoforms, orange are significant hits, gray are non-matching RefSeq mRNAs.

Full-Length Transcriptome Comparison With the Transcript Annotation of the Genome Applying SQANTI3

Figure 9 shows the distribution of shared isoform transcripts when comparing our full-length transcriptome with the mRNAs in the *Salmo salar* genome annotation. The figure illustrates that 17 782 EC-HQ mRNA transcripts were full splice matches (identical) to already annotated transcripts in the RefSeq version of the *Salmo salar* genome assembly. The majority of mRNA isoforms predicted in the *Salmo salar* genome annotation (81%, *Salmo salar* mRNAs with no FSM in **Figure 9**) could not be verified by our long-read mRNA transcripts. Furthermore, there were 53 674 mRNAs in our final dataset (75%, Novel mRNAs in **Figure 9**) that represented novel isoforms that mapped to the genome (61%, non-FSM categories in *Salmo salar* in **Table 2**), or mRNAs that did not map at all (14%, *Salmo trutta* or Cogent categories in **Table 2**) due to inconsistencies at the genome sequence level.

The fact that there were considerably fewer FSMs (19%, **Figure 9**) than the number of identical isoforms (25%, **Figure 7**) could indicate that the genome sequence is the less reliable source for transcript sequences. This is also supported by the fact that 14% of the mRNA transcripts did not map at all (mapped to *Salmo trutta* or grouped by Cogent). Furthermore, 1268 of the transcripts that did not map to the *Salmo salar* genome were in the category identical isoforms in the blast comparisons against the RefSeq mRNAs. A similar comparison where the RefSeq transcripts not matching the genome reference sequence (4475 transcripts, methods, "Transcriptome Comparisons Applying BLAST Analysis and SQANTI3 Annotations") were compared to our *de novo* transcriptome showed that 673 of these were in the category identical isoforms. All together, these comparisons indicate that the full-length *de novo* transcriptome may contribute considerably in improving the current transcript annotation quality and aid in improving the reference genome assembly.

More Than 80% of the Transcripts Were Assigned GO Terms and Subsets Revealed Organ Specific Expression Patterns

Figure 10 shows the distribution of OmicsBox annotation results in the final mRNA dataset. Eighty-two percent of the transcripts

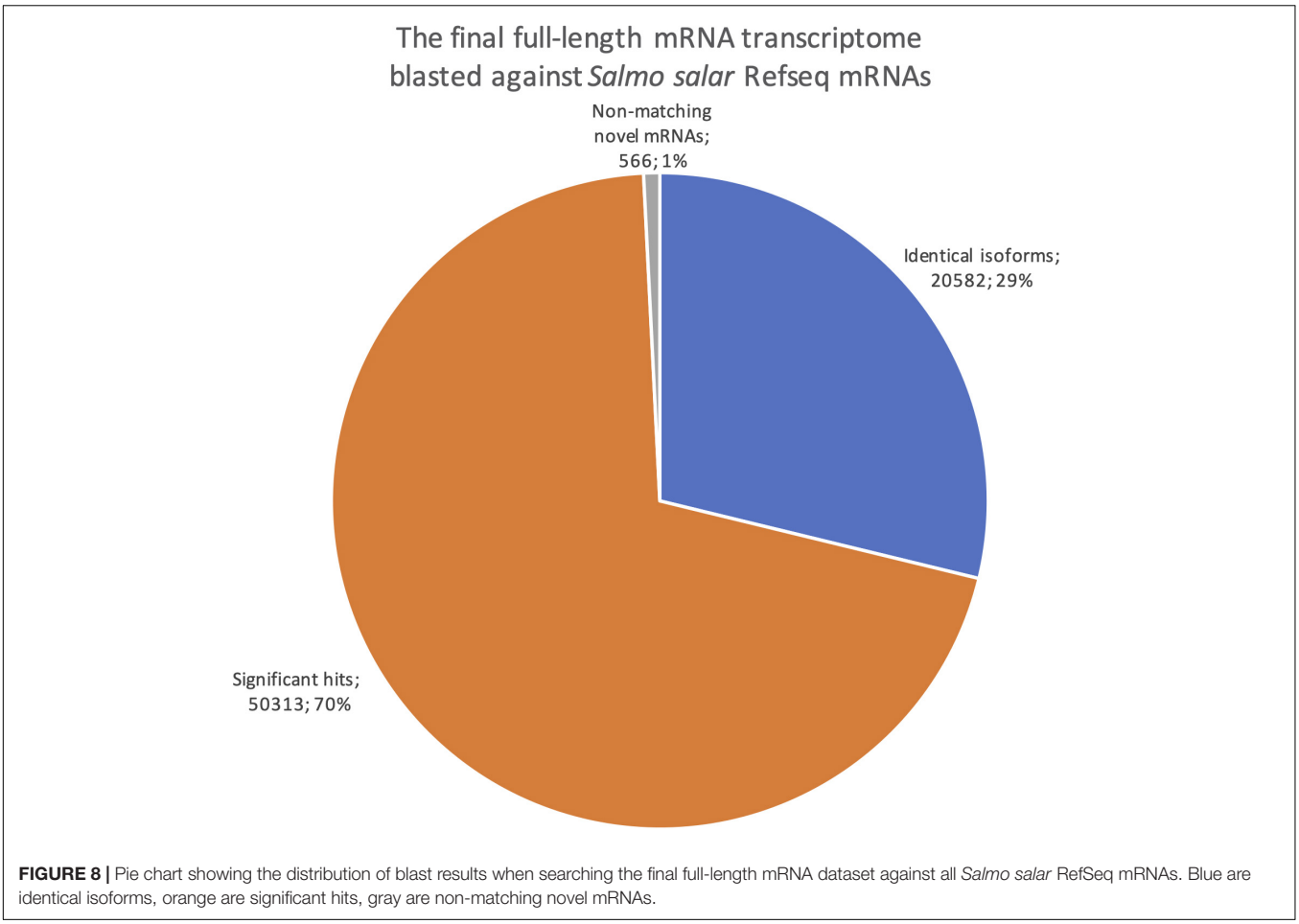


FIGURE 8 | Pie chart showing the distribution of blast results when searching the final full-length mRNA dataset against all *Salmo salar* RefSeq mRNAs. Blue are identical isoforms, orange are significant hits, gray are non-matching novel mRNAs.

were successfully annotated with at least one GO term (GOs > 0, **Figure 10**). The remaining 18% of the transcripts with no GO terms were distributed into ones with significant blast hits, but to proteins with no GO terms in the Gene Ontology database (9%), while the other half did not have any significant hits in the RefSeq protein database. Instead, they were supported as protein-coding by their CDS length and the support in Pfam. The functional annotation, including the GO terms and gene symbols are included in **Supplementary File 1**. The distribution of the number of GO terms assigned to each sequence showed that 50% of the transcripts in the dataset were assigned between 2 and 5 GO terms, while 14% were assigned even more. Together there was a solid level of functional annotation for about two thirds of our dataset. A total of 46 769 of these were also annotated with specific gene symbols based on the Blastp results from OmicsBox.

Ten percent of all FLNCs in the dataset were from only 17 mRNAs (listed in **Supplementary File 5**). This demonstrated that there were a few highly expressed protein-coding transcripts in the final mRNA transcriptome. The single most abundant transcript, *alb1*, constituted 3,8% of all FLNCs on its own, and the top five most abundant transcripts represented 6.5% of all FLNCs. Serum albumin (*alb1*) and the two other top expressed genes (*fgg* and *itih3*) were all encoding secretory proteins from liver tissues. Other highly expressed genes like two splice variants of actin (*actb*), selenoprotein P (*SelP*), and apolipoprotein Eb (*apoeb*) were expressed in all tissues. A surprising finding was that one of the highly expressed transcripts annotated as complement factor H-like (*cfhr5*) did not map to the Atlantic salmon genome sequence, but was a FSM in the *Salmo trutta* genome. Furthermore, two other highly expressed transcripts (prothrombin-like and transferrin A-like) were annotated as fusion-products by SQANTI. Again, this is likely due to incorrect genome-annotation rather than due to sequence artifacts, given the high number of FLNCs supporting these transcripts in several tissues.

The final mRNA transcripts were from three organs: liver, gills, and head-kidney. Several of the highly expressed transcripts were only present in liver samples, indicating that the transcriptome pipeline could identify transcripts expressed in an organ specific manner. Applying a conservative approach to identify such organ specific transcripts, we searched the full-length mRNA transcriptome for transcripts that were expressed in at least three samples from one organ, while they were absent in samples from either of the other two organs. This revealed that there were 2717 transcripts expressed in gills only, either from genes expressed only in gills (1811) or splice-variants expressed only in gills (906). In liver there were 1784 transcripts, of which 1113 were from liver-specific genes and 671 were liver-specific

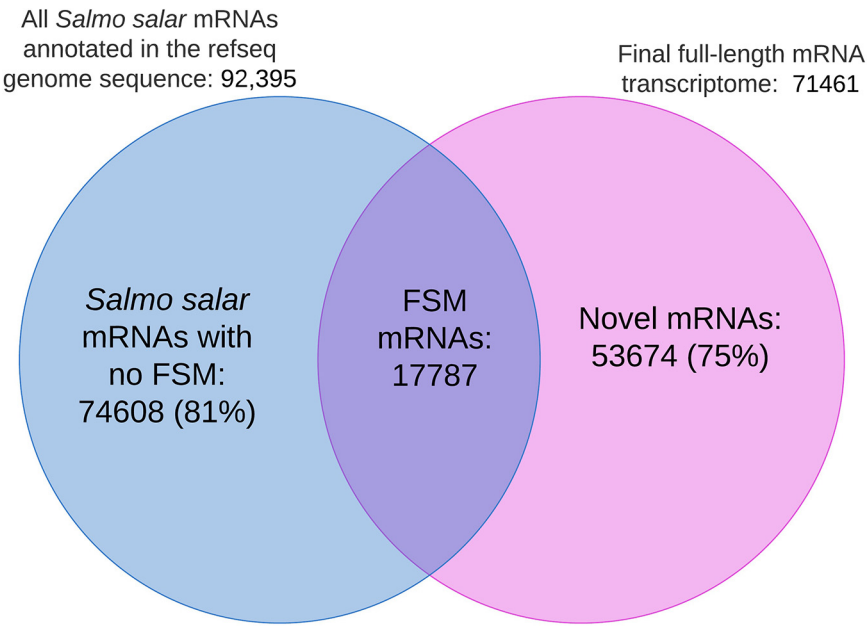


FIGURE 9 | Venn diagram illustrating the number of identical isoforms (Full Slice-Match) shared among the mRNAs in the genome annotation and the final full-length mRNA transcriptome. No FSM represents isoforms in the genome annotation with no identical match in the final full-length mRNA transcriptome. Novel mRNAs refers to transcript isoforms in the final mRNA dataset without an identical match to the sequences annotated to the *Salmo Salar* genome assembly.

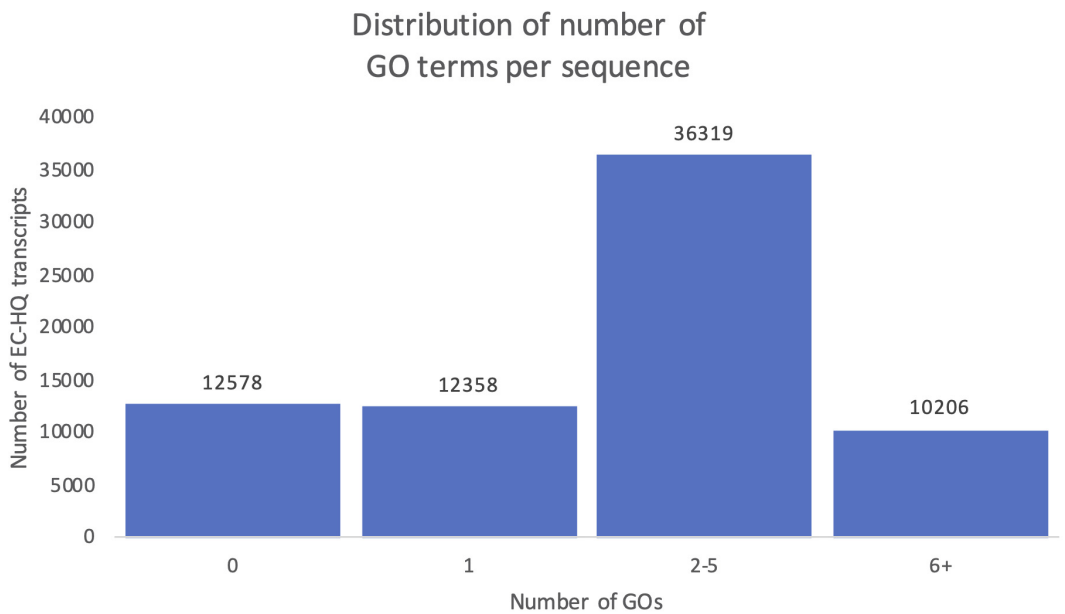


FIGURE 10 | Distribution of number of predicted Gene Ontology terms. Each column shows the number of transcripts falling within the given interval of Gene Ontology terms identified for the transcript.

splice-variants. In head-kidney there were 1757 transcripts, 700 from head-kidney specific genes and 1057 were head-kidney specific splice-variants. **Figures 11–13** show the distribution of the most specific common Biological Process GO terms (see Materials and Methods "Functional Annotation") for these three organ-specific groups of mRNAs. Each GO term indicates biological processes that are enriched among the organ-specific transcripts. The GO-terms in gill transcripts clearly pointed toward gill specific functions, such as ion transport and cell surface receptor signaling pathway, genes which take part in osmoregulation. Transcripts annotated as playing a role in system development (e.g., the development of specific tissues types

and regulation of DNA transcription) were also specifically expressed in the gill samples (**Figure 11**). The head-kidney in teleosts consists of several tissues with different distinct functions such as excretion, steroid biosynthesis, and immune response. Among the transcripts specifically expressed in head-kidney were those involved in biosynthesis of macromolecules, aromatic- and nitrogen compounds. Many were also annotated as involved in organelle organization and transport (**Figure 12**). There were 132 immune function and immune response genes expressed only in head-kidney, although not recognized as a uniform group of immune genes by the GO methods applied here (most specific common GO terms). Some examples of such transcripts include VIG2, different INFs, chemokines and toll-like receptors. The transcripts in liver (**Figure 13**), showed GO terms related to processes like metabolism, biosynthesis, and blood coagulation (e.g., the highly expressed *fgg* and *itih3*). Again, the transcripts exclusively expressed in this organ were, as expected, among those coding for proteins associated with liver function. Taken together, the results here showed the potential of our transcriptome pipeline to identify genes and splice variants that have particular organ-specific functions.

DISCUSSION

The Benefits of Applying Hybrid Error Correction of Single Molecule Long-Reads in Transcriptome Sequencing

Our pipeline aimed to generate a full-length mRNA transcriptome with reference level sequence accuracy from a variety of organ samples by processing PacBio long-reads that were hybrid error-corrected with Illumina paired-end reads from the same samples. Similar approaches have been used to generate high quality transcriptomes in other species (Feng et al., 2019; Puglia et al., 2020), but this is the first of its kind in Atlantic salmon. The strategy and the main functions of the pipeline is illustrated in **Figures 1–3**. First, the initial generation of consensus sequences from single molecules, removal of artifacts, and conservative clustering were achieved by use of the SMRTLink package IsoSeq3 (PacBio, 2020; **Figure 1**). This processing corrects for much of the high raw error rate associated with PacBio sequencing [estimated to be 11–14% (Roberts et al., 2013)]. The output is sequences (termed HQ reads) that are supported by at least two single sequenced molecules with a predicted accuracy of at least 99%.

The initial examination of the HQ reads in the materials indicated there were many cases of frame shift errors leading to incorrect CDS's or premature stop codons (data not shown). While we could not conclude for certain that these were sequencing errors in all cases, such errors were not unexpected given a sequence accuracy of 99% and the fact that PacBio reads are prone to numerical errors in homopolymers (Tedersoo et al., 2018). This accuracy would not be sufficient for our purpose, as we aimed to generate transcript sequences with a quality equally as high as (or better than) the Atlantic salmon mRNAs in RefSeq. Applying an approach where the long-reads were error-corrected using short-read sequencing data followed by filtering of sequences not supported by both datasets seemed to be the better solution (Au et al., 2012). The error correction approach takes advantage of the superior performance of the Pac Bio platform to identify structural isoforms (and differ between very similar paralogs) by long-read single molecule sequencing (Liang et al., 2016) while the accuracy is expected to be greatly increased due to a much higher read depth and phred quality contributed by the shorter reads. Also, the kind of errors most frequently acquired from the two platforms are not the same, and the probability of acquiring the same contaminating sequence from genomic DNA or other fusion artifacts when processing the same sample in two independent cDNA synthesis and different library prep methods is small. Together, transcript sequences generated from separate processing methods of sample RNA likely reduced the probability of retaining sequences with errors generated in each of the two sample processing pipelines in the final filtered EC-HQ reads (**Figure 2**).

A previous study (Sahraeian et al., 2017) comparing a variety of tools for RNA-seq analysis identified LoRDEC (Salmela and Rivals, 2014) as an efficient and accurate tool for hybrid error correction, and LoRDEC was successfully implemented in our pipeline. We applied a filtering of EC-HQ reads that assured that long-reads with internal gaps not supported by the shorter reads were removed. Together, the error correction approach with a very solid short-read support and the additional lower threshold for FLNC support assured that the final *de novo* transcriptome sequences had, in agreement with findings in similar studies (Au et al., 2012), an accuracy comparable with other Atlantic salmon transcript reference sources.

This pipeline focused on characterization of protein coding RNAs. The RepeatMasker software package was therefore implemented in our pipeline to ensure the final sequences would not contain transcripts from long interspersed repeats. A third of the sequences were identified as some kind of long interspersed repeat transcripts. This was a surprisingly large proportion of all transcripts. Future transcriptome projects in Atlantic salmon would benefit greatly from removing such transcripts prior to cDNA synthesis and library prep (Zhulidov, 2004). Also, after filtering with RepeatMasker, a substantial proportion (37%) of the unique EC-HQ reads were still not classified as protein coding transcripts. This revealed that our pipeline likely identified thousands of long non-coding RNAs (lncRNA). Characterization of lncRNAs by long-read approaches has emerged as the gold standard for studies of lncRNA (Wan et al., 2019), and characterization of the Atlantic salmon lncRNAs from these materials is now ongoing in a parallel project (manuscript in prep).

The Full-Length mRNA Transcriptome Substantially Increased Number of Isoforms

Whether comparing our data with the genome sequence annotation (SQANTI3) or the blast analysis against the Atlantic salmon RefSeq mRNAs, about 70% of the final mRNA

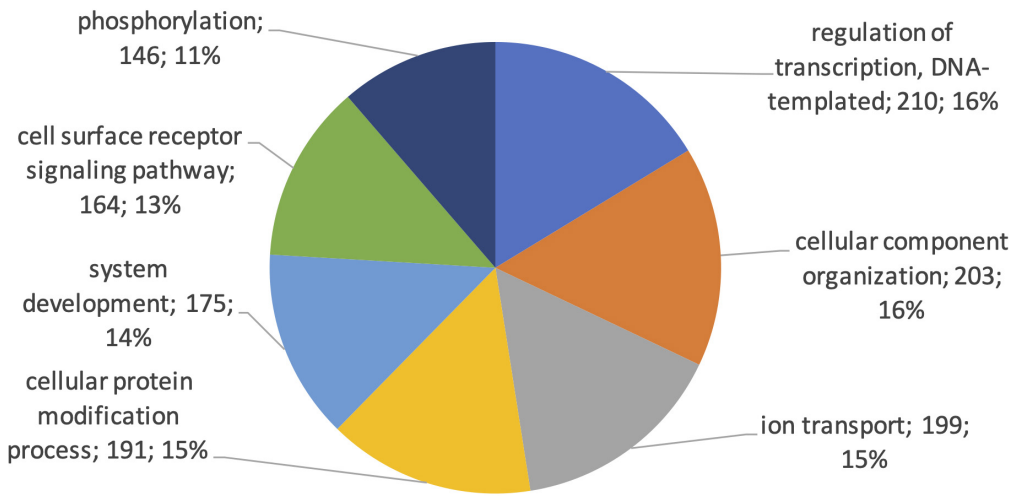


FIGURE 11 | Multilevel Gene Ontology chart, gill. The pie chart shows the most specific Gene Ontology terms occurring in at least 144 gill-specific transcripts in a non-redundant way (see also Materials and Methods "Functional Annotation").

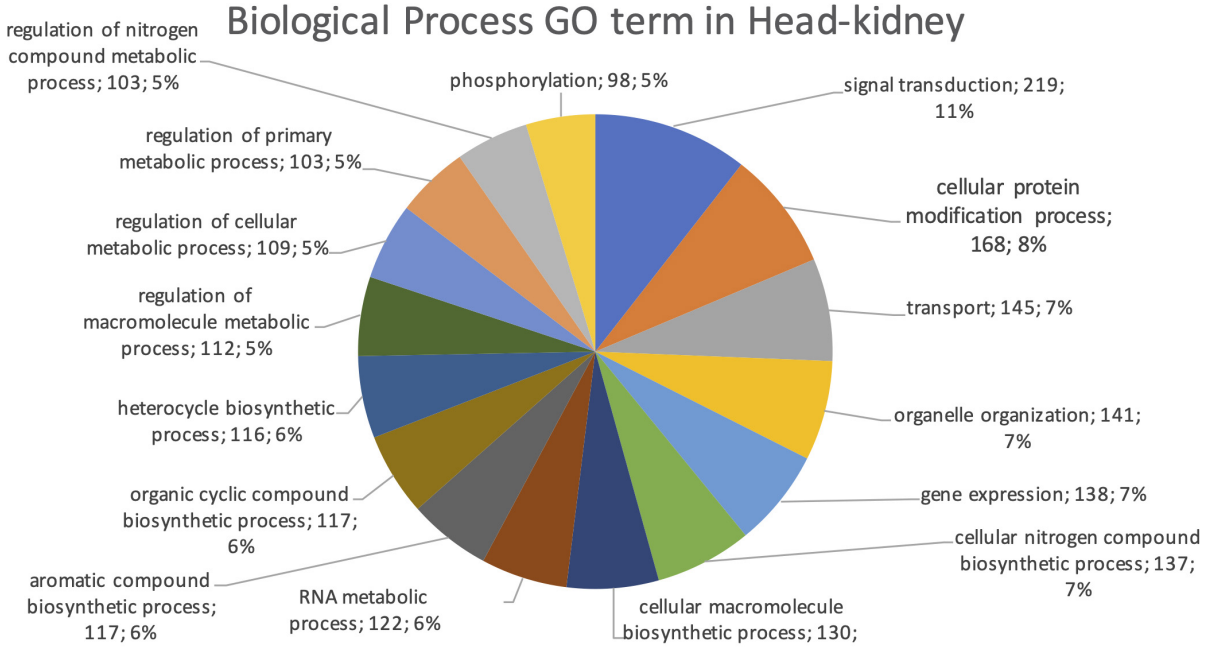


FIGURE 12 | Multilevel Gene Ontology chart, head-kidney. The pie chart shows the most specific Gene Ontology terms occurring in at least 98 head-kidney-specific transcripts in a non-redundant way (see also Materials and Methods "Functional Annotation").

transcriptome are novel isoforms. These novel isoforms are either splice variants, (each locus had on average three splice variants) or paralogs. This showed that our long-read transcriptome sequencing approach led to a substantial increase in number of Atlantic salmon transcript isoforms. Such high success rate in discovery of novel isoforms agrees well with findings from similar studies (Zhang et al., 2019). A cutoff of 5 supporting FLNCs was implemented to filter the remaining transcript categories.

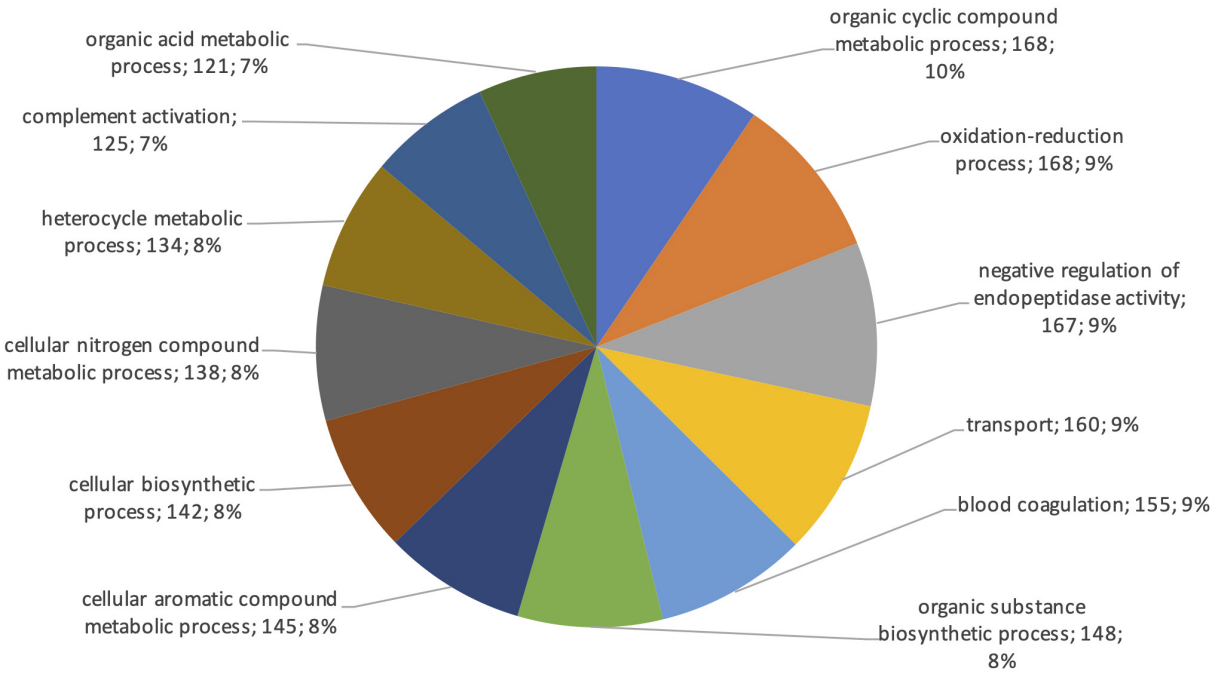


FIGURE 13 | Multilevel Gene Ontology chart, liver. The pie chart shows the most specific Gene Ontology terms occurring in at least 116 gill-specific transcripts in a non-redundant way (see also Materials and Methods "Functional Annotation").

Although the standard cutoff recommended by the developer is 10 FLNCs (Tseng, 2020a), other recent studies argue that 5 FLNCs is enough to support categories like fusion transcripts (Nattestad et al., 2018). We concluded that the hybrid error correction step with a minimum of three supporting Illumina reads across the entirety of the sequence provided the additional supporting evidence needed to accept the remaining SQANTI categories and the Cogent transcripts when supported by 5 FLNCs. Although these were categorized as dubious mRNA transcripts by SQANTI3 or not mapped at all, we find it likely that these are also true full-length mRNAs, but that they are not correctly annotated in the current genome sequence.

The Long-Read Transcriptome as a Reference to Study Expression of Splice Variants and Paralogs From Organs or Particular Conditions

A large fraction of the contributing FLNCs in this study belonged to a few transcripts with extremely high levels of expression (10% of FLNC reads were from the 17 most abundant transcripts). In future projects aiming to characterize all full-length mRNAs in a sample material, removal of such transcripts (as well as all interspersed repeats) would greatly increase the likelihood of identifying the more rarely expressed transcripts (Zhulidov, 2004).

The sequencing depth applying one Sequel II cell was approximately eight times higher than from one Sequel I cell. This is in agreement with other studies (Castaño et al., 2020; Lang et al., 2020). Combining the high read depth from Sequel II with normalization methods to remove abundant transcripts one would expect most transcripts expressed in a sample to be detected. The results from this study demonstrated that our pipeline had the ability to identify a large number of transcripts expressed uniquely in each of the three organ types included in the materials. The following functional annotation also showed that the most common GO terms annotated with the transcripts were largely consistent with the function of those organs. Furthermore, any materials investigated by this long-read approach may not only identify genes expressed uniquely, but also uniquely expressed splice variants. This was also demonstrated in the group of transcripts uniquely expressed in a single organ.

Full-length transcriptomes have a range of useful applications (Oikonomopoulos et al., 2020). Among those mentioned, we propose that our high-quality full-length transcriptomes may serve as references in expression analyses. The Atlantic salmon genome sequence may be less suitable as such a reference as a very large proportion of the transcripts in this study did not align properly. Differing between splice variants or paralogs by aligning short-reads to the genome sequence would be prone to error (if not impossible). Instead, the error-corrected long-read

transcriptome representing unique well-characterized transcript variants could be applied as a reference to which transcript sequences from short-read sequencing platforms, that provide greater read depth at affordable costs, could be aligned, counted and analyzed by tools like DESeq2 (Love et al., 2014). The added advantage of such analysis is that they would simultaneously detect SNP-variation. The UTRs are a rich source for such variation (Andreassen et al., 2010) and such mapping has the potential to reveal allele specific transcription, be applied to discover QTLs and even reveal causative variation leading to phenotypic differences in groups compared.

In conclusion, the hybrid corrected long-read pipeline employed here successfully generated high-quality full-length mRNA transcripts. The long-read approach led to the detection of novel splice variants and validated a quarter of all predicted Atlantic salmon mRNAs by transcripts originating from single molecule sequenced long-reads. Comprised solely of mRNAs with complete CDSs, more than 80% were assigned GO terms, and thousands of genes or splice variants from genes expressed in an organ specific manner were identifies. This full-length transcriptome will be an important resource for functional genomics in salmon aquaculture research.

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 National Research Authority in Norway (NARA).

AUTHOR CONTRIBUTIONS

SR, RA, BH, and T-KØ contributed to conceptualization and final editing of the manuscript. SR and RA contributed to analysis pipeline design, formal analysis, and wrote, reviewed and edited the manuscript. SR contributed to pipeline implementation, additional programming, and writing the original draft. RA, BH, and T-KØ contributed to resources. RA supervised and acquired funding. RA and BH contributed to project administration. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by Norwegian Research Council Grant to RA (280839/E40).

ACKNOWLEDGMENTS

We would like to thank Simon Weli at the Norwegian Veterinary Institute for providing the SAV challenge and control tissues used in this study. We would like to thank Tina Visnovska at the Bioinformatics Core Facility, University of Oslo for help with troubleshooting, technical assistance and advice while developing the pipeline. We are also grateful to Elizabeth Tseng at Pacific Biosciences for help and fruitful discussions during the development of our analysis pipeline, and for permission to use **Figure 1**.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.656334/full#supplementary-material

REFERENCES

Abdelrahman, H., ElHady, M., Alcivar-Warren, A., Allen, S., Al-Tobasei, R., Bao, L., et al. (2017). Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research. *BMC Genomics* 18:191. doi: 10.1186/s12864-017-3557-1

Adzhubei, A. A., Vlasova, A. V., Hagen-Larsen, H., Ruden, T. A., Laerdahl, J. K., and Hoyheim, B. (2007). Annotated expressed sequence tags (ESTs) from pre-smolt Atlantic salmon (*Salmo salar*) in a searchable data resource. *BMC Genomics* 8:209. doi: 10.1186/1471-2164-8-209

Allendorf, F. W., and Thorgaard, G. H. (1984). "Tetraploidy and the evolution of salmonid fishes," in *Evolutionary Genetics of Fishes*. Monographs in Evolutionary Biology, ed. B. J. Turner (Boston, MA: Springer), 1–53. doi: 10.1007/978-1-4684-4652-4_1

Andreassen, R., Lunner, S., and Hoyheim, B. (2009). Characterization of full-length sequenced cDNA inserts (FLIcs) from Atlantic salmon (*Salmo salar*). *BMC Genomics* 10:502. doi: 10.1186/1471-2164-10-502

Andreassen, R., Lunner, S., and Hoyheim, B. (2010). Targeted SNP discovery in Atlantic salmon (*Salmo salar*) genes using a 3'UTR-primed SNP detection approach. *BMC Genomics* 11:706. doi: 10.1186/1471-2164-11-706

Andreassen, R., Woldemariam, N. T., Egeland, I. O., Agafonov, O., Sindre, H., and Hoyheim, B. (2017). Identification of differentially expressed Atlantic salmon miRNAs responding to salmonid alphavirus (SAV) infection. *BMC Genomics* 18:349. doi: 10.1186/s12864-017-3741-3

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 2020)

Au, K. F., Underwood, J. G., Lee, L., and Wong, W. H. (2012). Improving PacBio long read accuracy by short read alignment. *PLoS One* 7:e46679. doi: 10.1371/journal.pone.0046679

Bernhardt, L. V., Myrmel, M., Lillehaug, A., Qviller, L., and Chioma Weli, S. (2021). Filtration, concentration and detection of salmonid alphavirus in seawater during a post-smolt salmon (*Salmo salar*) cohabitant challenge. *Dis. Aquat. Org.* 144, 61–73. doi: 10.3354/dao03572

BioBam (2019). *OmicsBox – Bioinformatics Made Easy [Online]*. Valencia: BioBam Bioinformatics

Bjørgen, H., and Koppang, E. O. (2021). Anatomy of teleost fish immune structures and organs. *Immunogenetics* 73, 53–63. doi: 10.1007/s00251-020-01196-0

Campbell, M. A., Hale, M. C., McKinney, G. J., Nichols, K. M., and Pearse, D. E. (2019). Long-term conservation of ohnologs through partial tetrasomy following whole-genome duplication in salmonidae. *G3 (Bethesda)* 9, 2017–2028. doi: 10.1534/g3.119.400070

Castaño, C., Berlin, A., Brandström Durling, M., Ihrmark, K., Lindahl, B. D., Stenlid, J., et al. (2020). Optimized metabarcoding with Pacific biosciences

enables semi-quantitative analysis of fungal communities. *New Phytol.* 228, 1149–1158. doi: 10.1111/nph.16731

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. doi: 10.1186/s13059-016-0881-8

FAO (2018). *The State of World Fisheries and Aquaculture 2018*. Rome: Food and Agriculture Organization of the United Nations.

Feng, S., Xu, M., Liu, F., Cui, C., and Zhou, B. (2019). Reconstruction of the full-length transcriptome atlas using PacBio Iso-seq provides insight into the alternative splicing in *Gossypium australe*. *BMC Plant Biol.* 19:365. doi: 10.1186/s12870-019-1968-7

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Giuffra, E., Tuggle, C. K., and Consortium, F. (2019). Functional Annotation of Animal Genomes (FAANG): current achievements and roadmap. *Annu. Rev. Anim. Biosci.* 7, 65–88. doi: 10.1146/annurev-animal-020518-114913

Gordon, S. P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., et al. (2015). Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One* 10:e0132628. doi: 10.1371/journal.pone.0132628

Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435. doi: 10.1093/nar/gkn176

Haas, B., and Papanicolaou, A. (2015). *TransDecoder 5.5.0 [Online]*. Available online at: https://github.com/TransDecoder/TransDecoder/wiki (accessed 2019)

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Hagen-Larsen, H., Laerdahl, J. K., Panitz, F., Adzhubei, A., and Hoyheim, B. (2005). An EST-based approach for identifying genes expressed in the intestine and gills of pre-smolt Atlantic salmon (*Salmo salar*). *BMC Genomics* 6:171. doi: 10.1186/1471-2164-6-171

Hjeltnes, B., Bang-Jensen, B., Bornø, G., Haukaas, A., and Walde, C. S. (eds) (2019). *The Health Situation in Norwegian Aquaculture 2018*. Oslo: Norwegian Veterinary Institute.

Hoar, W. S. (1988). 4 the physiology of smolting salmonids. *Physiol. Dev. Fish Viviparity Posthatching Juveniles*. 11, 275–343. doi: 10.1016/s1546-5098(08)60216-2

Houston, R. D., and Macqueen, D. J. (2019). Atlantic salmon (*Salmo salarL.*) genetics in the 21st century: taking leaps forward in aquaculture and biological understanding. *Anim. Genet.* 50, 3–14. doi: 10.1111/age.12748

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., et al. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Res.* 44, D81–D89. doi: 10.1093/nar/gkv1272

Koop, B. F., von Schalburg, K. R., Leong, J., Walker, N., Lieph, R., Cooper, G. A., et al. (2008). A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics* 9:545. doi: 10.1186/1471-2164-9-545

Lang, D., Zhang, S., Ren, P., Liang, F., Sun, Z., Meng, G., et al. (2020). Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *GigaScience* 9:giaa123. doi: 10.1093/gigascience/giaa123

Leong, J. S., Jantzen, S. G., von Schalburg, K. R., Cooper, G. A., Messmer, A. M., Liao, N. Y., et al. (2010). Salmo salar and Esox lucius full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome. *BMC Genomics* 11:279. doi: 10.1186/1471-2164-11-279

Li, H., and Birol, I. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liang, M., Raley, C., Zheng, X., Kutty, G., Gogineni, E., Sherman, B. T., et al. (2016). Distinguishing highly similar gene isoforms with a clustering-based bioinformatics analysis of PacBio single-molecule long reads. *BioData Min.* 9:13. doi: 10.1186/s13040-016-0090-8

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533, 200–205. doi: 10.1038/nature17164

Liu, S., Zhang, Y., Zhou, Z., Waldbieser, G., Sun, F., Lu, J., et al. (2012). Efficient assembly and annotation of the transcriptome of catfish by RNA-Seq analysis of a doubled haploid homozygote. *BMC Genomics* 13:595. doi: 10.1186/1471-2164-13-595

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Macqueen, D. J., and Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. Biol. Sci.* 281:20132881. doi: 10.1098/rspb.2013.2881

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:10. doi: 10.14806/ej.17.1.200

McLoughlin, M. F., and Graham, D. A. (2007). Alphavirus infections in salmonids - a review. *J. Fish Dis.* 30, 511–531. doi: 10.1111/j.1365-2761.2007.00848.x

Nattestad, M., Goodwin, S., Ng, K., Baslan, T., Sedlazeck, F. J., Rescheneder, P., et al. (2018). Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.* 28, 1126–1135. doi: 10.1101/gr.231100.117

NCBI (2015). *NCBI Salmo salar Annotation Release 100 Assemblies Report [Online]*. Available online at: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Salmo_salar/100/ (accessed June 26th 2020)

Oikonomopoulos, S., Bayega, A., Fahiminiya, S., Djambazian, H., Berube, P., and Ragoussis, J. (2020). Methodologies for transcript profiling using long-read technologies. *Front. Genet.* 11:606. doi: 10.3389/fgene.2020.00606

PacBio (2020). *IsoSeq v3: Scalable De Novo Isoform Discovery [Online]*. Available online at: https://github.com/PacificBiosciences/IsoSeq (accessed 2020)

Puglia, G. D., Prjibelski, A. D., Vitale, D., Bushmanova, E., Schmid, K. J., and Raccuia, S. A. (2020). Hybrid transcriptome sequencing approach improved assembly and gene annotation in *Cynara cardunculus* (L.). *BMC Genomics* 21:317. doi: 10.1186/s12864-020-6670-5

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genom. Proteom. Bioinf.* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biol.* 14:405. doi: 10.1186/gb-2013-14-6-405

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., et al. (2017). Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat. Commun.* 8:59. doi: 10.1038/s41467-017-00050-4

Salmela, L., and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30, 3506–3514. doi: 10.1093/bioinformatics/btu538

Shwe, A., Ostbye, T. K., Krasnov, A., Ramberg, S., and Andreassen, R. (2020). Characterization of differentially expressed miRNAs and their predicted target transcripts during smoltification and adaptation to seawater in head kidney of atlantic salmon. *Genes (Basel)* 11:1059. doi: 10.3390/genes11091059

Smit, A., Hubley, R., and Green, P. (2013). *RepeatMasker Open-4.0. [Online]*. Available online at: http://repeatmasker.org/ (accessed 2018)

Taksdal, T., Jensen, B. B., Bockerman, I., McLoughlin, M. F., Hjortaas, M. J., Ramstad, A., et al. (2015). Mortality and weight loss of Atlantic salmon, Salmon salar L., experimentally infected with salmonid alphavirus subtype 2 and subtype 3 isolates from Norway. *J. Fish Dis.* 38, 1047–1061. doi: 10.1111/jfd.12312

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., et al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* 28, 396–411. doi: 10.1101/gr.222976.117

Tedersoo, L., Tooming-Klunderud, A., and Anslan, S. (2018). PacBio metabarcoding of fungi and other eukaryotes: errors, biases and perspectives. *New Phytol.* 217, 1370–1385. doi: 10.1111/nph.14776

Tseng, E. (2020a). *cDNA_Cupcake [Online]*. Available online at: https://github.com/Magdoll/cDNA_Cupcake (accessed 2020)

Tseng, E. (2020b). *Cogent: COding GENome Reconstruction Tool [Online]*. Available online at: https://github.com/Magdoll/Cogent (accessed 2020)

Wan, Y., Liu, X., Zheng, D., Wang, Y., Chen, H., Zhao, X., et al. (2019). Systematic identification of intergenic long-noncoding RNAs in mouse retinas using full-length isoform sequencing. *BMC Genomics* 20:559. doi: 10.1186/s12864-019-5903-y

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., et al. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7:11708. doi: 10.1038/ncomms11708

Woldemariam, N. T., Agafonov, O., Hoyheim, B., Houston, R. D., Taggart, J. B., and Andreassen, R. (2019). Expanding the miRNA repertoire in Atlantic salmon; discovery of IsomiRs and miRNAs highly expressed in different tissues and developmental stages. *Cells* 8:42. doi: 10.3390/cells8010042

Woldemariam, N. T., Agafonov, O., Sindre, H., Hoyheim, B., Houston, R. D., Robledo, D., et al. (2020). miRNAs predicted to regulate host anti-viral gene pathways in IPNV-challenged atlantic salmon fry are affected by viral load, and associated with the major IPN resistance QTL genotypes in late infection. *Front. Immunol.* 11:2113. doi: 10.3389/fimmu.2020.02113

Yanez, J. M., Houston, R. D., and Newman, S. (2014). Genetics and genomics of disease resistance in salmonid species. *Front. Genet.* 5:415. doi: 10.3389/fgene.2014.00415

Zhang, G., Sun, M., Wang, J., Lei, M., Li, C., Zhao, D., et al. (2019). PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. *Plant J.* 97, 296–305. doi: 10.1111/tpj.14120

Zhulidov, P. A. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* 32:e37. doi: 10.1093/nar/gnh031

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.

Copyright © 2021 Ramberg, Høyheim, Østbye and Andreassen. 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

METHODS
published: 28 April 2021
doi: 10.3389/fgene.2021.650228

Accessing Livestock Resources in Ensembl

Fergal J. Martin, Astrid Gall, Michal Szpak and Paul Flicek*

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, United Kingdom

Genome assembly is cheaper, more accurate and more automated than it has ever been. This is due to a combination of more cost-efficient chemistries, new sequencing technologies and better algorithms. The livestock community has been at the forefront of this new wave of genome assembly, generating some of the highest quality vertebrate genome sequences. Ensembl's goal is to add functional and comparative annotation to these genomes, through our gene annotation, genomic alignments, gene trees, regulatory, and variation data. We run computationally complex analyses in a high throughput and consistent manner to help accelerate downstream science. Our livestock resources are continuously growing in both breadth and depth. We annotate reference genome assemblies for newly sequenced species and regularly update annotation for existing genomes. We are the only major resource to support the annotation of breeds and other non-reference assemblies. We currently provide resources for 13 pig breeds, maternal and paternal haplotypes for hybrid cattle and various other non-reference or wild type assemblies for livestock species. Here, we describe the livestock data present in Ensembl and provide protocols for how to view data in our genome browser, download via it our FTP site, manipulate it via our tools and interact with it programmatically via our REST API.

Keywords: Ensembl, genome browser, annotation, tutorial, livestock, farmed animals, Ensembl VEP, REST API

OPEN ACCESS

Edited by:
Hans Cheng,
United States Department of Agriculture, United States

Reviewed by:
Derek M. Bickhart,
United States Department of Agriculture, United States
Christophe Klopp,
Institut National de la Recherche Agronomique de Toulouse, France

****Correspondence:***
Fergal J. Martin
fergal@ebi.ac.uk

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

Received: *06 January 2021*
Accepted: *18 March 2021*
Published: *28 April 2021*

Citation:
Martin FJ, Gall A, Szpak M and Flicek P (2021) Accessing Livestock Resources in Ensembl. Front. Genet. 12:650228. doi: 10.3389/fgene.2021.650228

INTRODUCTION

Efficient management of livestock resources is key to global food security. Livestock production represents the largest land use sector worldwide and employs almost a billion people globally (Hurst et al., 2005; Abu Hatab et al., 2019). Livestock production is critical to developing countries, as it acts both as a major source of income and a means to escape poverty (Otte and Upton, 2005) and as a backup food source in the case of crop failures (Kabubo-Mariara, 2009; Rota and Sperandini, 2009). As the world's population continues to grow, so too does the demand for livestock source foods (LSFs). LSFs and other animal products account for approximately one-third of human protein consumption (Popp et al., 2010). The average per capita meat consumption is projected to grow from 34 kg in 2015 to 49 kg in 2050 (Yawson et al., 2017). At the same time, there is increasing competition for the use of key resources such as land and water and a need to move to less carbon intensive LSF production, especially in the face of climate change (Thornton, 2010; Yawson et al., 2017). Over the past decade, genomics has emerged as a key tool in the effort to

create more efficient LSF production, particularly the use of genomic selection to improve breeding programs (Hayes et al., 2009; Christensen et al., 2012; Cleveland and Hickey, 2013).

The livestock community has been at the forefront of genomics in terms of generating high quality genome assemblies and accompanying transcriptomic data, which are key to generating detailed genome annotations and exploring genomic variation among populations. Species such as pig, chicken, cow, horse, sheep, goat, salmon and herring all have chromosome-level genome assemblies, suitable for detailed annotation and downstream analyses (Jiang et al., 2014; Lien et al., 2016; Bickhart et al., 2017; Warren et al., 2017; Kalbfleisch et al., 2019; Pettersson et al., 2019; Warr et al., 2020).

In addition to reference genomes, an increasing number of alternative genome assemblies are available for analysis. Several breed-specific genomes have been sequenced and assembled including a large number of pig breeds (Fang et al., 2012; Li et al., 2017; Warr et al., 2020), black Bengal goat (Siddiki et al., 2019), Korean chicken (Sohn et al., 2018), and three strains of common carp (Xu et al., 2019). Livestock species are often at the cutting edge of genome assembly, exemplified by the recent trio binning approach (Koren et al., 2018) used to fully separate the maternal and paternal genomes from a *Bos indicus* × *Bos taurus* hybrid individual (Low et al., 2020).

With this wealth of high-quality genome sequences, it is crucial that the resulting genome annotation is carried out and presented in a clear and consistent manner. Ensembl is a genomics resource built to provide genome annotation and enable consistent interpretation of genomic variation both within and across species (Howe et al., 2021). The mission of Ensembl is to accelerate downstream science by providing pre-computed analyses, powerful genome interpretation tools and numerous ways of interacting with data through our extensive infrastructure. The data include genome sequences, gene annotation (Aken et al., 2016), comparative analyses (Herrero et al., 2016), variation (Hunt et al., 2018), and regulatory data (Zerbino et al., 2015). Tools such as the Ensembl Variant Effect Predictor (VEP; McLaren et al., 2016) and BLAST/BLAT services allow further interrogation of both the genome sequences and their annotations. Numerous ways are available to interact with the data including our genome browser, FTP site, REST APIs and BioMart querying tool.

In this article we present protocols for interacting with livestock data in Ensembl. We will examine several different livestock species from a variety of perspectives. These include investigating a genomic region of the cow, exploring a gene in chicken, viewing comparative data across pig breeds and annotating variants in goat. For more large-scale analyses, we provide examples of how to programmatically access the data and download the associated annotation files. In summary, readers will get a thorough understanding of the livestock data held in Ensembl and how to work with it.

MATERIALS

Computer and Internet Connection.

An Internet browser: recent versions of Firefox, Chrome, Safari, and Microsoft Edge are supported.

For working with the REST API, examples are presented in Python 3.

METHODS

The following protocols use Ensembl release 101 (August 2020)[1]. There may be updates to interfaces or data if a more recent release is used.

Exploring Genes and Genomes

The most fundamental data in Ensembl are the genome sequences and the gene annotation for each species. In this section we will look initially at how to view and explore a region of the cow genome and then examine a gene in the chicken genome. This will form the basis for later explorations of comparative and variation data, which build on data held in the genes and genomes.

Browsing a Genome

Much of the annotation in Ensembl corresponds to an underlying genomic region. Becoming familiar with how to browse these regions is key to understanding the annotation available. The following protocol describes how to examine a region of the cow reference genome.

1. Getting started: The Ensembl genome browser can be searched using a variety of terms including gene names, genomic coordinates, variant IDs or phenotypes. Go to the Ensembl's homepage, www.ensembl.org, and locate two search boxes: one in the upper right corner, and another in the middle of the page. Both the main and the corner search box can be used to search all species. Additionally, you can also refine the main search by choosing species of interest from a drop-down list.
2. Finding a region: Type "cow 2:20721000-20826000" into either box and press the return key. A Location tab will open with a "Region in detail" view displaying the region of interest spanning the *HOXD* gene cluster involved in limb development (see **Figure 1**). You will find available location displays in the left-hand side menu with blue tool buttons below. The "Region in detail" page has three images, each more detailed than the last: (1) the chromosome view at the top, (2) the 1MB region around the region of interest in the middle, (3) the region of interest corresponding to the specified genomic coordinates at the bottom. The region of interest is indicated by a red box in all three images.
3. Getting help: To get page specific help, click on the question mark (?) button next to the "Region in detail" heading. A pop-up help window will open with instructions on how to navigate this page. You will find a description of the page with screenshots and a tutorial video, as well as links to FAQ, glossary and the Ensembl helpdesk.

[1] http://e101.ensembl.org

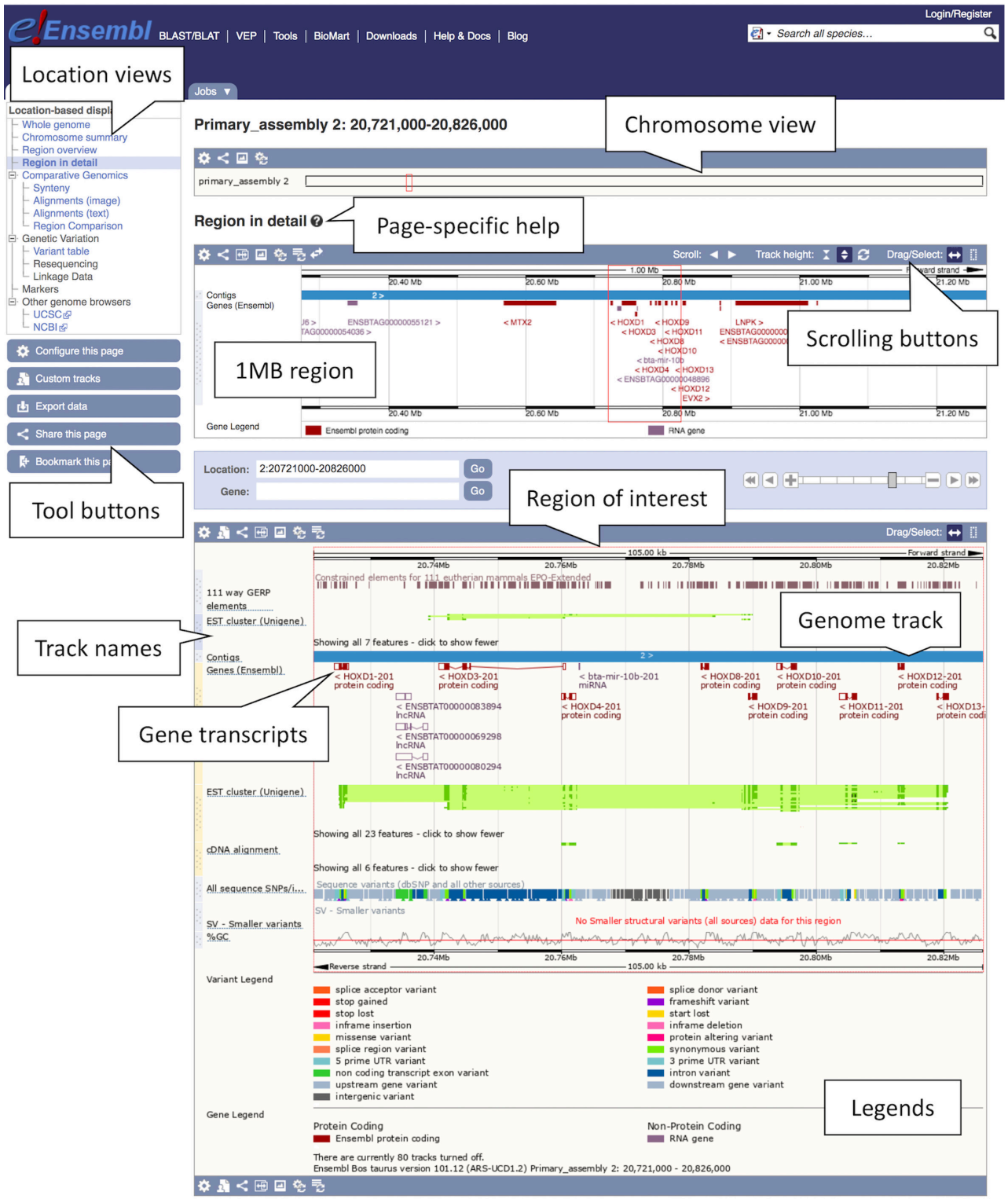


FIGURE 1 | Location view of the *HOXD* gene cluster in cow. In the figure, various tracks are displayed including the main gene track, SNP data, constrained elements and a GC content. Most tracks are disabled by default, tracks such as the tissue-specific short read alignments can be toggled on and off via the "Configure this page" option in the left-hand menu.

4. Navigating a region: There are several ways of navigating a genomic region. By clicking over the region and dragging the cursor, you can draw a box in all three images, which opens a pop-up menu with options to "Mark region" and "Jump to region." You can also scroll along the genome by using the "Scroll" arrow buttons in the middle image or by changing the mouse click mode to "Drag" (double headed arrow icon). The zoom scrollbar enables zooming in and out. Scroll along the genome in the middle image to change the current genomic location. As you scroll, the image below greys out and two blue buttons appear with options to "Update this image" or "Reset scrollable image."
5. Customising the view: The data in this view is organised in tracks plotted along the genome. You will find separate tracks for different data types such as genes, SNPs, structural variants or contigs representing the genome assembly. Click on the "Configure this page" button on the left to add more data to this view. A pop-up window with a menu listing all currently active tracks will open. You can find a list of all available tracks organised in different categories on the left with a search box above (with the text "Find a track"). Search for "Proteins from UniProtKB" and turn it on as "Labels." Click on the tick at the top left of the pop-up window or anywhere outside to save and close. You will find the protein track added to the view.
6. Exporting data: Click on the blue "Export data" button on the left to download data for this region.

A data export window will open with different output format options, including FASTA for sequences and various feature formats such as BED, CSV, TSV, GTF, and GFF3.

Exploring a Gene

Gene annotation is one of the most commonly used annotation types in Ensembl. It is a composite data type, representing underlying transcripts, exons, and protein products. In this section, we describe how to find and export information about the *SOX5* gene in the chicken genome.

1. Getting started: To search for a chicken gene, select "Chicken" from the species selector drop down list above the main search box on the Ensembl homepage, type the gene name, "*SOX5*," into the search box and click the "GO" button. A list of search results restricted to chicken and matching "*SOX5*" will be generated with the "*SOX5* (Chicken Gene)" at the top.
2. Studying a gene: Click on the "*SOX5* (Chicken Gene)" link to open the Gene tab.
   The Gene tab landing page contains summary information on *SOX5* including its Ensembl stable ID (ENSGALG00000032768), gene description, genomic position, and strand information, as well as the number of transcripts and an option to show them in the tabular format transcript table (see **Figure 2**). A graphical representation of all transcripts can be found at the bottom of the page. A number of gene-related displays providing additional data can be found in the left-hand menu. The gene overview at the top of this page is visible across all subsequent views.
3. Exploring a gene sequence: Click the "Sequence" display in the left-hand side menu.
   A sequence view of *SOX5* and its flanking region will be displayed in FASTA format with all exonic sequences in this region highlighted in peach colour. Exons of *SOX5* will be indicated by brown bold lettering on top of the peach highlight. You can manipulate this view and change display options by clicking the blue "Configure this page" button on the left. It will allow you to customise the length of the flanking sequence and show genetic variants in the sequence.
4. Downloading a gene sequence: The gene sequence can be downloaded by clicking the "Download sequence" button in the current "Sequence" view or by clicking the blue "Export data" button on the left in any other view. This will open a pop-up window with customisation options that allow to choose different sequence types, the length of the flanking region and the file format (FASTA or RTF).
5. Studying gene ontology (GO): You will find three GO categories under "Ontologies" in the left-hand menu: "GO: Cellular component," "GO: Molecular function" and "GO: Biological process." Browse all three views to learn about the gene function. GO terms describe the protein function using standardised vocabulary:

   (a) "GO: Biological process": what does it do? Example: "positive regulation of chondrocyte differentiation."
   (b) "GO: Molecular function": how does it do it? Example: "DNA binding."
   (c) "GO: Biological process": where is it located? Example: "nucleus."

   Each of the above GO categories lists terms associated with transcripts of the *SOX5* gene. The data are organised in a tabular format containing the GO term accession number, the corresponding description, evidence and annotation source, along with the associated transcript stable IDs. On exploring all three views, it emerges that *SOX5* gene encodes a nuclear transcription factor involved in the regulation of chondrogenesis.
6. Exploring external resources: Click "External references" in the left-hand side menu.
   Links to external databases such as Expression Atlas, NCBI, and WikiGene, as well as related RefSeq and UniProtKB/TrEMBL accession numbers can be found here.
7. Studying a transcript: Click the "Show transcript table" button at the top of the page and go to the Transcript tab by clicking the transcript stable ID link "ENSGALT00000105978.1" in the table. The transcript table is visible in any gene and transcript view. It lists all transcripts of the gene of interest, their corresponding name, stable ID, length, biotype and transcript flags indicating transcript quality. You can hover the cursor over the flags to find out more information. In this case the chosen transcript named "*SOX5*-201" has a flag

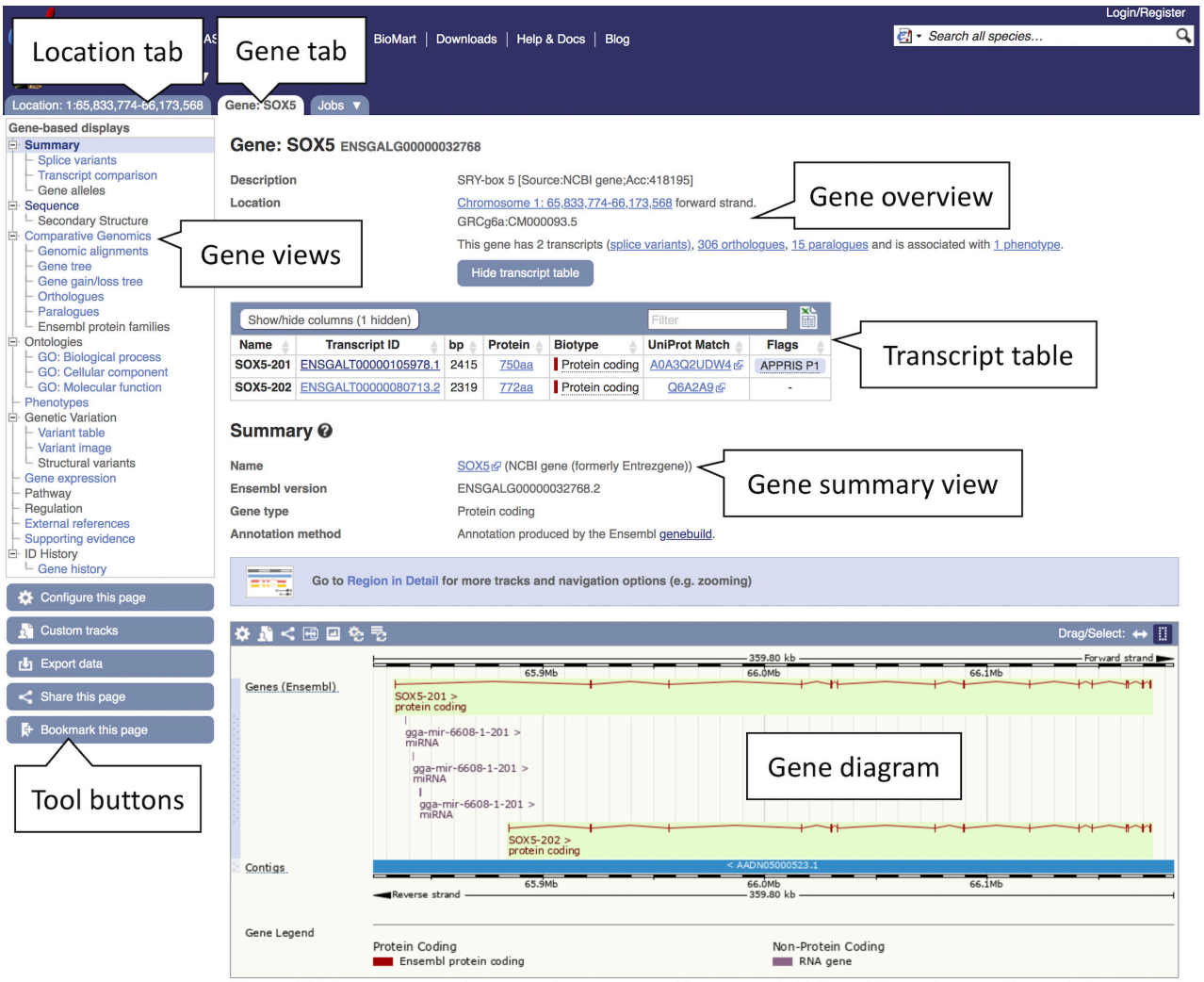


FIGURE 2 | Gene view of *SOX5*. The gene view tab shows a variety of information about the gene including details on the location, transcripts, orthologues, and paralogues. The main gene track is shown in the bottom panel with red blocks representing the exons and connecting red lines representing the introns. In this example a secondary track for liver transcriptomic data can be seen in blue. This and other tissue tracks are available via the "Configure this page" option in the left-hand menu.

"APPRIS P1." This means that it is predicted by the APPRIS database (Rodriguez et al., 2018) to be the most functionally important transcript of this gene based on protein structure, functional features and information from cross-species conservation. Similar to the Gene tab, the Transcript tab is also composed of several displays introducing different data types including "Sequence" views and "Protein Information."

8. Exploring exon sequences: Click "Exons" under "Sequence" in the left-hand side menu to see exon information in tabular format. The Exons view displays a table listing exons, their order in the transcript, genomic position, start and end phase, length, and sequence. Translated sequence is marked in blue, untranslated region (UTR) in orange, flanking sequence in green and introns in grey. This transcript has a 5′ UTR spanning the entire first exon and the beginning of the second exon. "Configure this page" will allow you to customise this view.

Comparing Genes and Genomes

Ensembl has a powerful comparative genomes infrastructure to deliver information about how genes and genomes relate to one another across our supported species. Here, we describe assessing comparative data, focusing on gene trees and whole genome alignments.

Examining Genes Trees and Orthologous Genes in Pig

A common way to assess the reliability of annotation of a gene, both in terms of structure and function, is to examine orthologous genes in other species. Genes that are present across a broad range of species, with high sequence similarity and

within syntenic regions, are likely to have equivalent functions. Understanding the level of conservation of a gene across species can assist with downstream inference and analysis. Here, we will examine a highly conserved gene, *FILIP1*, in pig and explore the associated gene tree as well as data for *FILIP1* orthologues in other species.

1. Getting started: Select "Pig" from the species selector drop down list above the main search box on the Ensembl homepage, type "*FILIP1*" into the search box and hit the "GO" button.
   A list of search results across all breeds will be generated. In the left-hand menu you will see options for filtering the search results. For species like pig, where there are alternative breeds/strains available, we have a defined reference. In general, for livestock species, the reference chosen is based on feedback from the community. In the case of pig, the current reference chosen by the community is the Sscrofa11.1 assembly of the Duroc breed. Try restricting the results to the *FILIP1* gene in the reference breed. First, in the left-hand menu under "Restrict breeds to" click on "Pig reference." Next, under "Restrict category to" click on "Gene." The initial results are now filtered to just genes in the reference breed matching the name "*FILIP1*." You will now have two results representing the *FILIP1* and *FILIP1L* genes. Click on the "*FILIP1* (Pig Gene, Breed: reference)" link on top to go to the Gene tab. This will present the gene view for the *FILIP1* gene in the reference pig genome.
2. Exploring a gene tree: In the left-hand side menu, click on the "Gene tree" display. An image showing a phylogenetic tree will be loaded. The current gene, marked in red, is shown in the context of homologous genes found across various clades including primates, rodents, birds, reptiles and even non-vertebrates such as *Caenorhabditis elegans* and *Drosophila melanogaster* (see **Figure 3**). Grey funnels indicate collapsed nodes, which can be expanded by clicking on them and selecting "expand this sub-tree" from the pop-up menu. A graphical representation of the protein alignment used to calculate this tree can be found on the right. Green colour indicates aligned sequences, while alignment gaps are shown in white. A consensus alignment is displayed for collapsed nodes, with two shades of green corresponding to the proportion of aligned collapsed sequences. Black vertical lines mark exon-intron boundaries.
3. Retrieving orthologues: Click "Orthologues" in the menu on the left. This will load the orthology table.
4. The table lists *FILIP1* orthologues found across a large range of species. To make it easier to examine the data click on the show details box beside "Laurasiatheria" in the species set list. This will restrict the data in the orthologue table below to just laurasiatherian species.
5. Examine the column headings in the orthologue table to see the types of information available in the table. Some columns have extra information that displays when the mouse cursor is left over them.
6. Scroll down to the orthologue for horse. The sequence identity of the pig *FILIP1* gene with the horse *FILIP1* gene is high. The query and target ID percentages, which represent how much of the pig *FILIP1* sequence matches the horse *FILIP1* sequence and vice-versa are both over 94%, indicating strong conservation. The gene order conservation score, which represents conserved orthology between the two nearest 5′ and 3′ genes flanking *FILIP1*, and whole genome alignment coverage are both 100. This implies strong gene order conservation and a high coverage pairwise alignment of the broader underlying genomic regions. As a result, the orthologue is listed as high confidence, as indicated by a "Yes" in the "High Confidence" column.
7. Exploring a protein alignment: Click on the "View Sequence Alignments" link to open a pop-up menu with options to view protein and cDNA alignment. Click "View Protein Alignment." An alignment of the gene of interest and its orthologue will be displayed in CLUSTAL W format. Click on the question mark button (?) next to the "Orthologue alignment" heading for more information on the conservation codes.

Viewing a Whole Genome Alignment of Pig Breeds

Ensembl provides a large number of pairwise and multiple whole genome alignments. Every species has a pairwise alignment against a reference species for its clade. The reference species for mammals, birds and fish are human, chicken, and zebrafish, respectively. For some species, additional pairwise alignments are generated. For example, rodent genomes are aligned against the mouse reference, while the pig reference has a pairwise alignment to the USMARC pig assembly. In addition to the pairwise alignments, various multiple whole genome alignments are available, including the 57 mammals and the 95 amniota vertebrates alignments.

Here, we will look at a multiple alignment generated for 13 pig breeds and three outgroup species: cow, horse, and sheep (texel). Using the *COL12A1* gene, we will see that this region is generally well conserved across the alignment, however the gene is truncated in the Tibetan breed. We will examine the alignment for potential explanations for this truncation.

1. Getting started: From the Ensembl frontpage type "pig 1: 90744429-90875118" into the search box and click the "GO" button. This will bring you directly to the genomic location of the *COL12A1* gene in the pig reference genome.
2. Using the left-hand side menu, click on "Alignments (text)" under "Comparative Genomics" to access available whole genome alignments for the pig reference genome.
3. Click on the "Select an alignment" button to see the alignment selector tool.
4. Choose "16 pig breeds EPO-Extended" from the multiple alignment category. This provides a multiple whole genome alignment of all pig genomes in Ensembl and the three outgroup species generated from the Enredo, Pecan, and Ortheus pipeline (Paten et al., 2008). A graphical representation of an expanded phylogenetic tree and

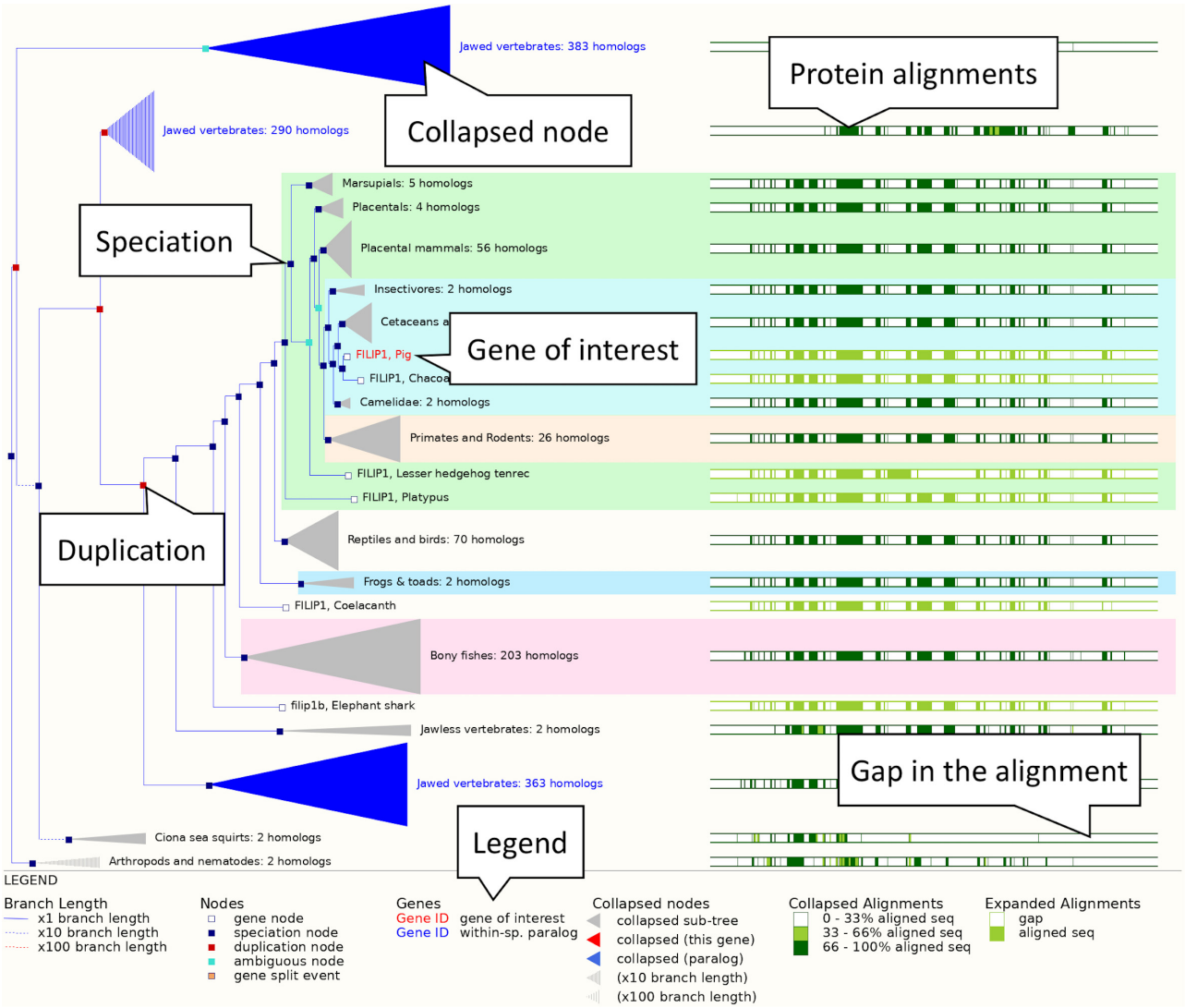


FIGURE 3 | The *FILIP1* gene tree. The pig FILIP1 gene is highlighted in red. Various other clades are collapsed into funnels to improve usability of the tree. Subtrees can be expanded and collapsed by clicking on the corresponding node (represented by a square on the tree) and selecting the appropriate option from the menu. The right-hand side shows a visual representation of alignment conservation of the corresponding protein sequences across the tree.

corresponding section of the whole genome alignment will load, followed by a list of the aligned regions and a preview of the sequence alignment.

5. Examine the list of aligned regions. The alignment from the reference to the USMARC assembly is represented by a single contiguous alignment block in the USMARC assembly, reflecting the high-quality and contiguity of the assembly. This is also true for the aligned regions of the three outgroup assemblies (cow, sheep, and horse). For other pig breeds, there are multiple alignment blocks due to the lower contiguity and completeness of the assemblies. In particular, the Tibetan and Wuzhishan alignments are fragmented across multiple genomic regions, implying the region is not correctly reconstructed in these assemblies.
6. Click on the "View an image of this alignment" link, located directly above the list of regions. This will load a more detailed view of the *COL12A1* gene structure across the aligned regions, as shown in **Figure 4**.
7. Examine the gene structure across the breeds. Note that the intron/exon structures are mostly well conserved across breeds. Scroll down to the Tibetan copy of the *COL12A1* gene. Note that the gene is heavily truncated in comparison to the other breeds. The alternating light and dark blue bands represent the boundaries of different alignment blocks and are labelled with the region of the genome each block comes from. For the Tibetan breed the annotated section of the gene lies on the AORO02005858.1 scaffold. The remainder of the alignment blocks are on the AORO02052718.1 scaffold. This provides strong evidence

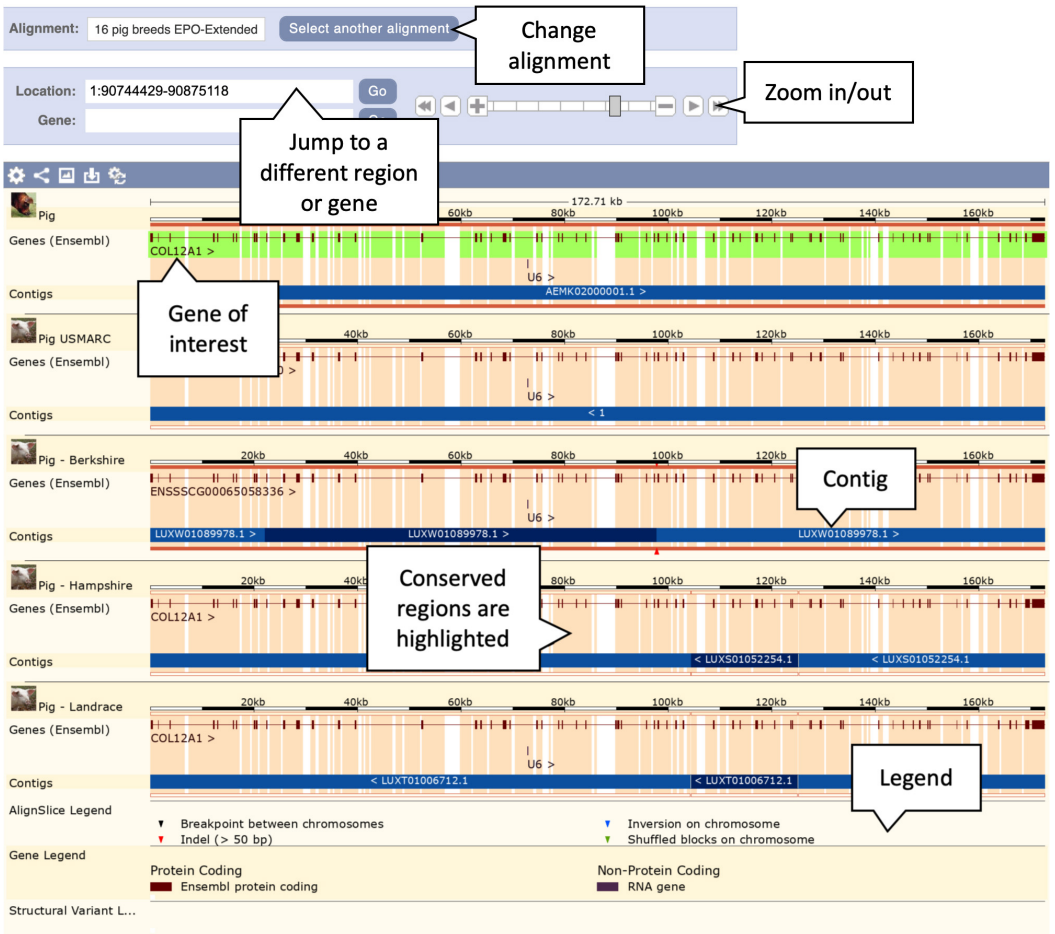


FIGURE 4 | Genomic alignment of pig breeds centred on the *COL12A* gene. This figure shows only the first five alignments including the reference pig at the top, the complete alignment contains all 13 pig breeds along with cow, horse, and sheep as outgroups. The pale orange highlighting in the background represents conservation. Exonic regions are naturally more conserved across the alignment.

that parts or all of these two scaffolds should have been joined in the Tibetan assembly. As a result of them not being joined the *COL12A1* gene is truncated at the end of the AORO02005858.1 scaffold, where the majority of the gene resides.

Viewing a Synteny Map of Pig Chromosome 6 to Human

Whole genome alignments can also be used to generate synteny maps between chromosomes of different species. These maps show genomic regions in which genes occur in the same order in two species. This view gives insight into how chromosomes have diverged between two species and any two species with pairwise whole genome alignments can be compared in this way. Here, we will describe viewing a synteny map between pig chromosome 6 and the corresponding regions in the human genome.

1. Getting started: From the Ensembl homepage click the dropdown box under the "All genomes" heading (the box will have "–Select a species–" by default).
2. This will produce a list of species grouped under headings, including major clades such as primates and rodents. Scroll down to "Laurasiatheria" and click on "Pig" to go to the species page for the reference pig genome
3. From the species page select "View karyotype." This will give the karyotype view of all chromosomes in the reference pig. Click anywhere within chromosome 6 on the image of the karyotype to see a pop-up window.
4. Select "Jump to region overview" in the pop-up window. This will bring you directly to the Location tab of the corresponding region on chromosome 6.
5. In the Location tab, select "Synteny" under the "Comparative Genomics" section in the left-hand side menu to bring up the synteny view for pig chromosome 6. You will now see an image of a synteny map between chromosome 6 in pig and the various chromosomes in human that it maps to (see **Figure 5**). The region you selected and the corresponding

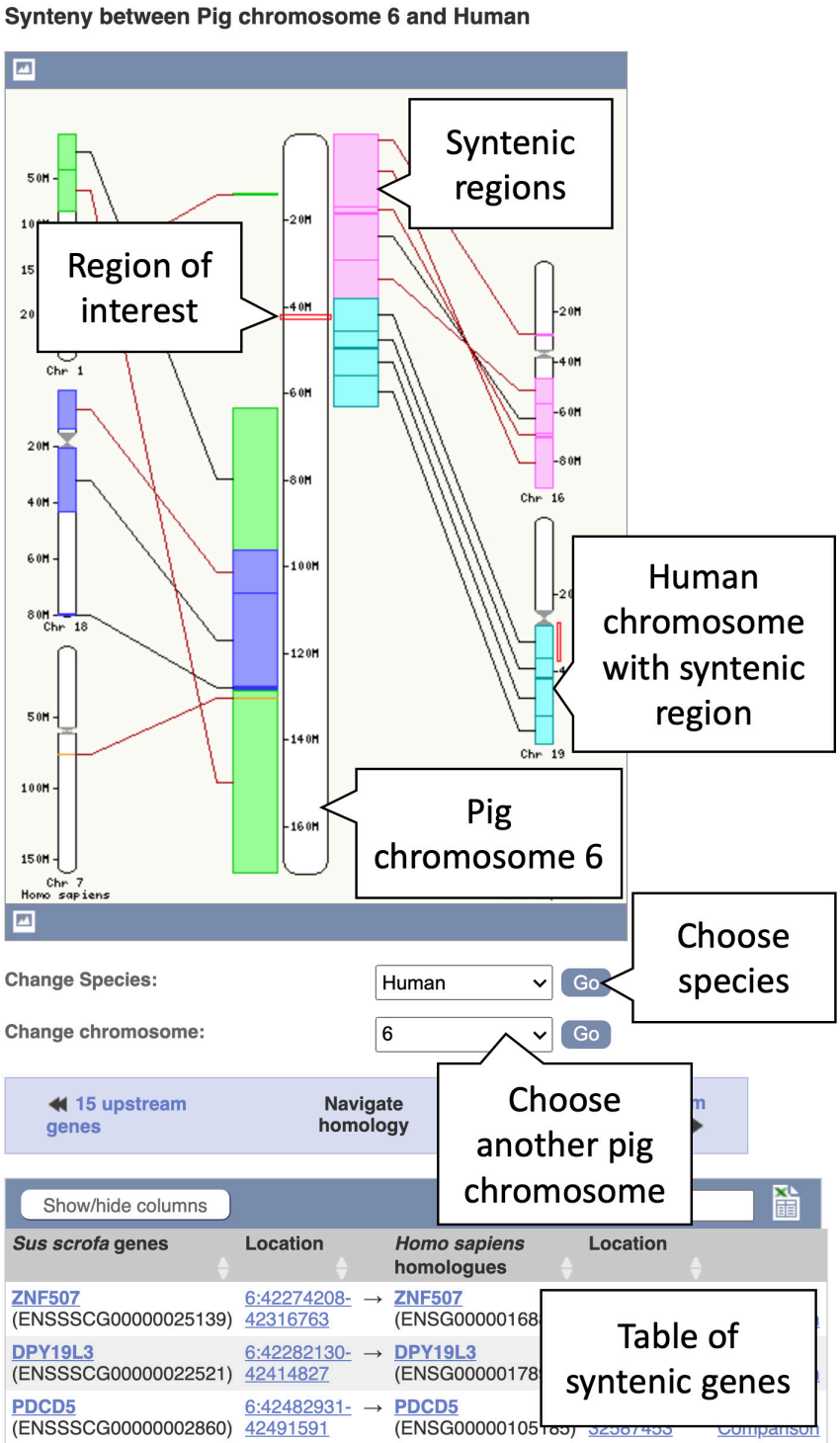


FIGURE 5 | A synteny map from pig chromosome 6 to human. Regions on pig chromosome 6 (located in the centre) are syntenic to multiple regions on human chromosomes 1, 7, 16, 18, and 19 (shown on the left- and right-hand sides). Syntenic blocks are coloured according to the chromosome numbers for human. The blocks are connected by lines, whereby black lines connect blocks with the same orientation and brown lines indicate blocks with the opposite orientation.

location in human are indicated by red boxes. Syntenic blocks are shown in different colours and connected by lines. You can change the chromosome or select species other than human, where a pairwise whole genome alignment is present, using the drop-down on the right.

Examining Genetic Variation and Annotating Variants

Several livestock species including chicken, goat, pig, and salmon have extensive variation data. Ensembl provides extensive views via the bowser and also analysis of variant data via the Ensembl Variant Predictor (VEP). In this section we will show how to browse and analyse variant data in the goat genome.

Exploring Variation Data

Variation data can be accessed in a number of ways through the browser including by selecting it in the configuration menu and by navigating to the "Genetic variation" section in the left-hand menu on the Location, Gene, and Transcript tab. There are also example entry points on each species page and support for searching variant identifiers from dbSNP and other databases. Here, we will examine different aspects of variation data including population frequencies, phylogenetic context, and consequences.

1. Getting started: Type "rs666529295" into the main or the upper right corner search box to search all species and hit return. The search result page will return two hits with "rs666529295 (Goat Variant, Breed: reference)" at the top, click on this link.
2. Studying a variant: You will be taken to the variant summary page containing variant overview information such as the most severe consequence, variant alleles as reference/alternative (here: "G/A"), highest population minor allele frequency (MAF), genomic location and strand.
3. Exploring variant allele frequencies: Click the "Population genetics" icon or the link in the left-hand menu to display allele frequencies from the NextGen Project (Alberto et al., 2018). The pie chart shows the allele frequency across all sequenced goat populations (see **Figure 6**). Click the " + " next to "Sub-populations" to reveal the allele frequencies in the included sub-populations.
4. Studying the phylogenetic context: Click "Phylogenetic context" to see the conservation of this variant and its flanking region across different taxa. Click "Select an alignment," then "Multiple." This should show a list of available multiple species alignments. Select "95 amniota vertebrates Mercator-Pecan" and click "Apply" to load the alignment. The resulting page displays a multiple alignment of 10 bp around the focus variant. In this case the variant is conserved across all taxa.
5. Transcript consequences: Click "Genes and regulation" under "Genomic context" in the menu to see which genes and transcripts are affected by this variant. In this case, we can see that a transcript in the *EDNRA* gene is affected with a consequence type of "Missense variant." The location in both the transcript and CDS sequence are shown, along with the codon and amino acid changed. The Sorting Intolerant From Tolerant (SIFT) pathogenicity prediction score (Kumar et al., 2009) is "0," indicating that the amino acid substitution is predicted to be deleterious to the function of the protein.
6. The Variant table: By clicking the Gene stable ID link "ENSCHIG00000019737," you can navigate directly to the Gene tab and be taken straight to the "Variant Table" display. This lists all variants in the Ensembl database that fall within the *EDNRA* gene (including a 5 kb flanking region). Data can be downloaded as a CSV file by clicking the Excel icon.
7. Studying a phenotype: From the left-hand menu in the Gene tab, click "Phenotypes" to explore the complete set of phenotypes, diseases and traits associated with the gene. This gene has been associated with "Coat colour, white spotting, *EDNRA*-related" according to OMIA (Lenffer et al., 2006). There are no variants for this gene currently associated directly with phenotypes. Toward the bottom of the page a list of phenotypes for orthologues of the gene is provided to help cross-species phenotypic comparison. Clicking on the link for any of the phenotypes listed on the page will provide a list of other loci present in the species that are associated with the same phenotype.

Annotating Variants With the Ensembl VEP

The Ensembl VEP (McLaren et al., 2016) classifies the impact of variants on genes, transcripts, and protein sequences and identifies known variants that match the input variants.

The Ensembl VEP is available as a web interface, a command line tool and through a REST API endpoint. The web interface is suitable for smaller amounts of data, while the command line tools is suitable for large-scale analysis and offers maximum flexibility, including the option to analyse variants for genomes that are not in Ensembl. Here, we'll look at the use of the point-and-click web interface to analyse six goat variants input in Variant Call Format (VCF).

1. From any page in Ensembl, click on the link to "VEP" at the top of the page.
2. From the VEP page, click on "Launch VEP" in the "Web interface" box to load the VEP input form (see **Figure 7**).
3. By default, human is selected as in the species list, click on the "X" beside "Homo_sapiens" to remove human from the species list. Now click "Add/remove species" to load the species selector box. When it pops up, begin to type "goat" and select "Goat reference (Capra_hircus)" and click "Apply." The reference goat genome should now be selected in the VEP input form.
4. Copy the below variant data into the "Input data" box.
   #CHROM POS ID REF ALT
   4 171761. AG A
   4 237277. A G
   17 60280444. G A
   17 60280445. C A
   29 5900081. G T
   29 5900083. G C
   In this case we will run with default parameters, but the reader is encouraged to take a look at the various options available for configuring the VEP.
5. Click on the "Run" button. The display shows the status of the job. It will say "Queued," then switch to "Done" when

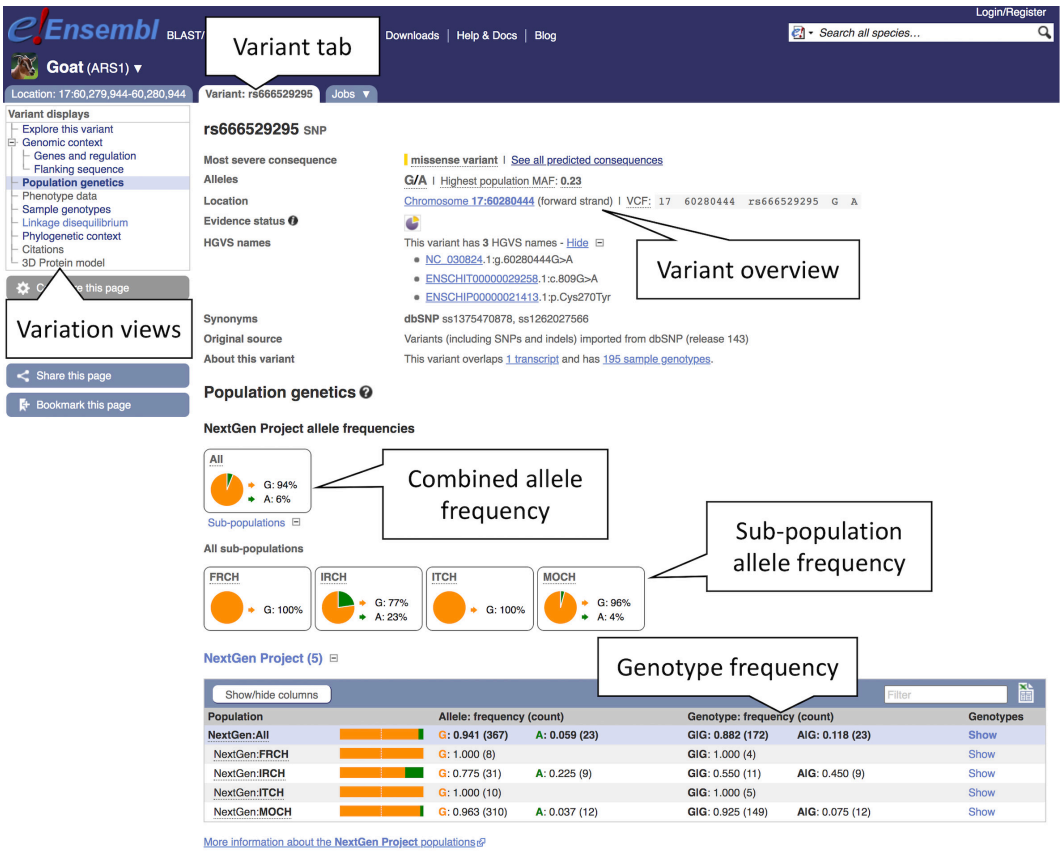


FIGURE 6 | Exploring variation data in goat. Here, we see the view of population genetics data within the variant tab. Pie charts for the overall allele frequencies along with frequencies for subpopulations are provided based off the NextGen Project data.

the job has finished. It is possible to save, edit, share or delete a job by using the icons on the right. If multiple jobs are submitted, they will appear in this table.

6. Click on "View Results" once the job is done. At the top of the results page, three sets of summary information are displayed (see **Figure 8**). The table shows that six variants have been processed, none has been filtered out, three existed already and that the variants overlap three genes and three transcripts. Pie charts show the proportions of total and coding consequences predicted: two missense variants, one intron, frameshift, and synonymous variant each, and one variant that introduces a stop codon. At the bottom of the page, a table with detailed results is displayed. It includes the alleles used for the predictions, the location of the variants, their consequences and other useful information.

Accessing and Downloading Large Data Sets

Up until this point we have focused on viewing data relating to individual loci in the context of the genome browser. To enable analysis of data over larger regions we provide a number of different methods for bulk data access. In this section, we will focus on three different methods: creating queries in our BioMart data exporter, programmatic access via our REST API and downloading files via our FTP site.

BioMart: Retrieving NCBI Gene IDs, GO Terms, and cDNA Sequences of Sheep Genes

BioMart (Kinsella et al., 2011) enables the creation of complex queries on data in Ensembl. The results can then be exported in different formats depending on the type of data queries. The underlying databases can also be accessed programmatically using R with the Bioconductor package biomaRt (Durinck et al., 2005). Data retrieval using BioMart is possible for medium to large datasets with hundreds of entries, but it is not suitable for whole genome-scale data.

The following BioMart queries first generate a CSV file with NCBI gene IDs and GO terms for the sheep genes *ESPN*, *USH1C*, *CISD2*, *THRB*, *GIPC3*, and *BRCA2* (query #1 below) and the get their cDNA sequences in FASTA format (query #2 below). In all BioMart queries a dataset must be selected, filters set (input – here the six gene names) and attributes (desired output) defined before the results can be exported.

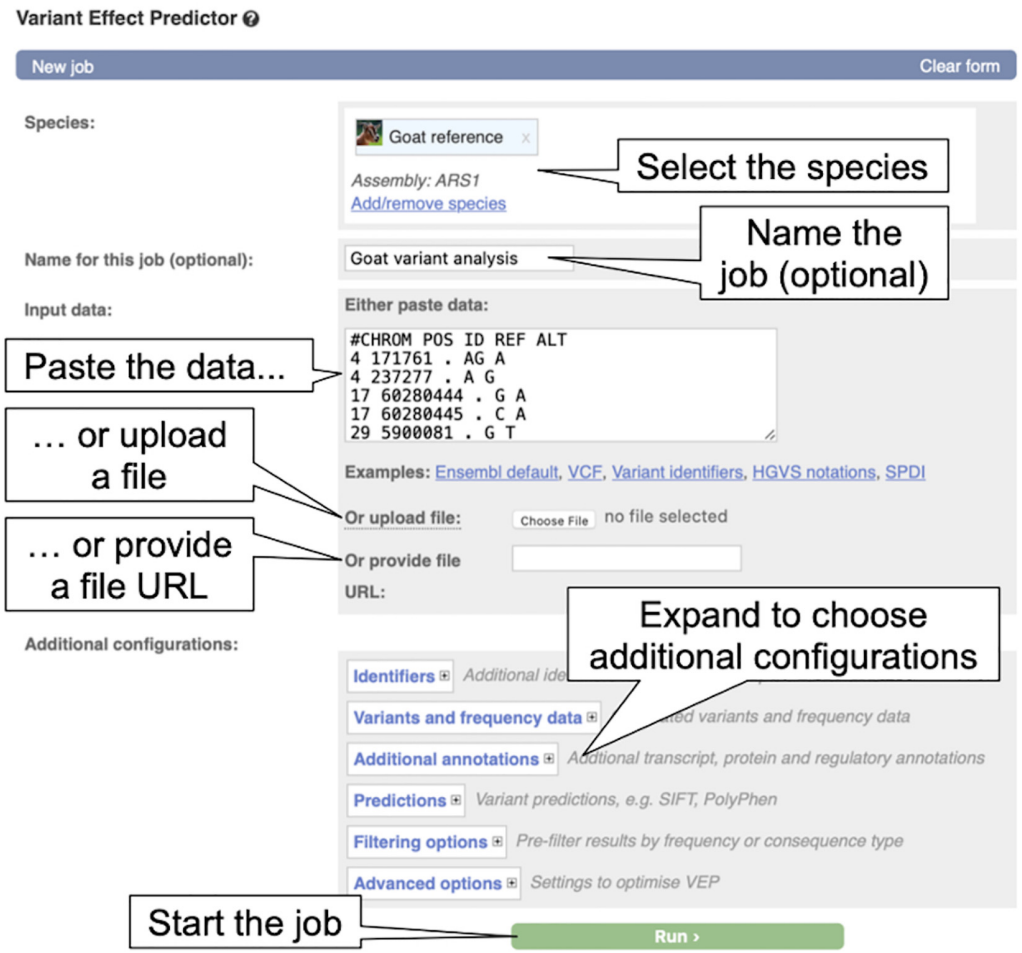


FIGURE 7 | The Variant Effect Predictor input form. The overlayed dialogue boxes provide a breakdown of the steps involved in submitting data to the VEP.

(1) Query 1: Click on "BioMart" at the top of any Ensembl page to load BioMart. You should see an interface similar to **Figure 9**.

1. Under "Dataset," Choose the "Ensembl Genes" database and the "Sheep (texel) genes (Oar_v3.1)" dataset from the respective drop-down menu.
2. Click on "Filters" in the left panel. Expand the "GENE" section by clicking on the " + " box. Select "Input external references ID list" and paste "*ESPN, USH1C, CISD2, THRB, GIPC3*, *BRCA2*" in the text box. Select "Gene Name(s)" from the drop-down menu.
3. Click on "Count" to check the Filters. This shows six genes.
4. Click on "Attributes" in the left panel. Expand the "EXTERNAL" section by clicking on the " + " box. Select "GO term accession," "GO term name," "GO term definition" and "NCBI gene (formerly Entrezgene) ID." Then expand the "GENE" section by clicking on the " + " box. The Ensembl "Gene stable ID" and "Transcript stable ID" are pre-selected. In addition, select "Gene name" to include the input in the CSV file.
5. Click "Results." Select Export all results to "File" and "CSV" from the drop-down menus. Click on the "Go" button to export the file.

(2) Query 2: Click on "Attributes" again. Do not change Dataset and Filters.

1. Select the "Sequences" attributes page at the top. Expand the "SEQUENCES" section by clicking on the "+" box. Select "cDNA sequences." Then expand the "HEADER INFORMATION" section by clicking on the "+" box. As before, Ensembl "Gene stable ID" and "Transcript stable ID" are pre-selected. Select "Gene name" to include the input in the FASTA file.
2. Click "Results." Select Export all results to "File" and "FASTA" from the drop-down menus. Click on the "Go" button to export the file.

FIGURE 8 | The Ensembl Variant Effect Predictor results page. The summary information at the top shows the consequence information for the uploaded variants. The results table at the bottom shows more detailed information for each uploaded variant including overlapping genes and transcripts.

REST API: Retrieving Homologues of a Horse Gene

The Ensembl REST API (Yates et al., 2015) is available at rest.ensembl.org and its user guide, including a Getting Started section, at http://github.com/Ensembl/ensembl-rest/wiki. Our REST API consists of a variety of endpoints. Endpoints can be considered parts of the API that allow retrieval of particular types of data. The data these endpoints provide includes transcript sequences, meta data, gene trees, variants, and a host of others.

The following Python script uses the GET homology/symbol/:species/:symbol endpoint to retrieve homologues of the horse *BRCA2* gene and print information about them in FASTA format. It uses a helper function to make the request, check for errors and decode the JSON response (the function returns text if the content_type is not JSON). The function can be integrated in any script to simplify these steps.

```
#!/usr/bin/env python
# Get the necessary Python modules
import requests, sys, json
# Define a helper function
def fetch_endpoint(server, request, content_type):
    # Make the request
    r = requests.get(server + request, headers = {"Accept": content_type})
    # Get the status of any failed query
    if not r.ok:
        r.raise_for_status()
        sys.exit()
    # Decode JSON, if used as content_type. If not, return text.
    if content_type = = "application/json":
        return r.json()
    else:
        return r.text
# Define the gene name
gene = "BRCA2"
# Define the general URL parameters
server = "http://rest.ensembl.org/"
ext_hom = "homology/symbol/horse/" + gene
con = "application/json"
# Submit the query by calling the helper function
get_hom = fetch_endpoint(server, ext_hom, con)
# Print some information about the homologues
    for data in get_hom["data"]:
        for homology in data["homologies"]:
            source_id = homology["source"]["id"]
            source_species = homology["source"]["species"]
            source_seq = homology["source"]["align_seq"]
            target_id = homology["target"]["id"]
            target_seq = homology["target"]["align_seq"]
            target_species = homology["target"]["species"]
            print (">," source_id + " " + source_species + "\n" + source_seq + "\n>," target_id + " " + target_species +"\n" + target_seq)
```

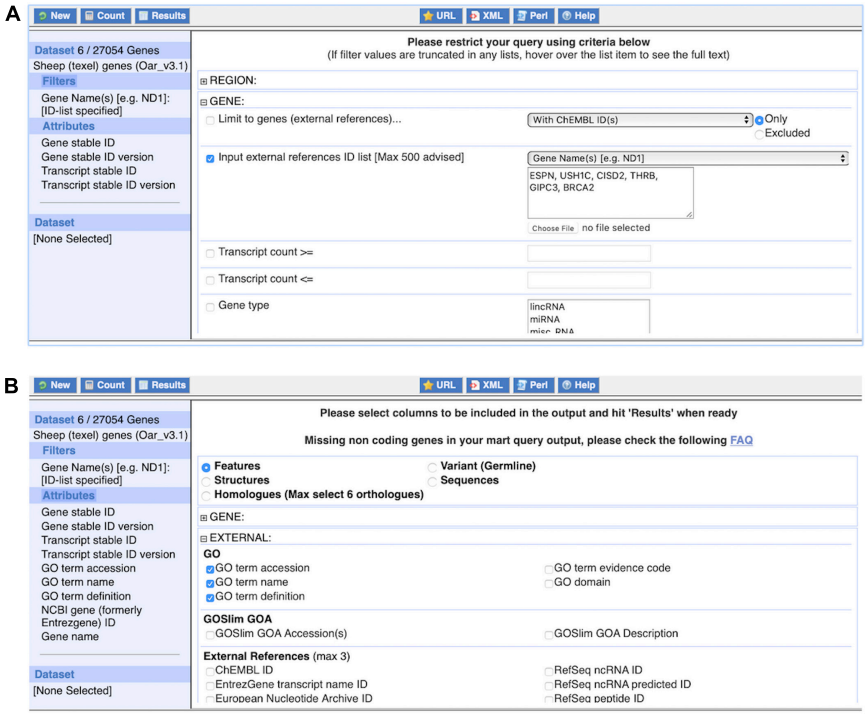


FIGURE 9 | The BioMart web interface. Panel **(A)** shows the application of one commonly used filter, an external references identifier list, here consisting of six gene names. The dataset size of six genes is displayed after clicking on "Count" at the top left. Panel **(B)** shows the selection of attributes, here "GO term accession," "GO term name," and "GO term definition." The complete list of attributes, i.e., what will be included in the results table, is shown on the left.

FTP: Downloading a GTF File for Atlantic Salmon

The Ensembl FTP site at ftp://ftp.ensembl.org/pub/can be accessed using the command line, a script, rsync, a web browser or FTP client. It provides files for all species in several formats, such as FASTA, GTF/GFF3 and VCF for the current and previous releases (going back to release 19). An overview of available data can be found at https://www.ensembl.org/info/data/ftp/index.html.

To download a GTF file with all annotated transcripts for Atlantic salmon using a web browser:

1. Navigate to ftp://ftp.ensembl.org/pub/
2. Click on current_gtf and then on salmo_salar
3. Click on the file Salmo_salar.ICSASG_v2.100.gtf.gz

RESULTS

Here, we present a summary of some of the key livestock data present in Ensembl, including the genomes, different annotation types and FTP files.

Livestock Species in Ensembl

Ensembl contains a large variety of livestock species ranging from cow, pig, and chicken to Arabian camel, African ostrich, and Siberian musk deer. In addition, there are some species with accompanying non-domesticated genomes, such as common and wild mallard or domestic and wild yak. **Table 1** shows ten of the major livestock species in Ensembl along with some accompanying information about the genome assembly and the Ensembl release which included the most recent update to the associated annotation.

Many of the livestock species, particularly the more recently sequenced ones, have high quality genome assemblies based on long read sequencing. Several species including chicken, cow, goat, and pig have genome assemblies with a contig N50 of over 10 Mb. That being said, there is considerable variability in the quality of livestock assemblies in Ensembl as many species were assembled using short read data. We see a range of contig N50 values, from approximately 30 kb to over 48 Mb, representing over a 1,000-fold variation in the level of contiguity of these assemblies.

Gene and Transcript Annotation Across Livestock Species

All livestock species have gene sets generated via the Ensembl gene annotation system (Aken et al., 2016). The counts of the coding and non-coding genes and transcripts across ten livestock species are shown in **Figure 10**. Across clades a consistent pattern emerges in terms of the expected number of protein-coding genes. Birds, with their smaller genomes and low number of repeat regions, have approximately 16,000 protein-coding genes, while mammals have closer to 20,000. Fish gene counts are highly variable, ranging between 20,000 and 60,000 protein-coding genes, which is reflective of the multiple rounds of whole genome duplication across fish species.

TABLE 1 | Assembly statistics for ten reference livestock species in Ensembl.

Species	Assembly	Accession	Contig N50	Date	Release
Chicken	GRCg6a	GCA_000002315.5	17655422	2018-03-27	95
Duck	CAU_duck1.0	GCA_002743455.1	88037	2017-11-03	96
Cow	ARS-UCD1.2	GCA_002263795.2	25896116	2018-04-11	98
Goat	ARS1	GCA_001704415.1	26244591	2016-08-24	92
Horse	EquCab3.0	GCA_002863925.1	1502753	2018-01-05	98
Pig	Sscrofa11.1	GCA_000003025.6	48231277	2017-02-07	98
Sheep	Oar_rambouillet_v1.0	GCA_002742125.1	2572683	2017-11-02	101
Herring	Ch_v2.0.2	GCA_900700415.1	1151065	2019-04-16	98
Seabream	fSpaAur1.1	GCA_900880675.1	2862625	2019-07-30	99
Salmon	ICSASG_v2	GCA_000233375.4	36085	2015-06-10	99

Date refers to the date the assembly was submitted to the public archives. Release is the Ensembl release number in which the annotation for the corresponding assembly was most recently updated.

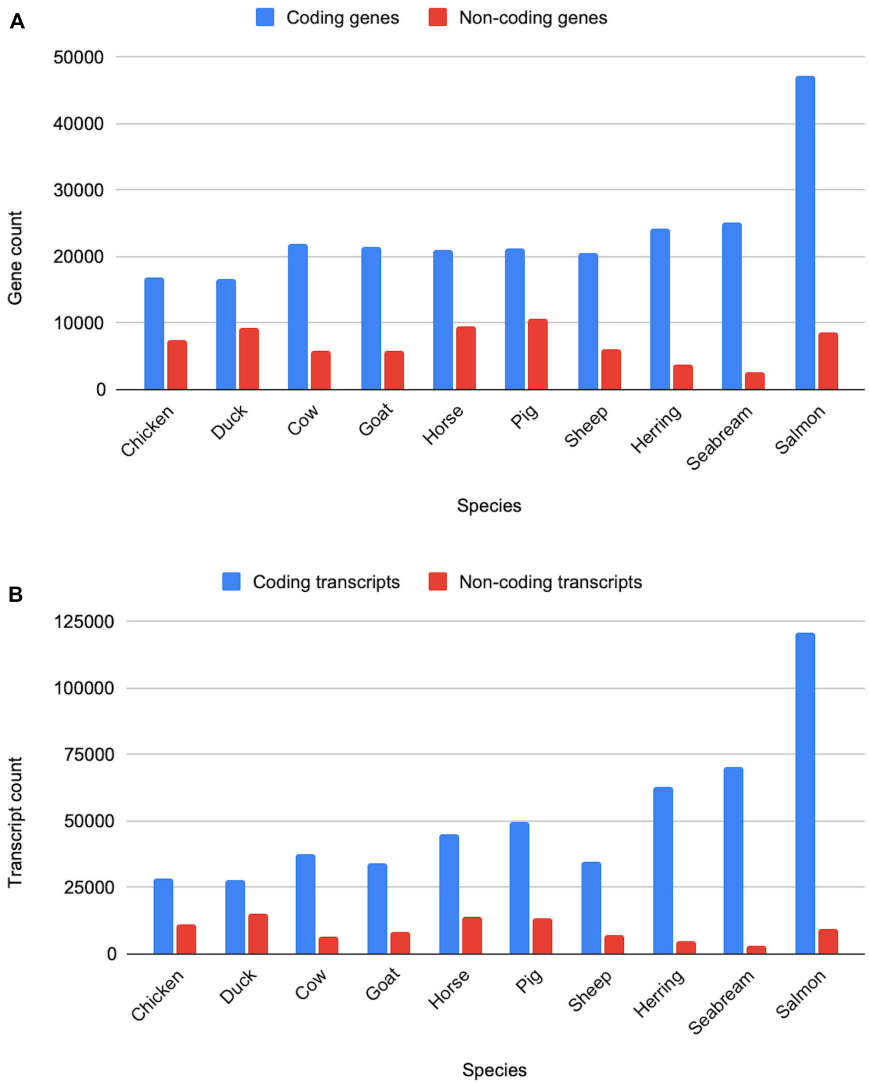


FIGURE 10 | Counts of coding and non-coding genes and transcripts for ten livestock species. Panel **(A)** shows the gene counts, while panel **(B)** shows the transcripts counts for various livestock species from the bird, mammal and fish clades.

Non-coding genes, which include pseudogenes, long and short non-coding RNAs, are considerably more variable in terms of overall counts. In particular, lncRNAs annotations are highly variable as they are detected through transcriptomic data and are often expressed at low levels and tissue-specific (Uszczynska-Ratajczak et al., 2018). As a result, species with large amounts of transcriptomic data, sampled across a broad set of tissues and development stages, have considerably more lncRNA annotations than species with a small amount of transcriptomic data. Sequence similarity to species with entries in miRBase (Kozomara et al., 2019) and Rfam (Kalvari et al., 2018) heavily affects annotation of small ncRNAs, species with high sequence similarity to species with good coverage in these databases will generally have higher counts of sncRNAs.

Comparative Livestock Data

All livestock species are included as part of our protein and ncRNA gene tree pipelines, which calculate gene tress across all species in Ensembl.

Table 2 provides more information on the other comparative data available across different sets of species. These data are calculated for subsets of the species in Ensembl; the selection is made based on the quality of the genome assembly and the clade the species belongs to. For example, the high-quality 57 mammals EPO alignments are clade-specific and require that the genome assemblies used in the alignment are at the chromosome-level, while the 111 eutherian mammals EPO-extended alignment is a broader sampling of mammals and extends to lower quality genome assemblies.

For pig, there are 13 breeds available in Ensembl and we have generated additional breed-specific comparative resources. **Table 3** shows data for a 16 species EPO-extended alignment between the 13 pig breeds, in addition to three outgroup species. We have also generated a set of gene trees using these 16 species.

Variation Livestock and Companion Animal Resources

Ensembl has variation data for ten livestock and companion animal species, summarised in **Table 4**. Between 9,000 and 104 million short variants are available for each of these species; in addition, structural variants are available for cow, dog, horse, pig, and sheep. Sources of short variants for livestock and companion animals in Ensembl are dbSNP, Pig SNP Consortium and EVA study PRJEB34225, while all structural variants are imported from DGVa. Sources of the phenotype data are OMIA, GOA and AnimalQTLdb, while allele frequencies are from NextGen Project, International Sheep Genome Consortium and EVA studies PRJEB34225, PRJEB24066, and PRJEB9799. Similar to the variant data, the amount of linked data available for the different species varies significantly. The richest linked data sets with all data types are available for Cow, Dog, Horse, and Sheep, while the linked data sets for Atlantic salmon, Cat and Turkey are most limited.

For each variant, we also identify all overlapping Ensembl transcripts and provide the most severe consequence of the

TABLE 2 | Availability of whole genome alignments and syntenies for livestock and companion animal species.

Species	16 pig breeds (EPO-Extended)	57 mammals (EPO)	111 eutherian mammals (EPO-Extended)	95 amniota vertebrates (Mercator-Pecan)	36 sauropsids (EPO)	69 sauropsids (EPO-Extended)	50 fish (EPO)	86 fish (EPO-Extended)	Pairwise alignments	Synteny maps
Chicken				×	×	×			75	24
Duck				×	×	×			2	2
Turkey				×	×	×			2	2
Cat		×	×	×					7	5
Cow	×	×	×	×					26	18
Dog		×	×	×					22	12
Goat		×	×	×					3	3
Horse	×	×	×	×					5	4
Pig	×	×	×	×					19	7
Sheep (texel)	×	×	×	×					6	6
Atlantic herring							×	×	3	2
Atlantic salmon							×	×	3	1
Seabream							×	×	3	2

X indicates inclusion of a given species in the listed multiple alignments. Methods used to compute multiple alignments are given in brackets. "Pairwise alignments" and "Synteny maps" columns specify the number of available alignments of each type.

variant, as defined by sequence ontology[2]. For missense variants of cat, chicken, cow, dog, goat, horse, pig, and sheep, SIFT scores are provided to help assess the potential pathogenicity of the variants.

Files Available for Livestock Species

Running analyses locally is made easier by the availability of data that has been exported from the Ensembl database into popular file formats for download. Files on the FTP site are organised by releases, with older releases stretching back almost 20 years. As such it is possible to retrieve files representing different reference assemblies or annotations through time. **Table 5** shows some of the most popular files available on the FTP site along with a brief description of their content.

[2]http://www.sequenceontology.org

DISCUSSION

Ensembl contains a wealth of livestock data that can be accessed in a number of different ways as described in the protocols presented here. The depth of available data reflects the commercial value of applying genomics to understanding these species and breeds and there is, of course, some variability in the data available for different species. We have certain key data types available for all livestock species including gene annotations, gene trees, inclusion in whole genome alignments

TABLE 3 | Summary statistics of the pig breeds whole genome alignment.

Species – breed	Genome len. (gb)	Genome cov. (gb)	Genome cov. (%)	Coding exon length (bp)	Coding exon coverage (bp)	Coding exon coverage (%)
Pig – Reference	2.50	2.35	93.82	35,828,571	34,187,392	95.42
Pig – USMARC	2.76	2.30	83.49	35,166,034	32,878,028	93.49
Pig – Wuzhishan	2.51	2.29	91.43	32,216,998	30,303,931	94.06
Pig – Tibetan	2.44	2.24	91.88	33,242,148	31,601,348	95.06
Pig – Meishan	2.47	2.29	92.96	34,405,547	33,144,362	96.33
Pig – Jinhua	2.45	2.30	93.54	34,359,345	33,112,715	96.37
Pig – Rongchang	2.46	2.30	93.47	34,688,475	33,474,298	96.5
Pig – Bamei	2.46	2.29	93.26	34,436,299	33,139,773	96.24
Pig – Largewhite	2.46	2.31	93.84	34,666,066	33,395,166	96.33
Pig – Pietrain	2.44	2.30	94.24	34,368,936	33,123,601	96.38
Pig – Berkshire	2.43	2.29	94.17	34,511,197	33,225,283	96.27
Pig – Hampshire	2.44	2.30	94.41	34,573,096	33,284,528	96.27
Pig – Landrace	2.44	2.30	94.1	34,646,469	33,337,432	96.22
Cow	2.72	2.52	92.83	34,983,666	33,272,562	95.11
Horse	2.51	2.28	90.93	37,559,221	34,821,694	92.71
Sheep (texel)	2.62	2.49	95	32,776,750	30,904,741	94.29

This alignment was initially generated in Ensembl release 98 and is composed of 11,721 blocks (up to 953,990 bp long). Cow, horse, and sheep were used as outgroups in the alignment.

TABLE 4 | Variation data counts for livestock and companion animal species.

Species	Short variants	Structural variants	Populations with allele frequency data/Total number of individuals	Variants with			Phenotype data	
				Population genotypes	Sample genotypes	SIFT scores	Phenotypes	Gene associations
Atlantic salmon	10.1 M	ND	3/80	10.1 M	10.1 M	ND	ND	ND
Cat	3.6 M	ND	ND	ND	ND	7 K	63	64
Chicken	24 M	ND	ND	3.2 M	3.2 M	229 K	225	5 K
Cow	104 M	18 K	1/8	10 K	10 K	2.1 M	549	98 K
Dog	5.9 M	104 K	1/219	727 K	727 K	50 K	257	258
Goat	37 M	ND	5/195	ND	ND	92 K	11	11
Horse	21 M	193 K	1/6	1.1	1.1 M	98 K	88	852
Pig	67 M	224 K	ND	175	15	218 K	394	20 K
Sheep	61 M	2	68/633	147	64	222 K	172	2 K
Turkey	9 K	ND	ND	48	ND	ND	29	42

The counts are based on Ensembl release 101. Sheep data are from the texel breed. Data sources: Short variants – EVA study PRJEB34225 (Atlantic salmon), dbSNP and Pig SNP Consortium (Pig), dbSNP (others); Structural variants – DGVa (all); Allele frequencies – EVA study PRJEB34225 (Atlantic salmon), NextGen Project (Cow, Goat), EVA study PRJEB24066 (Dog), EVA study PRJEB9799 (Horse), NextGen Project and International Sheep Genome Consortium (Sheep); Phenotypes – OMIA (Cat, Dog, and Goat), GOA and OMIA (Turkey), AnimalQTLdb and OMIA (others). ND, no data.

TABLE 5 | Selected files available on the Ensembl FTP site for livestock species.

Data	File types	Description
Genome sequences	FASTA, EMBL, GenBank	The non-redundant genome sequences (sometimes referred to as "toplevel"). Available in softmasked, hardmasked, and unmasked variations
Transcript sequences	FASTA	Sequences for all transcript structures including any 5′ and 3′ UTR present. A cDNA file contains all coding transcript sequences, while non-coding transcripts are represented in a separate ncRNA file
Peptide sequences	FASTA	Amino acid translations for each transcript with an annotated open reading frame
Gene annotation	GFF3, GTF, EMBL, GenBank	Files containing information on genes, transcripts exons and cds structures. The exact content varies from between formats, with GFF3 and GTF being less verbose than EMBL and GenBank formats
Genome alignments	lastZ Net, MAF	Pairwise lastZ genome alignments are available for all species against the reference for their clade. EPO multi-genome alignments are available for various clades including mammals and pig breeds
Variation data	VCF, VEP cache	Variation data is available in VCF format for several species including chicken, cow, goat, pig, and sheep. VEP cache files allow local installation of the VEP to speed up the analysis by using the data stored in the cache

Files can be found on ftp://ftp.ensembl.org/pub/, with files relating to the current release found in subdirectories prefixed with "current_." Files from all previous releases back to release 19 are also available.

and compatibility with tools such as our BLAST service and the Ensembl VEP. In addition to this, there are a significant number of additional data tracks, though this varies from species to species. For example, while all livestock species have associated transcriptomic data used to produce their gene annotations, the amount of transcriptomic data varies considerably. For pig, there is an extensive collection of both short and long read tissue data, all of which are available in the browser as tissue-specific gene tracks. For several species, we have variation data, with tracks for SNPs and structural variants as well as an LD calculation tool available for goat, salmon and sheep. With falling costs of generating these data, we expect that data availability will become more consistent across species. Ensembl will continue to integrate new data in line with the needs of the livestock community.

Working with Ensembl can be divided into two major approaches: working through the genome browser or fetching data and running local analyses. These two approaches are often done in tandem, with analysis starting in the browser, before continuing on to large-scale local analyses using the REST service and/or downloading files from the FTP site.

The browser provides access to many different views of the underlying data. The gene and location views support access to small scale information on a gene or set of genes, while more specialised views such as the synteny view allow for cross species comparisons across genomic regions. A knowledge of the available views is key to utilising the browser fully. The default option for each view tries to strike the right balance between delivering key information without overloading the view with too much visual noise. Additional tracks, that are not visible by default, are accessible from the configuration menu. An example of these are the transcriptomic data tracks, which are available in the gene and location views. For species with transcriptomic data, it is possible to switch on tracks showing the transcripts identified in each tissue/development stage along with intron support and BAM/BigWig coverage plots. These additional tracks can provide valuable extra information for each locus. In addition to understanding available views and tracks, efficiently working with the browser benefits from knowledge of the supported tools. The BLAST/BLAT service can be used to help identify unannotated genes or exons in a genomic region of interest, while the Ensembl VEP can be used to analyse uploaded variants against reference genomes, annotations and linked data. BioMart is a powerful tool for exporting complex data sets. Taking the time to become familiar with the available tools, their strengths and limitations is an important aspect of fully utilising the genome browser.

For large-scale or custom analyses, the path to interacting with Ensembl generally shifts from the browser to programmatic access via the API and bulk download of data via the FTP site. A typical workflow could involve downloading the softmasked genome sequences from the FTP site for a set of species to be analysed, followed by fetching annotation data, such as the gene sets, either from the files on the FTP site or via the REST API. The REST API is a powerful method to subsample or filter the data in various ways, such as only selecting genes on a certain chromosome or fetching a particular subclass of genes such as miRNAs. If no filtering or grouping of the data are required, the corresponding FTP files are generally the most straightforward and fastest way to get bulk access to annotation data. When combined, the FTP site and REST API give access to the vast wealth of data present in Ensembl. They act as a starting point for local workflows and a better understanding of what can be fetched directly from Ensembl can help accelerate downstream analysis.

While the protocols provided here give a comprehensive overview of what data are in Ensembl and how to interact with it, there is still much more to discover. Support is available in a variety of ways including a dedicated email helpdesk (helpdesk@ensembl.org) to field any inquiries about Ensembl. We are currently running virtual training courses during the COVID-19 pandemic, and will resume a full in-person and virtual training programme, including webinars, when possible. Our training materials are accessible online at https://training.ensembl.org. Ensembl courses are also available from the EMBL-EBI Train Online platform[3]. Tutorial videos and recorded webinars can be found both on our YouTube[4] and Youku[5] channels.

[3]https://www.ebi.ac.uk/training/online/

[4]https://www.youtube.com/user/EnsemblHelpdesk

[5]https://i.youku.com/i/UMzM1NjkzMTI0?spm=a2hzp.8244740.0.0

We invite the community to contact us via our helpdesk to ask questions regarding the use of our browser, tools and related resources, to request training events or to suggest features which would assist their work.

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

AG, MS, and FJM designed the protocols. AG, MS, FJM, and PF wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded in whole, or in part, by the Wellcome Trust (Grant number 108749/Z/15/Z). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This infrastructure has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. (817923) (AquaFAANG), No. (817998) (GENE-SWitCH), and No. (815668) (BovReg). We acknowledge funding from the Biotechnology and Biological Sciences Research Council (BB/M011615/1 and BB/S020152/1) and the European Molecular Biology Laboratory.

REFERENCES

Abu Hatab, A., Cavinato, M. E. R., and Lagerkvist, C. J. (2019). Urbanization, livestock systems and food security in developing countries: a systematic review of the literature. *Food Secur.* 11, 279–299. doi: 10.1007/s12571-019-00906-1

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., et al. (2016). The Ensembl gene annotation system. *Database* 2016:baw093. doi: 10.1093/database/baw093

Alberto, F. J., Boyer, F., Orozco-terWengel, P., Streeter, I., Servin, B., de Villemereuil, P., et al. (2018). Convergent genomic signatures of domestication in sheep and goats. *Nat. Commun.* 9:813. doi: 10.1038/s41467-018-03206-y

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., et al. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* 49, 643–650. doi: 10.1038/ng.3802

Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T., and Su, G. (2012). Single-step methods for genomic evaluation in pigs. *Animal* 6, 1565–1571. doi: 10.1017/s1751731112000742

Cleveland, M. A., and Hickey, J. M. (2013). Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. *J. Anim. Sci.* 91, 3583–3592. doi: 10.2527/jas.2013-6270

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440. doi: 10.1093/bioinformatics/bti525

Fang, X. D., Mu, Y. L., Huang, Z. Y., Li, Y., Han, L. J., Zhang, Y. F., et al. (2012). The sequence and analysis of a Chinese pig genome. *Gigascience* 1, 1–16. doi: 10.1186/2047-217x-1-16

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/jds.2008-1646

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., et al. (2016). Ensembl comparative genomics resources. *Database* 2016:bav096. doi: 10.1093/database/bav096

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., et al. (2021). Ensembl 2021. *Nucleic Acids Res.* 49, D884–D891. doi: 10.1093/nar/gkaa942

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., et al. (2018). Ensembl variation resources. *Database* 2018:bay119. doi: 10.1093/database/bay119

Hurst, P., Termine, P., and Karl, M. (2005). *Agricultural Workers and Their Contribution to Sustainable Agriculture and Rural Development*. Rome: FAO.

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J. F., Faraut, T., et al. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* 344, 1168–1173. doi: 10.1126/science.1252806

Kabubo-Mariara, J. (2009). Global warming and livestock husbandry in Kenya: impacts and adaptations. *Ecol. Econ.* 68, 1915–1924. doi: 10.1016/j.ecolecon.2009.03.002

Kalbfleisch, T. S., Rice, E. S., DePriest, M. S., Walenz, B. P., Hestand, M. S., Vermeesch, J. R., et al. (2019). Improved reference genome for the domestic horse increases assembly contiguity and composition. *Commun. Biol.* 2:342. doi: 10.1038/s42003-019-0591-3

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., et al. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* 46, D335–D342. doi: 10.1093/nar/gkx1038

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011:bar030. doi: 10.1093/database/bar030

Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., et al. (2018). De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* 36, 1174–1182. doi: 10.1038/nbt.4277

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res.* 47, D155–D162. doi: 10.1093/nar/gky1141

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1081. doi: 10.1038/nprot.2009.86

Lenffer, J., Nicholas, F. W., Castle, K., Rao, A., Gregory, S., Poidinger, M., et al. (2006). OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res.* 34(Suppl._1), D599–D601. doi: 10.1093/nar/gkj152

Li, M. Z., Chen, L., Tian, S. L., Lin, Y., Tang, Q. Z., Zhou, X. M., et al. (2017). Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Res.* 27, 865–874. doi: 10.1101/gr.207456.116

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533, 200–205. doi: 10.1038/nature17164

Low, W. Y., Tearle, R., Liu, R. J., Koren, S., Rhie, A., Bickhart, D. M., et al. (2020). Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nat. Commun.* 11:2071. doi: 10.1038/s41467-020-15848-y

Otte, J., and Upton, M. (2005). "Poverty and livestock agriculture," in *Animal Production and Animal Science Worldwide*, eds A. Rosati, A. Tewolde, and C. Mosconi (Wageningen: Wageningen Academic Publisher), 281–296.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17:122. doi: 10.1186/s13059-016-0974-4

Paten, B., Herrero, J., Beal, K., Fitzgerald, S., and Birney, E. (2008). Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* 18, 1814–1828. doi: 10.1101/gr.076554.108

Pettersson, M. E., Rochus, C. M., Han, F., Chen, J. F., Hill, J., Wallerman, O., et al. (2019). A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. *Genome Res.* 29, 1919–1928. doi: 10.1101/gr.253435.119

Popp, A., Lotze-Campen, H., and Bodirsky, B. (2010). Food consumption, diet shifts and associated non-CO_2 greenhouse gases from agricultural production. *Glob. Environ. Chang.* 20, 451–462. doi: 10.1016/j.gloenvcha.2010.02.001

Rodriguez, J. M., Rodriguez-Rivas, J., Di Domenico, T., Vazquez, J., Valencia, A., and Tress, M. L. (2018). APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res.* 46, D213–D217. doi: 10.1093/nar/gkx997

Rota, A., and Sperandini, S. (2009). *Livestock and Pastoralists. Livestock Thematic Papers–Tools for Project Design*. Rome: International Fund for Agricultural Development (IFAD).

Siddiki, A. Z., Baten, A., Billah, M., Alam, M. A. U., Shawrob, K. S. M., Saha, S., et al. (2019). The genome of the Black Bengal goat (*Capra hircus*). *BMC Res. Notes* 12:362. doi: 10.1186/s13104-019-4400-3

Sohn, J. I., Nam, K., Hong, H., Kim, J. M., Lim, D., Lee, K. T., et al. (2018). Whole genome and transcriptome maps of the entirely black native Korean chicken breed Yeonsan Ogye. *Gigascience* 7, 1–14. doi: 10.1093/gigascience/giy086

Thornton, P. K. (2010). Livestock production: recent trends, future prospects. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 2853–2867. doi: 10.1098/rstb.2010.0134

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., and Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* 19, 535–548. doi: 10.1038/s41576-018-0017-y

Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., et al. (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience* 9:giaa051. doi: 10.1093/gigascience/giaa051

Warren, W. C., Hillier, L. W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., et al. (2017). A new chicken genome assembly provides insight into avian genome structure. *G3 Genes Genom. Genet.* 7, 109–117. doi: 10.1534/g3.116.035923

Xu, P., Xu, J., Liu, G. J., Chen, L., Zhou, Z. X., Peng, W. Z., et al. (2019). The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*. *Nat. Commun.* 10:4625. doi: 10.1038/s41467-019-12644-1

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R., et al. (2015). The Ensembl REST API: Ensembl data for any language. *Bioinformatics* 31, 143–145. doi: 10.1093/bioinformatics/btu613

Yawson, D. O., Mulholland, B. J., Ball, T., Adu, M. O., Mohan, S., and White, P. J. (2017). Effect of climate and agricultural land use changes on uk feed barley production and food security to the 2050s. *Land* 6:74. doi: 10.3390/land6040074

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., and Flicek, P. R. (2015). The ensembl regulatory build. *Genome Biol.* 16:56. doi: 10.1186/s13059-015-0621-5

Conflict of Interest: PF is a member of the Scientific Advisory Boards of Fabric Genomics, Inc. and Eagle Genomics, Ltd.

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.

Copyright © 2021 Martin, Gall, Szpak and Flicek. 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: 30 April 2021
doi: 10.3389/fgene.2021.610116

Network Analyses Predict Small RNAs That Might Modulate Gene Expression in the Testis and Epididymis of *Bos indicus* Bulls

*Andressa O. de Lima[1], Juliana Afonso[2], Janette Edson[3], Esteban Marcellin[4], Robin Palfreyman[4], Laercio R. Porto-Neto[5], Antonio Reverter[5] and Marina R. S. Fortes[3]**

[1] Department of Production and Animal Health, School of Veterinary Medicine, São Paulo State University (UNESP), Araçatuba, Brazil, [2] Department of Animal Science, University of São Paulo/ESALQ, Piracicaba, Brazil, [3] School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD, Australia, [4] Australian Institute for Bioengineering and Nanotechnology (AIBN), The University of Queensland, St Lucia, QLD, Australia, [5] CSIRO Agriculture and Food, Queensland Bioscience Precinct, St. Lucia, QLD, Australia

OPEN ACCESS

Edited by:
Christopher K. Tuggle, Iowa State University, United States

Reviewed by:
Terje Raudsepp, Texas A&M University, United States
Wansheng Liu, Pennsylvania State University (PSU), United States
Gregorio Miguel Ferreira De Camargo, Federal University of Bahia, Brazil

***Correspondence:**
Marina R. S. Fortes
m.fortes@uq.edu.au

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

Received: *25 September 2020*
Accepted: *19 March 2021*
Published: *30 April 2021*

Citation:
de Lima AO, Afonso J, Edson J, Marcellin E, Palfreyman R, Porto-Neto LR, Reverter A and Fortes MRS (2021) Network Analyses Predict Small RNAs That Might Modulate Gene Expression in the Testis and Epididymis of Bos indicus Bulls. Front. Genet. 12:610116. doi: 10.3389/fgene.2021.610116

Spermatogenesis relies on complex molecular mechanisms, essential for the genesis and differentiation of the male gamete. Germ cell differentiation starts at the testicular parenchyma and finishes in the epididymis, which has three main regions: head, body, and tail. RNA-sequencing data of the testicular parenchyma (TP), head epididymis (HE), and tail epididymis (TE) from four bulls (three biopsies per bull: 12 samples) were subjected to differential expression analyses, functional enrichment analyses, and co-expression analyses. The aim was to investigate the co-expression and infer possible regulatory roles for transcripts involved in the spermatogenesis of *Bos indicus* bulls. Across the three pairwise comparisons, 3,826 differentially expressed (DE) transcripts were identified, of which 384 are small RNAs. Functional enrichment analysis pointed to gene ontology (GO) terms related to ion channel activity, detoxification of copper, neuroactive receptors, and spermatogenesis. Using the regulatory impact factor (RIF) algorithm, we detected 70 DE small RNAs likely to regulate the DE transcripts considering all pairwise comparisons among tissues. The pattern of small RNA co-expression suggested that these elements are involved in spermatogenesis regulation. The 3,826 DE transcripts (mRNAs and small RNAs) were further subjected to co-expression analyses using the partial correlation and information theory (PCIT) algorithm for network prediction. Significant correlations underpinned the co-expression network, which had 2,216 transcripts connected by 158,807 predicted interactions. The larger network cluster was enriched for male gamete generation and had 15 miRNAs with significant RIF. The miRNA bta-mir-2886 showed the highest number of connections (601) and was predicted to down-regulate *ELOVL3*, *FEZF2*, and *HOXA13* (negative co-expression correlations and confirmed with TargetScan). In short, we suggest that bta-mir-2886 and other small RNAs might modulate gene expression in the testis and epididymis, in *Bos indicus* cattle.

Keywords: bovine, RNA-sequencing, systems biology, spermatozoa, miRNA, bta-mir-2886, small RNAs, spermatogenesis

INTRODUCTION

Spermatozoid is the most specialized cell in mammalian organisms. Spermatogenesis, the differentiation of male germ cells, relies on a complex network of specialized molecular mechanisms that are critical to male fertility (MacLean and Wilkinson, 2005; Marengo, 2008; Hermann et al., 2018). During spermatogenesis, three sequential phases of cell proliferation and differentiation occur, where there is an extensive multiplication and proliferation of spermatogonial stem cells, followed by a meiotic division, and finally a remodeling of the nuclear and cellular components forming sperm cells (Abou-Haila and Tulsiani, 2000). Spermatogenesis starts with the multiplication of spermatogonial stem cells followed by their meiotic division into spermatids, which then differentiate into spermatozoa that are released into the lumen of seminiferous tubules in the testis (Staub and Johnson, 2018). Spermatozoa leaving the testis transit through the epididymis, where they further mature, acquiring motility and the ability to fertilize the egg (Cornwall, 2009). The epididymis is composed of *caput* (head), *corpus* (body), and *cauda* (tail), consisting of region-specific characteristics, including a region-specific luminal protein profile (Cornwall, 2009). The spermatozoa from the testis pass to the epididymis, which contributes to their maturation (Belleannée et al., 2012). In the epididymis, secreted luminal proteins, water, and solute balance contribute to the luminal environment necessary for sperm maturation (Huang et al., 2006). Fully formed mature sperm cells emerge from the tail epididymis and are stored until the ejaculation event in the *vas deferens*.

Recently, Hombach and Kretz (2016) have proposed a role for small RNAs in the testis and epididymis: they may be key regulators of gene expression in spermatogenesis, as they are in most cellular processes. RNA polymerase II transcribes small RNAs, and because of this, their expression is mostly regulated by mechanisms that regulate RNA polymerase II activity, such as the interaction of transcription factors and specific DNA sequences (Fuda et al., 2009). Some classes of small RNAs, such as micro (miRNAs), small nuclear (snRNAs), and small nucleolar (snoRNAs) RNAs, play a role in spermatogenesis by being involved in meiosis (Pradillo and Santos, 2018). Small RNAs regulate sperm maturation through mRNA-silencing mechanisms (Nixon et al., 2019), such as destabilizing mRNAs via deadenylation complexes (Bartel, 2018). In addition, miRNAs are important to maintain the epididymis homeostasis and function (Nixon et al., 2019). Small RNAs are present in epididymosomes (Sullivan, 2016) and can modulate mRNA expression in spermatozoa during the epididymal transit (Belleannée, 2015).

Considering the different roles played by the testis and epididymis, some studies investigate the pattern of gene expression of male tract reproductive tissues to shed light on the biological processes related to each specific tissue. Among these studies, there was a characterization of epididymis gene expression in humans (Thimon et al., 2007; Browne et al., 2015) and yak (Zhao et al., 2019). Guyonnet et al. (2009) have reported the differences in the expression pattern between the testis and epididymis in boar. However, knowledge of gene expression patterns in the testis or epididymis of *Bos indicus* bulls is lacking, and the hypothesized role of small RNAs in these tissues remains to be confirmed.

By sampling biopsies from testicular parenchyma (TP), head epididymis (HE), and tail epididymis (TE), we obtained different cell groups that are representative of spermatogenesis in three different stages. In the TP, Sertoli, Leydig, and differentiating male germ cells represent a group of cells with the DNA still bound to histones. In TE and HE, sperm cells are further along their differentiation process, and protamines instead of histones are observed, which is typical of mature sperm cells as described before (Fortes et al., 2014). Therefore, when sampling these tissues, we opened a window to investigate spermatogenesis. Our aim was to combine RNA sequencing, differential gene expression, functional enrichment, and co-expression analyses to investigate potential transcript interactions in the male reproductive system, using *Bos indicus* bulls as a model organism.

Studies on the testicular transcriptome, such as this one, are not only useful for understanding male fertility but also very helpful for genome annotation. Testicular tissue may be under less evolutionary pressure and this can be promoting duplication of protein-coding events and an overabundance of non-coding RNAs (ncRNAs), and not all the protein-coding genes expressed are functional (Soumillon et al., 2013). It is generally reported that the testes have higher gene expression than other tissues (Soumillon et al., 2013; Uhlén et al., 2015). The data reported on this study is available through the Functional Annotation of Animal Genomes (FAANG) Consortium for further research[1].

MATERIALS AND METHODS

Samples and Data

All the experimental procedures were conducted and approved by the ethics committee of the University of Queensland, Brisbane, Australia (protocol number: ANRFA/SCMB/094/16). Tissue samples were collected after euthanasia of cattle for commercial purposes, as part of normal beef industry activities. Testicular samples ($n = 4$) from mature Brahman bulls (approx. 2 years old) were collected shortly after slaughter and delivered to the research team, who performed the biopsies. For each bull, we performed three biopsies: testicular parenchyma (TP), head epididymis (HE), and tail epididymis (TE). Each biopsy (approximately 50 mg of tissue) was collected in Eppendorf tubes with 1 ml of RNA*later*® (RNA stabilizing reagent, Ambion Inc., Austin, TX, United States). The biopsies were left to stabilize in a cold room overnight. After that, the RNA*later*® fluid was pipetted out, and the tubes with tissue samples were stored in a −80°C freezer until RNA extraction.

RNA Extraction and Integrity

Biopsy samples were homogenized with Precellys 25 system with zirconium oxide beads (Bertin Technologies SAS, Montigny-le-Bretonneux, France). Following homogenization, RNA was extracted using the total RNA extraction protocol, with the RNeasy kit (QIAGEN Pty Ltd., Melbourne, VIC, Australia). After

[1] https://data.faang.org/home

DNAse treatment, using TURBO DNAse I, each sample was purified using the Zymo Clean and Concentrator Kit as per the manufacturer's instructions (Zymo Research, CA, United States). The RNA concentration was measured by a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, United States). Samples without the optimal 260:280 ratio, which was between 1.8 and 2.1, were excluded from the experiment. The RNA integrity was verified by Agilent Bioanalyzer (Agilent, Santa Clara, CA, United States), and only samples with an RNA integrity number (RIN) above eight (RIN $>$ 8) were used for RNA sequencing. When needed, RNA extraction was repeated to achieve this quality and integrity.

RNA Sequencing, Data Processing, and Quantification

Library preparation and RNA sequencing were performed following the standard Illumina protocols for the HiSeq platform (Illumina, San Diego, CA, United States). The library prep kit was the Illumina stranded total RNA kit with Ribo-Zero Gold (Illumina, San Diego, CA, United States). Pair-end 125-base pair (bp) sequencing was conducted across three lanes of an Illumina HiSeq 2000 v4 analyzer (Illumina Inc., San Diego, CA, United States) using standard protocols, generating approximately 60 to 100 million reads per sample. All the samples were run across all the lanes used, in order to avoid any lane effect on our dataset. The quality control procedure included removing adaptors and short reads. The software TrimGalore 0.4.5 was used for trimming adaptors and for the removal of short reads, where one of the pair-end reads was shorter than 20 base pairs[2]. Before trimming, all reads were 126-bases long, and after trimming, lengths ranged from 20 to 126. Trimming was run in paired mode to avoid unpaired reads after trimming. The quality of trimmed reads was high as evaluated with FastQC 0.11.7[3], and no quality cut off was required.

The sequencing reads were aligned to the *Bos taurus* genome assembly (UMD 3.1 assembly available in Ensembl database) using the HISAT2 v.2.1.0 (Kim et al., 2015) following the mapping evaluation by Qualimap 2.2.1 (Okonechnikov et al., 2016), reporting only known transcripts from the current bovine annotation. The "reads per kilobase per million mapped reads" (RPKM = total exon reads/mapped reads in millions $\times$ exon length in kilobase) were calculated and log2 transformed for data normalization (Mortazavi et al., 2008). To further normalize the gene expression values, we used a mixed model approach that considered the effects of library, tissue, and gene-by-tissue interaction as previously detailed (Reverter et al., 2005; Cánovas et al., 2014). In brief, the mixed model contained the sequencing library treated as a fixed effect, while the interaction of tissue, gene, and animal were fitted as random effects. Fitting this animal, gene and tissue interaction is a robust methodology, commonly used in gene expression experiments to reduce the noise. We were able to fit tissue as we had three different tissues per animal: TP, HE, and TE. The VCE6 software[4] was used to solve the mixed model equations and to estimate variance components associated with random effects. The normalized gene expression values were used in all subsequent analyses, including differential gene expression.

Differential Expression Analysis

To identify differentially expressed (DE) transcripts (protein-coding and small non-coding RNAs) in specific regions of the epididymis (head and tail) and in the testis (testicular parenchyma), we carried out pairwise comparisons among the epididymis (head and tail) and testis (testicular parenchyma) tissues.

Testis and epididymis expression data comprised over 21,000 transcripts, with at least 10 counts per million reads in the data. Among expressed transcripts, 20,155 were small non-coding RNAs (miRNAs, snRNAs, and snoRNAs) and protein-coding RNAs (mRNA); for more detail, see **Figure 1**. Prior to differential expression analysis, transcripts with less than two RPKM in at least three samples were removed. After filtering, we considered 17,221 transcripts for differential expression and subsequent analyses, which investigated the co-expression relationships between protein-coding RNA (mRNA) and small non-coding RNAs (miRNAs and snoRNAs) in testicular and epididymis tissues. We performed differential expression analysis contrasting the three tissues sampled, in pairwise comparisons: HE vs. TE, HE vs. TP, and TE vs. TP. To identify the DE transcripts (mRNAs and small RNAs), we used the Limma package in R (Ritchie et al., 2015) to compute the moderated *t*-statistics, using the empirical Bayes methods (eBayes) and the default parameters. The DE transcripts with adjusted *P* value $\leq$ 0.05 (Benjamini and Hochberg, 1995) and fold change $\geq$ 2 were considered significant. We generated three lists of DE transcripts, one for each pairwise comparison: HE/TE, HE/TP, and TE/TP.

Functional Enrichment Analysis of DE Transcripts

The three lists of DE transcripts were the target lists for functional enrichment analyses. We performed the enrichment analysis using the ClueGO v. 2.5.1 bioinformatics tool (Bindea et al., 2009), a plug-in of the Cytoscape software (Shannon et al., 2003). The background gene list for functional enrichment was based on the *Bos taurus* genome, available as a default database in ClueGO. In this analysis, we identified the gene ontology terms (GO terms) and pathways [from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database] that were over-represented in the target DE list. Redundant GO terms were clustered, considering a Kappa score = 0.4, and adjusted *P* values $\leq$ 0.05 (Bonferroni step-down method) were observed when reporting on significant GO terms or pathways. To improve the functional annotation of the DE transcripts, we cross-checked these lists with the manually curated database for bovine transcription factors (TF) (de Souza et al., 2018).

[2]https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

[3]https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[4]https://www.openagrar.de/servlets/MCRFileNodeServlet/openagrar_derivate_00022208/

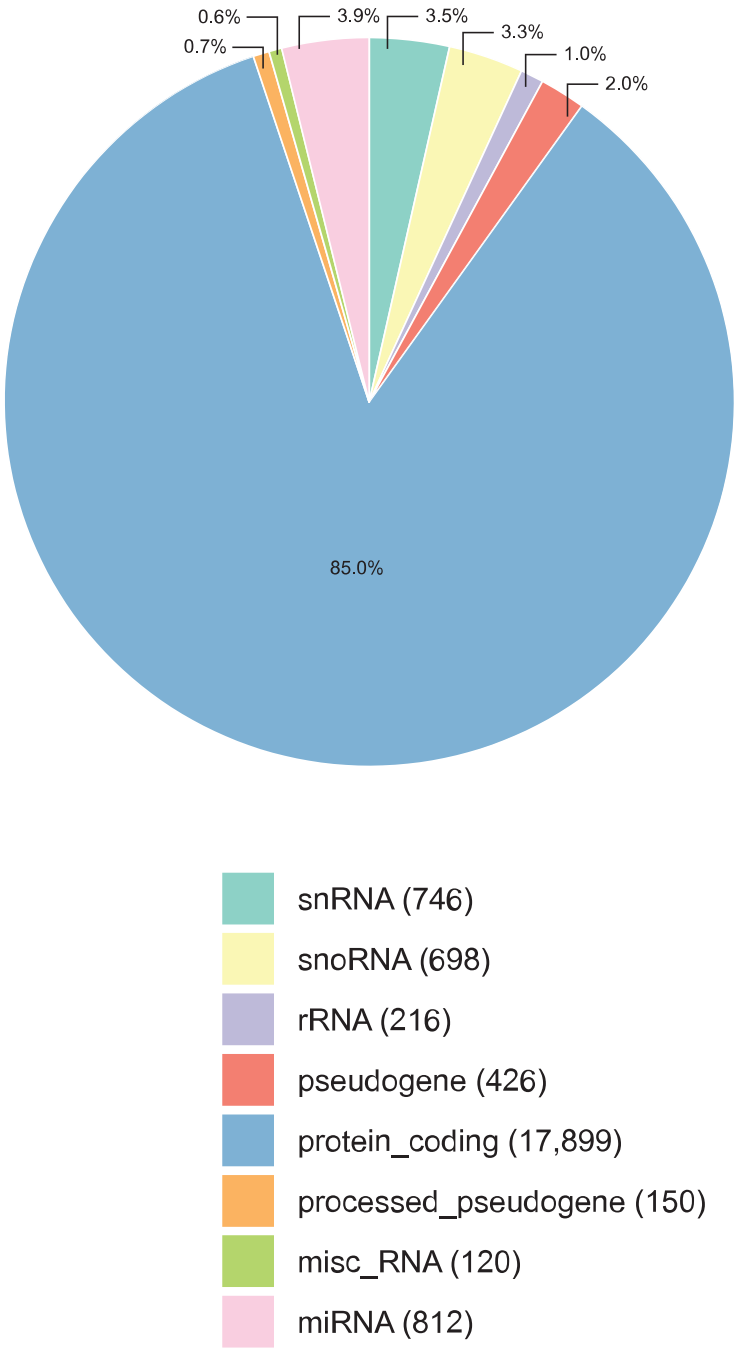


FIGURE 1 | Percentage of transcripts detected per annotation category, in all samples of the male reproductive tract (epididymis and testis) collected from four *Bos indicus* bulls. The majority of the transcripts detected were protein coding mRNAs. The absolute number of transcripts detected per category is given in parenthesis, next to its classification. Three types of small RNAs were detected: micro RNAs (miRNAs), small nuclear RNAs (snRNAs), and small nucleolar RNAs (snoRNAs). Misc_RNA stands for miscellaneous types of RNAs.

miRNA Target Genes Prediction

We predicted the target genes for the DE miRNA using the TargetScan function of R package hoardeR[5]. This function uses all the information stored in the database targetscan.org (release 7.2) that is available for the *Bos taurus* genome in terms of miRNA data. TargetScan predicts the targets of miRNAs by searching for the presence of conserved 8mer, 7mer, and 6mer sites that match the seed region of each miRNA (Lewis et al., 2005). Release 7 of TargetScan uses an improved method to predict targeting efficacy (the context + + model) (Agarwal et al., 2015), uses 3′ UTR

[5]https://cran.r-project.org/web/packages/hoardeR/index.html

profiles that indicate the fraction of mRNA containing each site (Nam et al., 2014), and uses updated miRNA families curated by Chiang et al. (2010) and Fromm et al. (2015). Of note, TargetScan is limited to known sites and 3′ UTR profiles and so it cannot predict all possible interactions between miRNA and target genes.

Co-expression Network Analysis

We performed a co-expression analysis with the log-normalized expression values of all transcripts (mRNAs and small RNAs that were DE) using the partial correlation and information theory (PCIT) algorithm (Reverter and Chan, 2008). Among the significant correlations according to PCIT, we prioritized the most extreme correlations (higher than 0.95 or lower than -0.95) as stronger evidence of interaction between transcripts. These significant and extreme correlations were used to construct a co-expression network, visualized with the Cytoscape software (Shannon et al., 2003). In the network, we marked as attributes the small RNAs (miRNAs, snoRNA, and snRNA), transcription factors (TFs), the tissue comparison in which the transcript was DE, and the small RNAs presenting significant regulatory impact factor (RIF) values for at least one tissue comparison (Reverter et al., 2010). Also, we pointed out hub transcripts in the network. Hubs are transcripts with higher than the average significant correlations, beyond two standard deviations (i.e., hubs are hyper-connected). In the same context, hub centrality elements are transcripts in the network with higher betweenness centrality than the average (more than two standard deviations), meaning that they tend to link different parts of the network.

Regulatory Impact Factor Analysis

The regulatory impact of each DE small RNAs over the DE genes for the same comparison analysis was estimated with the RIF algorithm (Reverter et al., 2010). The original application of the RIF algorithm was to determine the regulatory impact of TFs over selected genes (targets) related to a given trait through their expression values between contrasting groups (Reverter et al., 2010). In our experiment, for each pairwise comparison between sampled tissues, we used the RIF algorithm to determine the regulatory impact of each DE small RNA over the DE genes that were identified in the same pairwise comparison. For example, DE small RNA in the TP/HE comparison were tested as potential regulators of the DE genes identified in the TP/TE analyses. The RIF algorithm was selected for this analyses as its predicted regulatory roles have been showcased and validated in previous studies (Bottje et al., 2017; Nolte et al., 2019). A limitation of our analyses is that *in vitro* validations for the predicted co-expression and regulatory relationships were beyond the project scope. To mitigate this limitation, we used RIF in combination with PCIT, TargetScan, and *in silico* analyses of the minimum free energy of miRNA-target hybridization.

Minimum Free Energy: miRNA and Target Hybridization

The miRNA that were DE, significant according to RIF, and had potential targets with negative co-expression correlations were hypothesized as down-regulators of their targets. When the hypothesis was supported by the identification of binding sites confirmed with TargetScan, the miRNA were subjected to a final analysis: we estimated the minimum free energy (mfe) of the hybridization between the selected miRNA and their confirmed target genes, using RNAhybrid tool (Rehmsmeier et al., 2004). For this, we retrieved the miRNA mature sequence from the miRBase sequence database[6] and the cDNA sequences of the genes from BioMart (Durinck et al., 2009). A transcript with a mfe less than -20 kcal/mol can be considered a potential target for the miRNA in question (Yen et al., 2019).

RESULTS

Samples from the head and tail epididymis (HE and TE) and the testicular parenchyma (TP) of *Bos indicus* bulls were used for RNA sequencing. A total of 3,826 DE transcripts (mRNAs and small RNAs) were identified across the tissues in three pairwise comparisons: HE/TE, HE/TP, and TE/TP. A co-expression network was predicted and analyzed, with emphasis on investigating potential regulators of DE genes in these tissues. The network was enriched for male gamete generation and so we infer that the potential regulators of the identified DE genes might contribute to spermatogenesis.

Transcript Expression Patterns in Male Reproductive Tissues

Reads from RNA-sequencing of HE, TE, and TP were mapped to the genome, and the expression data was summarized per transcript category (**Figure 1**). All samples considered, the RNA sequencing data comprised of 85.0% mRNAs (17,899) and 10.7% small RNAs, including 812 miRNAs, 746 snRNAs, and 698 snoRNAs. In the bovine reference genome, approx. 13% of all transcripts are small RNAs, and so this is not too far from the 10% identified here. Ribosomal RNA (rRNA) were not well represented as expected in view of the library preparation methods. The library preparation allowed quantifying the expression of mRNAs and small RNAs, but it is also a limitation of this study since it did not enrich for small RNAs and no discovery of small RNAs was conducted. Mitochondrial RNA is not included in **Figure 1** because they were less than 1% of the distribution. After the quality control, we kept 17,221 transcripts expressed that were quantified across tissues for all subsequent analyses. The expression pattern of TP samples was different from the epididymis samples (both HE and TE) according to the principal component analysis (PCA) performed, see **Supplementary Figure 1**.

Differentially Expressed Transcripts and Functional Enrichment Analysis

The number of DE transcripts identified (FDR $\leq$ 0.05 and log2 fold-change $>$ 2) in each pairwise comparison between HE, TE, and TP are reported in **Table 1**. The full details on all DE transcripts are provided in **Supplementary Table 1**. In

[6]http://www.mirbase.org/

TABLE 1 | Summarized differentially expressed (DE) genes and small RNAs in each comparison of male reproductive (epididymis and testis) tissues of *Bos indicus* bulls.

Differentially expressed (DE)	HE/TE[1]	HE/TP[2]	TE/TP[3]	Total
mRNA	268	2,614	2,761	3,442
miRNA	10	108	111	144
snRNA	12	129	127	162
snoRNA	4	56	64	78
Total	**294**	**2,907**	**3,063**	**3,826**

[1]Comparison between HE and TE; [2]comparison between HE and TP; [3]comparison between TE and TP. HE, head epididymis; TE, tail epididymis; TP, testicular parenchyma.

Supplementary Table 1, positive and negative signals of the log-transformed fold change indicate if the transcript is up- or down-regulated for the first tissue in each comparison (for HE/TE and HE/TP comparisons, a positive fold change represents up-regulation in HE; in TE/TP comparison, a positive fold-change means the transcript was up-regulated in TE). **Figure 2** showcases the transcript expression patterns as volcano plots with the fold change plotted against the significance for each transcript, in each of the comparisons. We identified 40 DE transcripts that were in common for all the comparisons (**Figure 2D** and **Supplementary Table 2**). Our DE analysis identified a total of 3,826 transcripts that were DE in at least one of the three comparisons.

Enriched GO terms and KEGG pathways for the total of 3,826 DE transcripts are shown in **Figure 3** (details in **Supplementary Table 3**). The DE genes identified between HE and TE formed a target list that was enriched for nine GO terms and four KEGG pathways. The most significant GO term in the HE/TE comparison was *detoxification of copper ion* (corrected $P = 2.62 \times 10^{-7}$). DE genes identified between HE and TP were enriched for 46 GO terms and one KEGG pathway. In the third comparison, TE/TP, the DE genes were enriched for 36 GO terms and two KEGG pathways. When TP was compared to the epididymis regions, some of the most significant GO terms were *gated channel activity*, *cellular protein modification*, *male gamete generation*, *neuroactive ligand-receptor interaction*, *spermatogenesis*, and *acrosomal vesicle* (**Supplementary Table 3**).

Small RNAs With Regulatory Potential and Co-expression Networks

Among the 3,826 DE transcripts, 384 were small RNAs and 3,442 were mRNA genes. We identified 71 small RNAs that might modulate the DE genes, according to the significant RIF score (RIF 1 or 2 higher than | 1.96|; **Supplementary Table 4**). The expression pattern of these 71 small RNAs with regulatory potential differed between samples, across the male reproductive tract (**Figure 4**). Overall, we observed that small RNAs showed an expression pattern in the testis that was different from their epididymis expression. The difference between the head and tail epididymis was less pronounced, and this is similar to the PCA results for all transcripts.

The co-expression network was inferred using significant correlations ($> |0.95|$). This meant that 3,639 transcripts were nodes linked by 175,052 edges in the network, which is available as a Cytoscape file (**Supplementary Material**, cys file). The co-expression network was formed by multiple clusters, not all connected to each other (**Supplementary Figures 2**, **3**). A larger cluster with 2,216 transcripts connected through 158,807 edges was the prominent feature in the network. This large cluster was functionally enriched for *male gamete generation*, *germ cell development*, and *sperm capacitation*, among other GO terms (**Figure 5** and **Supplementary Table 5**).

In the large cluster of the co-expression network, the RIF significant small RNA bta-mir-2886 was the hub. Significant correlations suggested 601 co-expressed transcripts for bta-mir-2886, including genes and other small RNA. This is the highest number of connections for a RIF regulator in the network (**Supplementary Table 4**). Considering the connections between these 601 potential targets and their first neighbors, we observed a total of 1,035 transcripts that were directly or indirectly linked to bta-mir-2886 in the network (**Supplementary Table 6**). Among the 601 directly co-expressed transcripts, we identified one isoform of U4 spliceosomal RNA with significant RIF value and 38 transcription factors (TF).

Considering the first neighbors of bta-mir-2886 in the network, we identified a total of 241 negative correlations and 360 positive correlations. Among the negative correlations, 204 were mRNA genes. Only three of these potential targets were confirmed to have a site for hybridization with bta-mir-2886 according to TargetScan. The confirmed targets were *ELOVL3*, *FEZF2*, and *HOXA13*. All three had a mfe that is further evidence for bta-mir-2886 acting as their down-regulator: −28.5 kcal/mol for *ELOVL3*, -34.95 kcal/mol for *FEZF2*, and -35.75 kcal/mol for *HOXA13*. All TargetScan results for bta-mir-2886 are provided in **Supplementary Table 7**. TargetScan analyses of all DE miRNAs were performed and provided evidence for 1,846 DE genes that can be proposed as targets of miRNA regulation in the male reproductive tract. However, TargetScan analyses could not explain all the co-expression observed between miRNA and genes. This result is expected, since co-expression is not necessarily caused by direct hybridization and regulation, as there are many – and complex – molecular mechanisms that can lead to co-expression (Fionda, 2019).

DISCUSSION

Transcript Expression Patterns in Male Reproductive Tissues

In this study, we identified 17,221 transcripts quantified in bovine samples of the head and tail epididymis and the testicular parenchyma (HE, TE, and TP). This amounts to 64% of the genome (or 17,221 of 26,740 transcripts). The expression of a relatively large number of genes and small RNAs confirmed previous reports that suggest the testis as a good sample for functional genome annotation. In humans, the testis expressed a larger number of genes in comparison to other tissues (Uhlén et al., 2015). Harhay et al. (2010) has shown a cluster of genes exclusively expressed in the testis of bovine. The data we reported on is available through the Functional Annotation of Animal Genomes (FAANG)

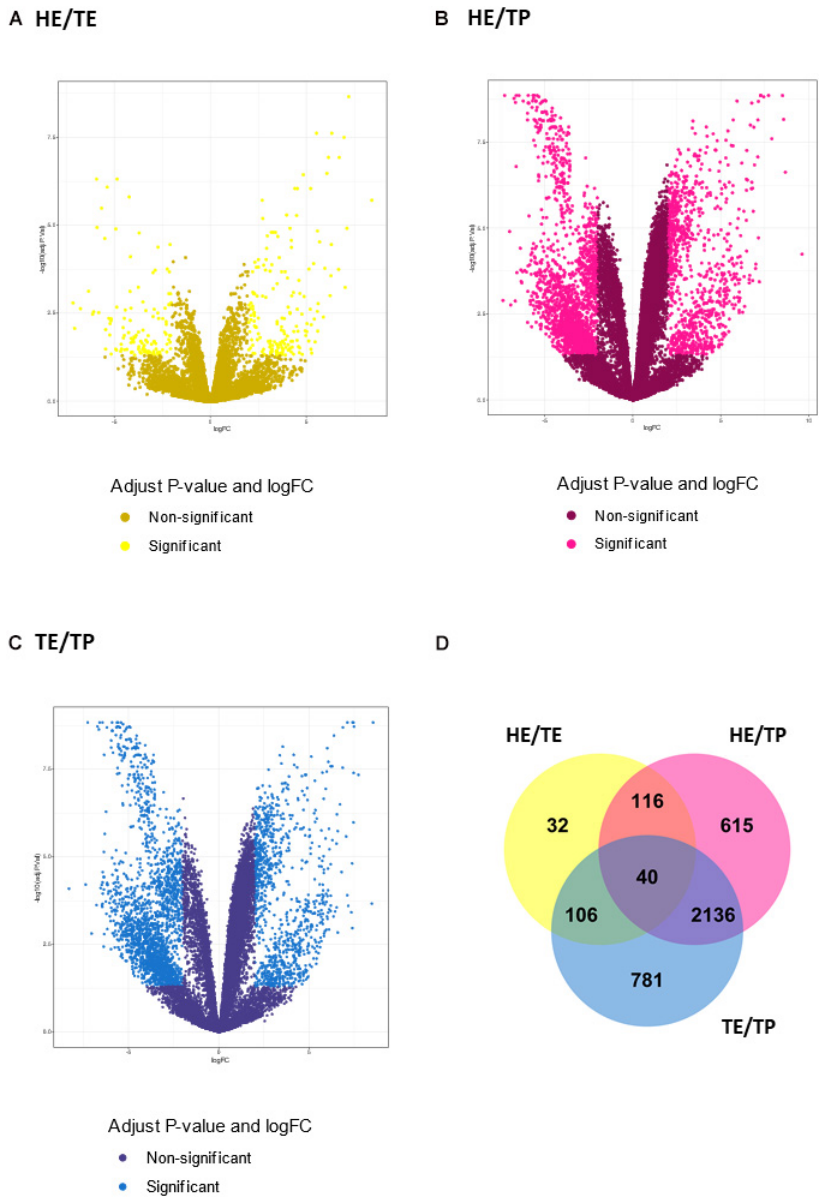


FIGURE 2 | Transcript expression patterns displayed as volcano plots with the log2 fold change in the *x*-axis and the –log10 (*P* value) in the *y*-axis for three pairwise comparisons: **(A)** quantified transcripts in the head versus tail epididymis comparison (HE/TE, in yellow). **(B)** quantified transcripts in the head epididymis versus testicular parenchyma comparison (HE/TP, in pink). **(C)** quantified transcripts in the tail epididymis versus testicular parenchyma comparison (TE/TP, in blue). In **(A–C)**, the light shade dots represent the significantly different genes, while the dark shade dots are not significant. In **(D)**, a Venn diagram summarizes the significantly different transcripts identified in each comparison, including their overlaps.

Consortium for future reference and further use (see text footnote 1). Herein, we focus on differential expression analyses, functional enrichment, co-expression network analyses, and regulatory impact metrics (RIF and additional *in silico* tests) that point to potential modulators of transcription in male reproductive tissues.

We identified 71 small RNAs with significant RIF values, interpreted as potential contributors to the modulation of over 3,000 DE transcripts. The expression patterns of these 71 potential regulators were similar to the overall pattern observed, with the testis expression contrasting with the epididymis expression. The expression patterns in all epididymis samples were relatively similar. These expression patterns might reflect the specific function and distinguished cell populations of the studied samples. In boar samples, Guyonnet et al. (2009) has observed different patterns of gene expression in the testis and epididymis. A different role for small regulatory RNAs has been proposed, specific for each region of the male reproductive system, associated with regional function. For example, Guyonnet et al. (2009) has found that genes related

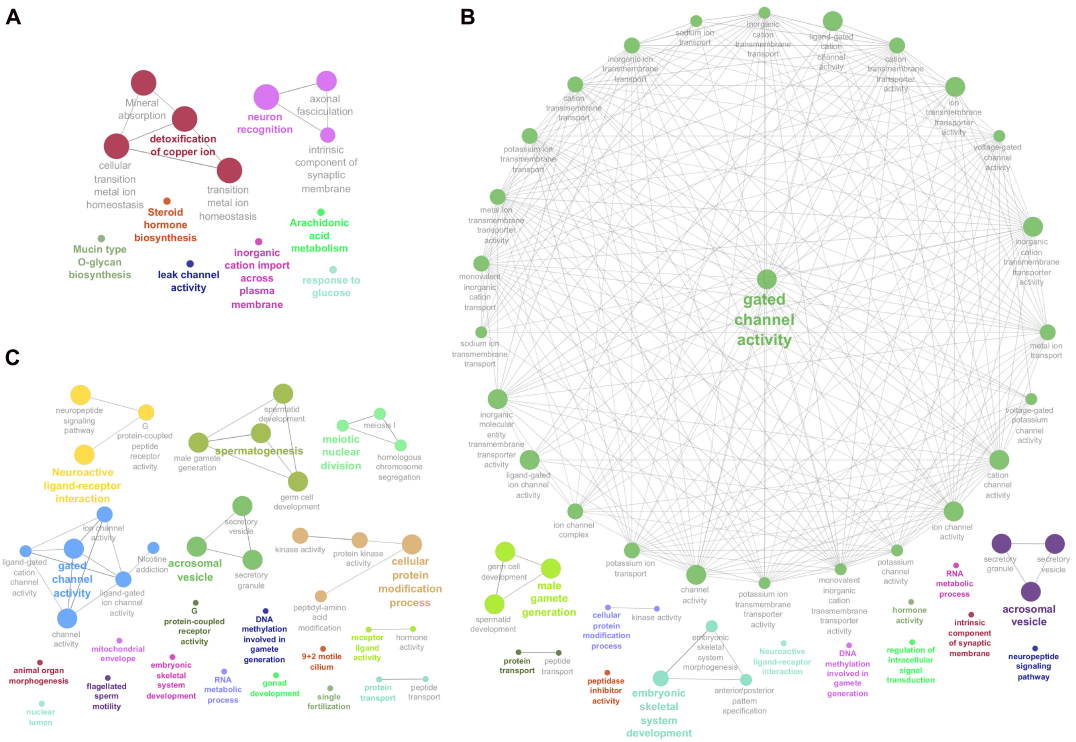


FIGURE 3 | Enriched Genome Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways identified by the ClueGO software for each comparison of differentially expressed (DE) analysis between the male reproductive tissues (HE/TE: head vs. tail; HE/TP: head vs. testicular parenchyma; TE/TP: tail vs. testicular parenchyma) of *Bos indicus* bulls. **(A)** Enriched GO terms and KEGG pathways of HE/TE comparison. **(B)** Enriched GO terms and KEGG pathways of HE/TP comparison. **(C)** Enriched GO terms and KEGG pathways of TE/TP comparison. GO terms and KEGG pathways are represented by circles.

to spermatogenesis are more prominent in the testis, compared to the epididymis. Their observation corroborates our findings. In the HE/TE comparison, only two of the 71 small RNAs detected with RIF were DE and relatively fewer genes were DE. Still, the epididymis regions have different roles in the biological processes involved in sperm maturation and transit, and this has been linked to the regionalization of gene expression patterns (Guyonnet et al., 2009; Belleannée et al., 2012; Browne et al., 2015). Our results indicated that specific small RNAs might play regulatory roles that contribute to the regionalization of gene expression in the reproductive system of *Bos indicus* bulls.

Detoxification of Copper Ion in the Head Epididymis

The regionalization of gene expression in the reproductive system can be discussed in light of the enriched GO terms and pathways associated with the DE transcripts. The most significant GO term in the HE/TE comparison was *detoxification of copper ion*. Genes associated with *detoxification of copper ion* were up-regulated in the head, including three members of the metallothionein family: *MT1A*, *MT2A*, and *MT1E* (Wong et al., 2017). These genes are involved in metal homeostasis and metal detoxification (Schulkens et al., 2014) and can protect cells from oxidative stress (Sutherland and Stillman, 2011). The expression of *MT1A*, *MT2A*, and *MT1E* is up-regulated in the presence of copper in adult human prostatic cell lines (Cu treatment vs. untreated) (Bigagli et al., 2010). In dogs with hepatitis, *MT1A* and *MT2A* expression levels decrease together with a copper concentration in hepatic cells (Dirksen et al., 2017). Protection from oxidative stress is important for sperm cells; in fact, the molecular environment of the epididymis is crucial for sperm maturation and capacitation (Belleannée et al., 2012). Increased dietary copper is associated to improved spermatozoa mobility and quality in bulls (Hidiroglou, 1979). However, high levels of copper can affect cell homeostasis and be detrimental to sperm quality and its fertilization capacity (Roblero, 1996). High levels of copper can disturb the integrity of the epididymis and affect sperm maturation (Xu et al., 1985). In this context, our results point to *MT1A*, *MT2A*, and *MT1E* as genes that may assist with copper homeostasis in the head epididymis and might have a role in sperm maturation.

The regionalization of gene expression in the male reproductive system is likely a consequence of regulatory mechanisms, including small RNAs that target genes post-transcriptionally. Bta-mir-362 had a negative co-expression correlation with *MT1E* and *MT1A* (lower than −0.95) and might down-regulate these genes. Bta-mir-362 is reported to contribute to spermatogenesis processes in pigs too (Ran et al., 2018). In short, this miRNA might modulate genes involved with *detoxification of copper ion* in the epididymis, and as a consequence, it might affect sperm quality.

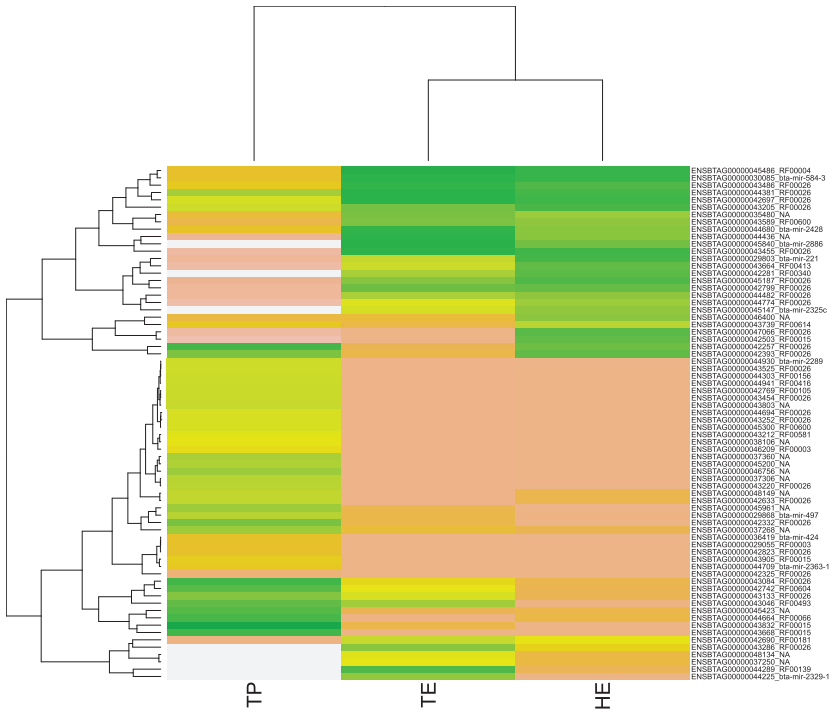


FIGURE 4 | Heatmap of pattern gene expression of small RNAs (snRNAs and miRNAs) with regulatory potential, according to significant values in the regulatory impact factor (RIF) metric. Small RNAs are shown on the *Y*-axis and the total of expression per biopsies of epididymis head (HE), epididymis tail (TE), and testicular parenchyma (TP). The colors correspond to the intensity of expression per tissue. The intensity is scaled by the colors white, yellow, and green. The low expression is represented by white, followed by salmon, yellow, and green (high expression).

Gated Channel Activity, Ions, and Water Transport

In our study, the GO term *gated channel activity* and other terms related to ion transport and channel activity were significant for the comparisons between the epididymis and testis. In both the HE/TP and the TE/TP comparisons, DE genes suggest that ion channels are relevant to spermatogenesis. This result is in agreement with previous knowledge, because ions such as Ca^{2+} and Na^{+} contribute to the acrosomal reaction, hyper-activation, sperm capacitation, and sperm quality (Mirnamniha et al., 2019).

Differentially expressed and enrichment analyses suggest regionalization of expression patterns in the male reproductive system for genes that code proteins related to ion channels. Specific proteins related to ion channels and solute transporters are responsible for the epididymis homeostasis and for the luminal environment that is adequate to sperm maturation (Belleannée et al., 2012). In short, ion channels, anions, cations, and water transport molecules (i.e., AQPs) are involved in the control of the luminal fluid (Browne et al., 2015).

The DE genes *ATP6V0A4*, *ATP6V0D2*, *ATP6V1G3*, *CFTR*, *SCNN1G*, and *SCNN4A* are related to ion channel activity and were up-regulated in the epididymis when compared to the testis. The *ATP6V0A4*, *ATP6V0D2*, and *ATP6V1G3* genes were up-regulated in both HE/TP and TE/TP comparisons. They code for proteins that compose the subunit of vacuolar H + – ATPase (V-ATPase) (Wagner et al., 2004). The V-ATPase is a multi-subunit ATP-driven proton pump (Pamarthy et al., 2018), with influence in the acidification of luminal fluid (Brown et al., 1992; Roy et al., 2013) that may help sperm maturation (Pamarthy et al., 2018). Failure in luminal fluid acidification can result in poor sperm maturation and infertility (Breton et al., 1996). In this context, a higher expression of these genes in the epididymis suggests they are relevant to sperm maturation in *Bos indicus* bulls.

The cystic fibrosis transmembrane conductance regulator (*CFTR*) gene was up-regulated in the head epididymis when compared to the testis. The CFTR channel is involved in sperm capacitation (Touré, 2019) and can contribute to the Cl^{-} and bicarbonate fluxes (Touré, 2019). The DE gene *SLC26A3*, up-regulated in the epididymis, is essential to bicarbonate fluxes and interacts with the CFTR channel (Touré, 2019). The *SLC26A3* knockout mice present lesions in the epididymis and sperm reduction (El Khouri et al., 2018). CFTR and epithelial Na^{+} channel (ENaC) contribute to sperm capacitation (Visconti et al., 2011; Sharma and Hanukoglu, 2019). ENaC is probably involved in the uptake of Na^{+} ions from the epididymal lumen into the cells. Like CFTR channels, ENaC channels are observed in patterns along the length of the mouse and rat epididymis (Sharma and Hanukoglu, 2019). Two genes that code for ENaC proteins, *SCNN1G* and *SCNN4A*, were up-regulated in the epididymis when compared to the testis. *SCNN1G* was up-regulated in the head (HE/TP) while *SCNN4A* was up-regulated

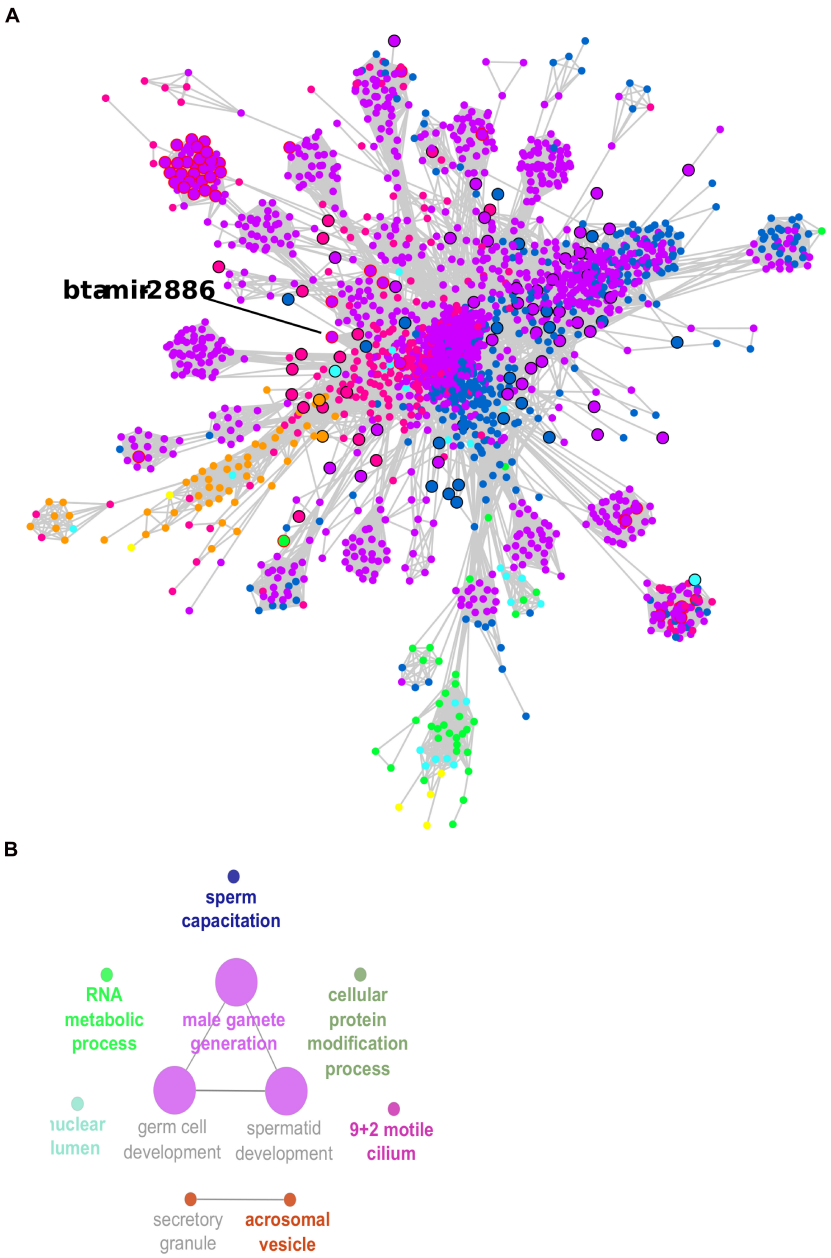


FIGURE 5 | Co-expression network large cluster of differentially expressed (DE) transcripts enriched for male gamete generation. **(A)** Node sizes are proportional to the number of connections (highly connected nodes appear bigger). The nodes are the DE transcripts and the edges are the significant correlations (partial correlation and information theory, PCIT > | 0.95|). Black borders represent hub transcripts. Red borders represent small RNAs with significant regulatory impactor factor values. Node color: yellow for DE only in the head and tail epididymis comparison (HE/TE); pink for DE only in the head epididymis and testicular parenchyma comparison (HE/TP); blue for DE only in the tail epididymis and testicular parenchyma comparison (HE/TP); orange for DE in the HE/TE and the HE/TP comparisons; green for DE in the HE/TE and the TE/TP comparisons; purple for DE in the HE/TP and the TE/TP comparisons; and turquoise for DE in all comparisons. **(B)** Enriched GO terms and KEGG pathways for the cluster represented in **(A)**.

in both the head (HE/TP) and tail (TE/TP). The regional regulation of genes coding for ENaC proteins might be an evidence of their role in sperm maturation. The bovine DE patterns discussed here conform to the expectations from studies in other species and might reveal mechanisms that are relevant to male fertility across mammals.

Another gene that was up-regulated in the head and tail epididymis when compared to the testis was *AQP9*, an aquaporin. Aquaporins (AQPs) are channels of proteins that facilitate the movement of water across the plasma membrane and contribute to epididymal sperm concentration (Belleannée et al., 2012; Schimming et al., 2017). *AQP9* has been previously reported as

expressed in the epididymis and related to water resorption in the proximal epididymis (Belleannée et al., 2012). Gene expression patterns of AQPs have been related to disorders of the male reproductive system in mammals (Huang et al., 2006). Therefore, we hypothesize that *AQP9* might affect water transport in the epididymis in *Bos indicus*.

Small RNAs might modulate DE genes involved in *gated channel activity*, ion channels, and water transport. We predicted interactions between 24 small RNAs (15 miRNAs and 9 snRNA) and the DE genes discussed above. For example, bta-mir-2886 was co-expressed with *AQP9* and *CFTR*. The snRNA RF00026 (U6 spliceosomal RNA) was co-expressed with *ATP6V0D2*, *AQP9*, *CFTR*, and *SLC26A3*. It is possible then that bta-mir-2886 and RF00026 among other small RNAs modulate *gated channel activity*, ions, and water transport, affecting epididymis function.

Neuroactive Ligand-Receptors and Spermatogenesis

Among the DE transcripts, we identified genes related to prolactin, GABA, and muscarinic acetylcholine receptors, all part of the enriched *neuroactive ligand-receptor interaction* pathway. Overall, this pathway seems more important to testicular function than to the epididymis, with a few exceptions as discussed below.

Prolactin is a peptide hormone that acts via its transmembrane receptor (Raut et al., 2019). In our study, nine DE genes related prolactin signaling – *PRL*, *PRLH*, *PRP1*, *PRP14*, *PRP2*, *PRP4*, *PRP6*, *PRP8*, and *PRP9* – were up-regulated in the testis when compared to the epididymis. Prolactin receptors are expressed in the testis of humans (Jabbour and Kelly, 1997) and bulls (Pratt et al., 2015). Prolactin signaling affects the male reproduction system (Hermanns and Hafez, 1981) as it interferers with steroidogenesis and spermatogenesis (Jabbour and Kelly, 1997). Also, prolactin signaling relates to testosterone concentration (Franchimont, 1983). In summary, the DE analyses recapitulated some known biology of testicular function and suggested prolactin genes that were regulated in male tissues.

GABA receptors were previously reported in the testis and sperm of mice (He et al., 2001, 2003). We identified 10 GABA receptors as DE genes in the HE/TP and TE/TP comparisons: *GABBR2*, *GABRA2*, *GABRA3*, *GABRA4*, *GABRA5*, *GABRB1*, *GABRB3*, *GABRG2*, *GABRP*, and *GABRQ*. Three receptors (*GABRG2*, *GABRB1*, and *GABRA5*) had higher expression in the epididymis, while the other seven were up-regulated in the testis. In the testis, GABA receptors affect the Leydig cell function, influence germ cell maturation (Geigerseder et al., 2003), and regulate spermatogenesis (Geigerseder et al., 2003; Kanbara et al., 2005; Du et al., 2013). The function of these DE GABA receptors, with expression that is specific to each region of the male system examined, requires further research.

Three DE genes, *CHRM1*, *CHRM2*, and *CHRM3*, are receptors related to muscarinic acetylcholine signaling. Muscarinic acetylcholine receptors (or MAChRs) are part of the regulatory mechanisms in the male reproductive system (Borges et al., 2001; Avellar et al., 2010). MAChRs regulate testicular cell function (Borges et al., 2001) and can influence the luminal fluid composition (Avellar et al., 2010). In our study, *CHRM1*, *CHRM2*, and *CHRM3* were up-regulated in the testis compared to the epididymis, and so we suggested that MAChRs might contribute to testicular function in *Bos indicus* bulls.

Among DE genes of the *neuroactive ligand-receptor interaction* pathway, 10 were connected to the larger network cluster and directly or indirectly linked to bta-mir-2886. The prolactin signaling genes *PR2*, *PRP14*, and *PRP9* were predicted to interact directly with bta-mir-2886, while *GABRA5* was a second neighbor of this same miRNA. In short, it is possible that bta-mir-2886 and other small RNAs regulate genes in the *neuroactive ligand-receptor interaction* pathway that might affect spermatogenesis.

Co-expression Network, Small RNAs, and Male Gamete Generation

Differentially expressed transcripts associated with the GO term *male gamete generation* were enriched in the comparisons between the epididymis and testis, being the third most significant term in both HE/TP and TE/TP comparisons. This same GO term was significant for transcripts in the larger cluster of the co-expression network. Four well-known regulators of spermatogenesis were among the DE transcripts associated with *male gamete generation*: *RFX2*, *HORMAD1*, *CCDC36*, and *DAZL*. The gene *RFX2* is an essential transcription factor in the regulation of spermatogenesis (Kistler et al., 2015; Pandey et al., 2019), which is expressed in spermatocytes and spermatids in mice (Pandey et al., 2019). The *RFX2*-deficient mice have completely blocked spermatogenesis (Kistler et al., 2015). *HORMADA1* is key during the meiosis and it possibly interacts with *CCDC36* (Stanzione et al., 2016). *DAZL* stands for "deleted in azoospermia like," and it codes for a RNA-binding protein that is localized to the nucleus of spermatogonia, but relocates to the cytoplasm during meiosis, where it persists in spermatids and spermatozoa. *DAZL* is highly expressed in the testis of sheep with sexual maturity (Yuan et al., 2020) and may have a role in sex differentiation (Rossitto et al., 2015). All four genes were up-regulated in the testis when compared to the epididymis, which is expected since *male gamete generation* is a characteristic of the testis. Two of these well-known regulators, *RFX2* and *CCDC36*, were also nodes in the larger network cluster.

In the larger network cluster, five out of 10 significant GO terms were very specific to spermatogenesis: *male gamete generation*, *spermatid development*, *germ cell development*, *acrosomal vesicle*, and *sperm capacitation.* Therefore, the small RNAs that were identified as potential regulators of this cluster of highly connected DE genes might be proposed as potential regulators of spermatogenesis. We identified 228 small RNAs in the larger network cluster and 43 of these had significant RIF values; they were 20 snRNAs, 8 snoRNAs, and 15 miRNAs. One of the significant miRNAs was bta-mir-2886, up-regulated in the epididymis for two pairwise comparisons (HE/TP and TE/TP), with a high fold change in both (approx. 7). We propose that bta-mir-2886, through its 601 co-expressed transcripts and additional 434 first neighbors, might affect spermatogenesis in *Bos indicus* bulls.

In our study, most of the correlations between bta-mir-2886 and predicted targets were positive, including for *AQP9* and *CFTR*. We speculated that this miRNA might be indirectly regulating the expression of co-expressed transcripts, with which it presents positive correlations, by inhibiting their negative regulators. This indirect mechanism was suggested previously by Ritchie et al. (2009). We observed 204 predicted negative correlations with bta-mir-2886. Among negative correlations, we identified three genes that may be down-regulated by bta-mir-2886, which were confirmed by TargetScan and had mfe below -20 kcal/mol. They were *ELOVL3*, *FEZF2*, and *HOXA13*.

HOXA13 was among 23 genes of the Hox family that were DE in our study. Hox family transcription factors are expressed in the male reproductive tract (Lindsey and Wilkinson, 1996), including the human testis (Zhu et al., 2016) and mice epididymis (Bomgardner et al., 2003; Raines et al., 2013). Hox genes act in spermatogenesis and sperm maturation (Lindsey and Wilkinson, 1996). Zhu et al. (2016) have reported Hox genes as regulators of meiosis in the human testis. Mutations in *HOXA13* were associated to male infertility in mice (Post and Innis, 1999). Further studies could investigate the role of *HOXA13* in bull fertility.

CONCLUSION

Our results indicate that bta-mir-2886, among other small RNAs, are co-expressed with DE genes that may contribute to spermatogenesis and sperm maturation in the testis and epididymis of *Bos indicus* bulls. Although our study predicts potential regulators of gene expression in the testis and the epididymis of *Bos indicus* bulls, further work is necessary to confirm our findings and detail the roles played by small RNAs in spermatogenesis.

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: https://data.faang.org/organism, SAMEA104495807.

ETHICS STATEMENT

The animal study was reviewed and approved by the Ethics Committee of the University of Queensland, Brisbane, Australia (protocol number: ANRFA/SCMB/094/16).

AUTHOR CONTRIBUTIONS

MF conceived the idea of this study. MF, EM, and LP-N performed field and laboratory work. JE prepared and sequenced the RNA libraries. AdL, RP, JA, and AR performed the bioinformatics and data analysis. AdL, JA, and MF drafted the manuscript. All authors contributed to the interpretation of results, discussion, review of the concepts in this manuscript, and agreed to be responsible for the content of this study.

FUNDING

This work was supported by the University of Queensland provided an Early Career Research grant to MF.

ACKNOWLEDGMENTS

We acknowledge the genomics facility of the University of Queensland for preparing and sequencing the cDNA libraries used in this study. We acknowledge Marina Naval Sanchez for her assistance with submitting the RNA-sequencing data to the FAANG Consortium. Elements of this research used the Queensland node of Metabolomics Australia.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.610116/full#supplementary-material

REFERENCES

Abou-Haila, A., and Tulsiani, D. R. (2000). Mammalian sperm acrosome: formation, contents, and f Function. *Arch. Biochem. Biophys.* 379, 173–182. doi: 10.1006/abbi.2000.1880

Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4, 1–38. doi: 10.7554/eLife.05005

Avellar, M. C. W., Siu, E. R., Yasuhara, F., Maróstica, E., and Porto, C. S. (2010). Muscarinic acetylcholine receptor subtypes in the male reproductive tract: expression and function in rat efferent ductules and epididymis. *J. Mol. Neurosci.* 40, 127–134. doi: 10.1007/s12031-009-9268-6

Bartel, D. P. (2018). Metazoan MicroRNAs. *Cell* 173, 20–51. doi: 10.1016/j.cell.2018.03.006

Belleannée, C. (2015). Extracellular microRNAs from the epididymis as potential mediators of cell-to-cell communication. *Asian J. Androl.* 17, 730–736. doi: 10.4103/1008-682X.155532

Belleannée, C., Thimon, V., and Sullivan, R. (2012). Region-specific gene expression in the epididymis. *Cell Tissue Res.* 349, 717–731. doi: 10.1007/s00441-012-1381-0

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.* 57, 289–300. doi: 10.2307/2346101

Bigagli, E., Luceri, C., Bernardini, S., Dei, A., and Dolara, P. (2010). Extremely low copper concentrations affect gene expression profiles of human prostate epithelial cell lines. *Chem. Biol. Interact.* 188, 214–219. doi: 10.1016/j.cbi.2010.06.009

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinform* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Bomgardner, D., Hinton, B. T., and Turner, T. T. (2003). 5′ Hox Genes and Meis 1, a Hox-DNA binding cofactor, are expressed in the adult mouse epididymis1. *Biol. Reprod.* 68, 644–650. doi: 10.1095/biolreprod.102.009324

Borges, M. O. R., Abreu, M. L. C., Porto, C. S., and Avellar, M. C. W. (2001). Characterization of muscarinic acetylcholine receptor in rat sertoli cells. *Endocrinology* 142, 4701–4710. doi: 10.1210/endo.142.11.8465

Bottje, W., Kong, B. W., Reverter, A., Waardenberg, A. J., Lassiter, K., and Hudson, N. J. (2017). Progesterone signalling in broiler skeletal muscle is associated with divergent feed efficiency. *BMC Syst. Biol.* 11:396.

Breton, S., Smith, P. J. S., Lui, B., and Brown, D. (1996). Acidification of the male reproductive tract by a proton pumping (H+)-ATPase. *Nat. Med.* 2, 470–472. doi: 10.1038/nm0496-470

Brown, D., Lui, B., Gluck, S., and Sabolic, I. (1992). A plasma membrane proton ATPase in specialized cells of rat epididymis. *Am. J. Physiol. Cell Physiol.* 263, 913–916. doi: 10.1152/ajpcell.1992.263.4.c913

Browne, J. A., Yang, R., Leir, S. H., Eggener, S. E., and Harris, A. (2015). Expression profiles of human epididymis epithelial cells reveal the functional diversity of caput, corpus and cauda regions. *Mol. Hum. Reprod.* 22, 69–82. doi: 10.1093/molehr/gav066

Cánovas, A., Reverter, A., DeAtley, K. L., Ashley, R. L., Colgrave, M. L., Fortes, M. R. S., et al. (2014). Multi-tissue omics analyses reveal molecular regulatory networks for puberty in composite beef cattle. *PLoS One* 9:e0102551. doi: 10.1371/journal.pone.0102551

Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeung, V. C., Spies, N., Baek, D., et al. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev.* 24, 992–1009. doi: 10.1101/gad.1884710

Cornwall, G. A. (2009). New insights into epididymal biology and function. *Hum. Reprod.* 15, 213–227. doi: 10.1093/humupd/dmn055

de Souza, M. M., Zerlotini, A., Geistlinger, L., Tizioto, P. C., Taylor, J. F., Rocha, M. I. P., et al. (2018). A comprehensive manually-curated compendium of bovine transcription factors. *Sci. Rep.* 8, 1–12. doi: 10.1038/s41598-018-32146-2

Dirksen, K., Spee, B., Penning, L. C., Van Den Ingh, T. S. G. A. M., Burgener, I. A., Watson, A. L., et al. (2017). Gene expression patterns in the progression of canine copper-associated chronic hepatitis. *PLoS One* 12:e0176826. doi: 10.1371/journal.pone.0176826

Du, Y., Du, Z., Zheng, H., Wang, D., Li, S., Yan, Y., et al. (2013). GABA exists as a negative regulator of cell proliferation in spermaogonial stem cells. *Cell. Mol. Biol. Lett.* 18, 149–162. doi: 10.2478/s11658-013-0081-4

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomaRt steffen. *Nat. Protoc.* 4, 1184–1191. doi: 10.1038/nprot.2009.97.Mapping

El Khouri, E., Whitfield, M., Stouvenel, L., Kini, A., Riederer, B., Lores, P., et al. (2018). Slc26a3 deficiency is associated with epididymis dysplasia and impaired sperm fertilization potential in the mouse. *Mol. Reprod. Dev.* 85, 682–695. doi: 10.1002/mrd.23055

Fionda, V. (2019). "Networks in biology," in *Encyclopedia of Bioinformatics and Computational Biology*, eds S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach (Oxford: Academic Press), 915–921.

Fortes, M. R. S., Satake, N., Corbet, D. H., Corbet, N. J., Burns, B. M., Moore, S. S., et al. (2014). Sperm protamine deficiency correlates with sperm DNA damage in Bos indicus bulls. *Andrology* 2, 370–378. doi: 10.1111/j.2047-2927.2014.00196.x

Franchimont, P. (1983). Regulation of gonadal androgen secretion. *Horm. Res. Paediatr.* 18, 7–17. doi: 10.1159/000179774

Fromm, B., Billipp, T., Peck, L. E., Johansen, M., Tarver, J. E., King, B. L., et al. (2015). A uniform system for the annotation of human microRNA genes and the evolution of the human microRNAome bastian. *Annu. Rev. Genet.* 49, 213–242. doi: 10.1146/annurev-genet-120213-092023.A

Fuda, N. J., Ardehali, M. B., and Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* 461, 186–192. doi: 10.1038/nature08449

Geigerseder, C., Doepner, R., Thalhammer, A., Frungieri, M. B., Gamel-Didelon, K., Calandra, R. S., et al. (2003). Evidence for a GABAergic system in rodent and human testis: local GABA production and GABA receptors. *Neuroendocrinology* 77, 314–323. doi: 10.1159/000070897

Guyonnet, B., Marot, G., Dacheux, J. L., Mercat, M. J., Schwob, S., Jaffrézic, F., et al. (2009). The adult boar testicular and epididymal transcriptomes. *BMC Genom.* 10:369. doi: 10.1186/1471-2164-10-369

Harhay, G. P., Smith, T. P. L., Alexander, L. J., Haudenschild, C. D., Keele, J. W., Matukumalli, L. K., et al. (2010). An atlas of bovine gene expression reveals novel distinctive tissue characteristics and evidence for improving genome annotation. *Genome Biol.* 11:102. doi: 10.1186/gb-2010-11-10-r102

He, X. B., Hu, J. H., Wu, Q., Yan, Y. C., and Koide, S. S. (2001). Identification of GABAB receptor in rat testis and sperm. *Biochem. Biophys. Res. Commun.* 283, 243–247. doi: 10.1006/bbrc.2001.4732

He, X., Zhang, Y., Yan, Y., Li, Y., and Koide, S. S. (2003). Identification of GABABR2 in rat testis and sperm. *J. Reprod. Dev.* 49, 397–402. doi: 10.1262/jrd.49.397

Hermann, B. P., Cheng, K., Singh, A., Roa-De, La Cruz, L., Mutoji, K. N., et al. (2018). The mammalian spermatogenesis single-cell transcriptome, from spermatogonial stem cells to spermatids. *Cell Rep.* 25, 1650–1667. doi: 10.1016/j.celrep.2018.10.026

Hermanns, U., and Hafez, E. S. E. (1981). Prolactin and male reproduction. *Syst. Biol. Reprod. Med.* 6, 95–125. doi: 10.3109/01485018108987351

Hidiroglou, M. (1979). Trace element deficiencies and fertility in ruminants: a review. *J. Dairy Sci.* 62, 1195–1206. doi: 10.3168/jds.S0022-0302(79)83400-1

Hombach, S., and Kretz, M. (2016). "Non-coding RNAs: classification, biology and functioning," in *Non-Coding RNAs in Colorectal Cancer*, eds O. Slaby and G. A. Calin (Cham: Springer), 3–17. doi: 10.1007/978-3-319-42059-2

Huang, H. F., He, R. H., Sun, C. C., Zhang, Y., Meng, Q. X., and Ma, Y. Y. (2006). Function of aquaporins in female and male reproductive systems. *Hum. Reprod.* 12, 785–795. doi: 10.1093/humupd/dml035

Jabbour, H. N., and Kelly, P. A. (1997). Prolactin receptor subtypes: a possible mode of tissue specific regulation of prolactin function. *Rev. Reprod.* 2, 14–18. doi: 10.1530/ror.0.0020014

Kanbara, K., Okamoto, K., Nomura, S., Kaneko, T., Shigemoto, R., Azuma, H., et al. (2005). Cellular localization of GABA and GABAB receptor subunit proteins during spermiogenesis in rat testis. *J. Androl.* 26, 485–493. doi: 10.2164/jandrol.04185

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kistler, W. S., Baas, D., Lemeille, S., Paschaki, M., Seguin-Estevez, Q., Barras, E., et al. (2015). RFX2 is a major transcriptional regulator of spermiogenesis. *PLoS Genet.* 11:e1005368. doi: 10.1371/journal.pgen.1005368

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20. doi: 10.1016/j.cell.2004.12.035

Lindsey, S., and Wilkinson, M. F. (1996). Homeobox genes and male reproductive development. *J. Assist. Reprod. Genet.* 13, 182–192. doi: 10.1007/BF02072542

MacLean, J. A., and Wilkinson, M. F. (2005). Gene regulation in spermatogenesis. *Curr. Top. Dev. Biol.* 5, 131–197. doi: 10.1016/S0070-2153(05)71005-X

Marengo, S. R. (2008). Maturing the sperm: unique mechanisms for modifying integral proteins in the sperm plasma membrane. *Anim. Reprod. Sci.* 105, 52–63. doi: 10.1016/j.anireprosci.2007.11.018

Mirnamniha, M., Faroughi, F., Tahmasbpour, E., Ebrahimi, P., and Beigi Harchegani, A. (2019). An overview on role of some trace elements in human reproductive health, sperm function and fertilization process. *Rev. Environ. Health* 19, 1–10. doi: 10.1515/reveh-2019-0008

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Nam, J., Rissland, O. S., Koppstein, D., Abreu-goodger, C., Jan, C. H., Agarwal, V., et al. (2014). Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol. Cell* 53, 1031–1043. doi: 10.1016/j.molcel.2014.02.013.Global

Nixon, B., De Iuliis, G. N., Dun, M. D., Zhou, W., Trigg, N. A., and Eamens, A. L. (2019). Profiling of epididymal small non-protein-coding RNAs. *Andrology* 7, 669–680. doi: 10.1111/andr.12640

Nolte, W., Weikard, R., Brunner, R. M., Albrecht, E., Hammon, H. M., Reverter, A., et al. (2019). Biological network approach for the identification of regulatory long Non-coding RNAs associated with metabolic efficiency in cattle. *Front. Genet.* 10:1130. doi: 10.3389/fgene.2019.01130

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32, 292–294. doi: 10.1093/bioinformatics/btv566

Pamarthy, S., Kulshrestha, A., Katara, G. K., and Beaman, K. D. (2018). The curious case of vacuolar ATPase: regulation of signaling pathways. *Mol. Cancer* 17, 1–9. doi: 10.1186/s12943-018-0811-3

Pandey, A., Yadav, S. K., Vishvkarma, R., Singh, B., Maikhuri, J. P., Rajender, S., et al. (2019). The dynamics of gene expression during and post meiosis sets the sperm agenda. *Mol. Reprod. Dev.* 86, 1921–1939. doi: 10.1002/mrd.23278

Post, L. C., and Innis, J. W. (1999). Infertility in adult hypodactyly mice is associated with hypoplasia of distal reproductive structures1. *Biol. Reprod.* 61, 1402–1408. doi: 10.1095/biolreprod61.6.1402

Pradillo, M., and Santos, J. L. (2018). Genes involved in miRNA biogenesis affect meiosis and fertility. *Chromosom. Res.* 26, 233–241. doi: 10.1007/s10577-018-9588-x

Pratt, S. L., Calcatera, S. M., Stowe, H. M., Dimmick, M. A., Schrick, F. N., Duckett, S. K., et al. (2015). Identification of bovine prolactin in seminal fluid, and expression and localization of the prolactin receptor and prolactin-inducible protein in the testis and epididymis of bulls exposed to ergot alkaloids. *Theriogenology* 83, 662–669. doi: 10.1016/j.theriogenology.2014.10.031

Raines, A. M., Adam, M., Magella, B., Meyer, S. E., Grimes, H. L., Dey, S. K., et al. (2013). Recombineering-based dissection of flanking and paralogous Hox gene functions in mouse reproductive tracts. *Development* 140, 2942–2952. doi: 10.1242/dev.092569

Ran, M., Weng, B., Cao, R., Li, Z., Peng, F., Luo, H., et al. (2018). miR-362 regulates the proliferation and apoptosis of porcine immature Sertoli cells through targeting the ZNF644 gene. *Yi Chuan* 40, 572–584. doi: 10.16288/j.yczz.18-028

Raut, S., Deshpande, S., and Balasinor, N. H. (2019). Unveiling the role of prolactin and its receptor in male reproduction. *Horm. Metab. Res.* 51, 215–219. doi: 10.1055/a-0859-1144

Rehmsmeier, M., Steffen, P., Ho, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *Bioinformatics* 10, 1507–1517. doi: 10.1261/rna.5248604.and

Reverter, A., and Chan, E. K. F. (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 24, 2491–2497. doi: 10.1093/bioinformatics/btn482

Reverter, A., Barris, W., McWilliam, S., Byrne, K. A., Wang, Y. H., Tan, S. H., et al. (2005). Validation of alternative methods of data normalization in gene co-expression studies. *Bioinformatics* 21, 1112–1120. doi: 10.1093/bioinformatics/bti124

Reverter, A., Hudson, N. J., Nagaraj, S. H., Pérez-Enciso, M., and Dalrymple, B. P. (2010). Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data. *Bioinformatics* 26, 896–904. doi: 10.1093/bioinformatics/btq051

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

Ritchie, W., Rajasekhar, M., Flamant, S., and Rasko, J. E. J. (2009). Conserved expression patterns predict microRNA targets. *PLoS Comput. Biol.* 5:513. doi: 10.1371/journal.pcbi.1000513

Roblero, L. (1996). Effect of copper ion on the motility, viability, acrosome reaction and fertilizing capacity of human spermatozoa in vitro. *Reprod. Fertil. Dev.* 8, 871–874. doi: 10.1071/RD9960871

Rossitto, M., Philibert, P., Poulat, F., and Boizet-Bonhoure, B. (2015). Molecular events and signalling pathways of male germ cell differentiation in mouse. *Semin. Cell Dev. Biol.* 45, 84–93. doi: 10.1016/j.semcdb.2015.09.014

Roy, J. W., Hill, E., Ruan, Y. C., Vedovelli, L., Păunescu, T. G., Brown, D., et al. (2013). Circulating aldosterone induces the apical accumulation of the proton pumping V-ATPase and increases proton secretion in clear cells in the caput epididymis. *Am. J. Physiol. Cell Physiol.* 305, 436–446. doi: 10.1152/ajpcell.00410.2012

Schimming, B. C., Baumam, C. A. E., Pinheiro, P. F. F., de Matteis, R., and Domeniconi, R. F. (2017). Aquaporin 9 is expressed in the epididymis of immature and mature pigs. *Reprod. Domest. Anim.* 52, 617–624. doi: 10.1111/rda.12957

Schulkens, I. A., Castricum, K. C. M., Weijers, E. M., Koolwijk, P., Griffioen, A. W., and Thijssen, V. L. (2014). Expression, regulation and function of human metallothioneins in endothelial cells. *J. Vasc. Res.* 51, 231–238. doi: 10.1159/000365550

Shannon, P., Markiel, A., Owen Ozier Nitin, S., Baliga Jonathan, T., Wang, D. R., Amin, N., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13:6. doi: 10.1101/gr.1239303.metabolite

Sharma, S., and Hanukoglu, I. (2019). Mapping the sites of localization of epithelial sodium channel (ENaC) and CFTR in segments of the mammalian epididymis. *J. Mol. Histol.* 50, 141–154. doi: 10.1007/s10735-019-09813-3

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., et al. (2013). Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* 3, 2179–2190. doi: 10.1016/j.celrep.2013.05.031

Stanzione, M., Marek, B., Frantzeskod, P., Dereli, I., Julian, L., Ramlal, A., et al. (2016). Meiotic DNA break formation requires the unsynapsed chromosome axis-binding protein IHO1 (CCDC36) in mice. *Nat. Cell Biol.* 18, 1208–1220. doi: 10.1038/ncb3417

Staub, C., and Johnson, L. (2018). Review: Spermatogenesis in the bull. *Animal* 12, s27–s35. doi: 10.1017/S1751731118000435

Sullivan, R. (2016). Epididymosomes: role of extracellular microvesicles in sperm maturation. *Front. Biosci. Sch.* 8:106–114. doi: 10.2741/S450

Sutherland, D. E. K., and Stillman, M. J. (2011). The "magic numbers" of metallothionein. *Metallomics* 3, 444–463. doi: 10.1039/c0mt00102c

Thimon, V., Koukoui, O., Calvo, E., and Sullivan, R. (2007). Region-specific gene expression profiling along the human epididymis. *Mol. Hum. Reprod.* 13, 691–704. doi: 10.1093/molehr/gam051

Touré, A. (2019). Importance of slc26 transmembrane anion exchangers in sperm post-testicular maturation and fertilization potential. *Front. Cell Dev. Biol.* 7:230. doi: 10.3389/fcell.2019.00230

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Tissue-based map of the human proteome. *Science* 347:1260419. doi: 10.1126/science.1260419

Visconti, P. E., Krapf, D., De La Vega-Beltrán, J. L., Acevedo, J. J., and Darszon, A. (2011). Ion channels, phosphorylation and mammalian sperm capacitation. *Asian J. Androl.* 13, 395–405. doi: 10.1038/aja.2010.69

Wagner, C. A., Finberg, K. E., Breton, S., Marshansky, V., Brown, D., and Geibel, J. P. (2004). Renal vacuolar H+-ATPase. *Physiol. Rev.* 84, 1263–1314. doi: 10.1152/physrev.00045.2003

Wong, D., Merrifield-MacRae, M., and Stillman, M. (2017). "9. Lead(II) Binding in Metallothioneins," in *Lead: Its Effects on Environment and Health*, eds A. Sigel, H. Sigel, and R. Sigel (Berlin: De Gruyter), 241–270. doi: 10.1515/9783110434330-009

Xu, Y., Xiao, F. -L., Xu, N., and Qian, S. -Z. (1985). Effect of intra-epididymal injection of copper particles on fertility, spermatogenesis, and tissue copper levels in rats. *Int. J. Androl.* 8, 168–174. doi: 10.1111/j.1365-2605.1985.tb00830.x

Yen, P. S., Chen, C. H., Sreenu, V., Kohl, A., and Failloux, A. B. (2019). Assessing the potential interactions between cellular mirna and arboviral genomic RNA in the yellow fever mosquito, aedes aegypti. *Viruses* 11:540. doi: 10.3390/v11060540

Yuan, Z., Luo, J., Wang, L., Li, F., Li, W., and Yue, X. (2020). Expression of DAZL gene in selected tissues and association of its polymorphisms with testicular size in Hu sheep. *Animals* 10, 1–12. doi: 10.3390/ani10040740

Zhao, W., Mengal, K., Yuan, M., Quansah, E., Li, P., Wu, S., et al. (2019). Comparative RNA-Seq analysis of differentially expressed genes in the epididymides of Yak and cattleyak. *Curr. Genom.* 20, 293–305. doi: 10.2174/1389202920666190809092819

Zhu, Z., Li, C., Yang, S., Tian, R., Wang, J., Yuan, Q., et al. (2016). Dynamics of the transcriptome during human spermatogenesis: predicting the potential key genes regulating male gametes generation. *Sci. Rep.* 6, 1–15. doi: 10.1038/srep19069

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.

Copyright © 2021 de Lima, Afonso, Edson, Marcellin, Palfreyman, Porto-Neto, Reverter and Fortes. 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: 10 May 2021
doi: 10.3389/fgene.2021.665153

Comparative Analysis of the Circular Transcriptome in Muscle, Liver, and Testis in Three Livestock Species

Annie Robic[1], Chloé Cerutti[1], Christa Kühn[2,3]* and Thomas Faraut[1]*

[1] INRAE, ENVT, GenPhySE, Université de Toulouse, Castanet-Tolosan, France, [2] Institute Genome Biology, Leibniz Institute for Farm Animal Biology (FBN), Dummerstorf, Germany, [3] Faculty of Agricultural and Environmental Sciences, University of Rostock, Rostock, Germany

Circular RNAs have been observed in a large number of species and tissues and are now recognized as a clear component of the transcriptome. Our study takes advantage of functional datasets produced within the FAANG consortium to investigate the pervasiveness of circular RNA transcription in farm animals. We describe here the circular transcriptional landscape in pig, sheep and bovine testicular, muscular and liver tissues using total 66 RNA-seq datasets. After an exhaustive detection of circular RNAs, we propose an annotation of exonic, intronic and sub-exonic circRNAs and comparative analyses of circRNA content to evaluate the variability between individuals, tissues and species. Despite technical bias due to the various origins of the datasets, we were able to characterize some features (i) (ruminant) liver contains more exonic circRNAs than muscle (ii) in testis, the number of exonic circRNAs seems associated with the sexual maturity of the animal. (iii) a particular class of circRNAs, sub-exonic circRNAs, are produced by a large variety of multi-exonic genes (protein-coding genes, long non-coding RNAs and pseudogenes) and mono-exonic genes (protein-coding genes from mitochondrial genome and small non-coding genes). Moreover, for multi-exonic genes there seems to be a relationship between the sub-exonic circRNAs transcription level and the linear transcription level. Finally, sub-exonic circRNAs produced by mono-exonic genes (mitochondrial protein-coding genes, ribozyme, and sno) exhibit a particular behavior. Caution has to be taken regarding the interpretation of the unannotated circRNA proportion in a given tissue/species: clusters of circRNAs without annotation were characterized in genomic regions with annotation and/or assembly problems of the respective animal genomes. This study highlights the importance of improving genome annotation to better consider candidate circRNAs and to better understand the circular transcriptome. Furthermore, it emphasizes the need for considering the relative "weight" of circRNAs/parent genes for comparative analyses of several circular transcriptomes. Although there are points of agreement in the circular transcriptome of the same tissue in two species, it will be not possible to do without the characterization of it in both species.

Keywords: circular RNA, annotation, sub-exonic circRNA, intronic circRNAs, parent genes, circular transcriptome, exonic circRNA

OPEN ACCESS

Edited by:
James Reecy,
Iowa State University, United States

Reviewed by:
Paolo Zambonelli,
University of Bologna, Italy
Jacqueline Smith,
The University of Edinburgh,
United Kingdom

****Correspondence:***
Annie Robic
annie.robic@inrae.fr
Christa Kühn
kuehn@fbn-dummerstorf.de

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

Received: *07 February 2021*
Accepted: *07 April 2021*
Published: *10 May 2021*

Citation:
Robic A, Cerutti C, Kühn C and Faraut T (2021) Comparative Analysis of the Circular Transcriptome in Muscle, Liver, and Testis in Three Livestock Species. Front. Genet. 12:665153. doi: 10.3389/fgene.2021.665153

INTRODUCTION

The identification and functional characterization of all transcripts in livestock species is one of the major goals of the consortium for the Functional Annotation of Animal Genomes (FAANG[1]). An animal genome contains 20,000 to 30,000 genes but only a subset of these genes produce transcripts in a given tissue. Developments in high-throughput RNA-seq technology have enabled a deeper understanding of gene expression functions. The classical approach for studying the transcriptome uses the mRNA-seq protocol (sequencing of polyadenylated RNAs). A large number of mRNA-seq studies have demonstrated for example that part of the associated genes are transcribed in a tissue specific manner (Soumillon et al., 2013). However, datasets generated by mRNA-seq contain only a part of the transcripts. To overcome this drawback, it is possible to sequence RNAs after depletion of ribosomal sequences (total-RNA-seq) (Chen et al., 2020). Studies using the total-RNA-seq protocol have shown that a large number of protein-coding genes, long non-coding (lnc) RNA genes and intergenic elements are expressed in a tissue-specific manner (Soumillon et al., 2013; Clark et al., 2017; Yang et al., 2020).

Since 2012, advances in high throughput sequencing revealed the presence of circular RNAs (circRNAs) in total-RNA-seq datasets in addition to linear transcripts (Salzman et al., 2012). CircRNAs are probably a natural by-product of the transcription process in all eukaryotes (see Kristensen et al. (2019) for a review). To study circRNAs, it is important to identify the gene that is likely to generate the considered circRNA alongside the linear transcripts already described, namely the parent gene. The majority of studies have focused on exonic circRNA that are generated by the circularization of one or several exons through a back splicing process: the end of an exon is joined to the beginning of an upstream exon (Kristensen et al., 2019; Li et al., 2019). Exonic circRNAs can be produced by coding or non-coding genes (Salzman et al., 2012; Barrett et al., 2017; Robic et al., 2020) and by some snRNA genes (Kaur et al., 2018; Robic et al., 2020). Two types of circRNAs can be derived from intronic sequences (see Robic and Kühn (2020) for a review): (1) when intron lariats escape degradation due to the failure of intron debranching (Zhang et al., 2013), they may become circRNA precursors of lariat-derived circRNAs and (2) for very rare introns we can observe the circularization of the entire intronic sequences as intron circle (Taggart et al., 2017). These two types of intron derived circRNAs can be grouped as "intronic circRNAs." Sub-exonic circRNAs have been characterized as including only a part of the exon of some mono-exonic genes (Robic et al., 2020). Up to now, only intronic circRNA from protein-coding genes and only sub-exonic circRNAs from small-non-coding (snc) RNA have been reported in pigs (Robic et al., 2020). Understanding to what extent, these different subclasses of circRNAs are produced and what kind of genes are able to produce them is a question of interest.

In 2013, circRNAs were shown to have functional relevance (Jeck et al., 2013; Memczak et al., 2013) as reviewed by (Chen, 2020; Xiao et al., 2020). Since the landmark discovery of ciRS-7/CDR1as functioning as a miR-7 sponge in Hansen et al. (2013), a lot of studies focused on circRNAs action as microRNA decoys. However, as circRNA research expands, many divergent views have emerged (reviewed by Li et al., 2019) and our understanding of circRNAs, their production and their function, remains limited. The diversity of functions assigned to circRNAs is very large but concerns only some circRNAs. For example, recent studies have highlighted that the presence of a particular circRNA from *SLC45A4* is essential to keep neural cells in a progenitor state in the mammalian brain (Suenkel et al., 2020). Recently, the regulatory functions of two circRNAs produced respectively from a mitochondrial gene (Zhao et al., 2020) and from an intron (Das et al., 2020; Stoll et al., 2020) were characterized. These studies have also underlined the need to work on the conservation of circRNAs, and beyond exonic circRNAs.

The identification of circular RNAs in highly divergent species raises interesting questions about their evolutionary history, and functions (Wang et al., 2014; Ji et al., 2019; Zucko and Boris-Lawrie, 2020). Li et al. (2019), who reviewed this topic, reported wide discrepancies: some studies claimed that circRNAs are evolutionarily highly conserved molecules, while others believe they are species-specific. For our study, we take advantage of functional datasets produced within partners and also the FAANG consortium to study circular RNA in farm animals (cattle, pig, sheep). We studied the pervasiveness of circRNA transcription in three tissues (skeletal muscle, liver, and testis). As the transcriptomes of these three tissues present specific features (Yang et al., 2020), this choice seemed to us pertinent to compare and analyze circRNA production.

MATERIALS AND METHODS

Data Collection

For this study, we collected total-RNA-seq data produced by our groups and others from the literature. The whole dataset contains 33 bovine tissues, 15 ovine tissues and 19 pig tissues (**Table 1**). The considered datasets are originating from three SRA projects for bovine tissue and from four SRA projects for porcine tissues. All the ovine datasets were generated at Roslin Institute in a unique SRA biological project PRJEB19199 (Clark et al., 2017). We achieved to consider at least 70 giga bases (gb) of sequencing data for each tissue in a given species. In the following, a batch is defined as a collection of datasets from a single tissue of animals from the same species, same sex, same age, and originating from a unique SRA project. In **Table 1**, these datasets are clustered in 17 batches and one singleton. The dataset ssc_testis_1 was excluded from the batch, which constituted datasets from the SRA project PRJNA506525, because of its atypical behavior regarding the production of circRNAs (Robic et al., 2019). Among the 48 animals from the SRA project PRJEB34570 (Nolte et al., 2019), we chose six males and six females to obtain two batches balanced on known physiological traits. For bovine testis (Gao et al., 2019), we selected datasets from bulls at 13 months of age (bta_testis_4-6), which is assumed

[1]www.animalgenome.org/community/FAANG/

TABLE 1 | Samples characteristics.

Datasets	Species	Tissue	Age animal	Sex animal	Breed		Reads PE (bp)	SRA project	gb
bta_liver_1-6 [1]	cattle	liver	18 months	male	Charolais X Holstein-F2	FBN (Nolte et al., 2019)	2 × 100	PRJEB34570	68.1
bta_liver_7-12 [1]	cattle	liver	3.5 years	female	Charolais X Holstein-F2	FBN (Nolte et al., 2019)	2 × 100	PRJEB34570	68.3
bta_liver_13-15	cattle	liver	"adult"	male		EMBL-2017	2 × 10	PRJEB13074	26.4
bta_muscle_1-6 [1]	cattle	muscle	18 months	male	Charolais X Holstein-F2	FBN (Nolte et al., 2019)	2 × 100	PRJEB34570	55.1
bta_muscle_7-12 [1]	cattle	muscle	3.5 years	female	Charolais X Holstein-F2	FBN (Nolte et al., 2019)	2 × 100	PRJEB34570	56.5
bta_testis_1-3	cattle	testis	2 days	male	Angus	Yangling (Gao et al., 2019)	2 × 150	PRJNA47564	46.5
bta_testis_4-6	cattle	testis	13 months	male	Angus	Yangling (Gao et al., 2019)	2 × 150	PRJNA47564	43.9
ssc_liver_5-7 [2]	pig	liver	2 years	male		EMBL-2019	2 × 150	PRJEB33381	46.0
ssc_liver_8-10	pig	liver	"adult"	male		EMBL-2017	2 × 100	PRJEB13074	26.6
ssc_muscle_2-4 [2]	pig	muscle	2 years	male		EMBL-2019	2 × 150	PRJEB33381	87.8
ssc_testis_1 [3][4]	pig	testis	6 months	male	Pietrain (Pi)	INRAE (Robic et al., 2019)	2 × 125	PRJNA506525	17.3
ssc_testis_2-7 [4]	pig	testis	6 months	male	3 Pi & 3Pi X Large White	INRAE (Robic et al., 2019)	2 × 125	PRJNA506525	115.2
ssc_testis_8-10 [2]	pig	testis	2 years	male		EMBL-2019	2 × 150	PRJEB33381	56.6
oar_liver_1-3 [1]	sheep	liver	2 years	male	Scot. Blackface x Texel	Roslin (Clark et al., 2017)	2 × 125	PRJEB19199	90.8
oar_liver_4-6 [1]	sheep	liver	2 years	female	Scot. Blackface x Texel	Roslin (Clark et al., 2017)	2 × 125	PRJEB19199	90.8
oar_muscle_1-3 [1]	sheep	muscle	2 years	male	Scot. Blackface x Texel	Roslin (Clark et al., 2017)	2 × 125	PRJEB19199	99.4
oar_muscle_4-6 [1]	sheep	muscle	2 years	female	Scot. Blackface x Texel	Roslin (Clark et al., 2017)	2 × 125	PRJEB19199	91.8
oar_testis_1-3 [1]	sheep	testis	2 years	male	Scot. Blackface x Texel	Roslin (Clark et al., 2017)	2 × 125	PRJEB19199	92.9

Seven teen groups of datasets were collected, combining total RNAseq generated by our groups and others from the literature. Only groups of datasets generated by sequencing stranded RNA from healthy animals and containing at least three datasets produced in parallel were selected. [1]The animals from [bta_liver_1-6 and bta_muscle_1-6], [bta_liver_7-12 and bta_muscle_7-12], [oar_liver_1-6 and oar_muscle_1-6], and [oar_liver_4-6, oar_muscle_4-6 and oar_testis_1-3] were the same and were presented in the same order. [2]The three animals from [ssc_liver_5-7, ssc_muscle_2-4 and ssc_testis_8-10] were the same but they were not presented in the same order. [3]We chose to not associate the dataset ssc_testis_1 to other datasets from the SRA project PRJNA506525 because we know that this datasets is an outlier dataset in its origin group. [4]Datasets ssc_testis_1-7 were obtained from corresponding animals previously called Animal−31, −05, −54, −16, −65, −73, and −93 in Robic et al. (2019).

an age at the end of puberty (Rawlings et al., 2008; McGowan et al., 2018), to represent data from pubertal testis.

To analyze the impact of read length on the circular RNA detection, we produced 6 new datasets with 2 × 100 bp PE (Paired- End) sequencing from the 6 datasets from cattle testis samples, which had been previously sequenced for 2 × 150 bp PE.

Reads Mapping

The RNA-seq reads were aligned to the following genome reference assemblies: ARS-UCD1.2 (GCA_002263795.2), Oar_rambouillet_v1.0 (GCA_002742125.1), Sscrofa11.1 (GCA_000003025.6) for cow, sheep and pig respectively. We used the gene annotation (v-101) provided by Ensembl (Ensembl-Websites: Cattle, 2021; Pig, 2021; Sheep, 2021).

RNA-seq reads were mapped to the genome reference assemblies using the rapid splice-aware read mapper Spliced Transcripts Alignment (STAR) (Dobin et al., 2013). Two alignment modes were considered, single-end alignments (STAR-SE option, mates of each pair were mapped independently) and paired-end alignments (STAR-PE option). STAR was used with the previously proposed parameters (Cheng et al., 2016) that enable the highlighting of chimeric reads with only two segments and with a minimal size for the smallest mapped segment of 15 bp.

CircRNA Detection and Annotation

The first step in the detection of circRNAs is the identification of reads containing a circular junction (see Gao and Zhao, 2018 for a review). The analysis of these reads allows to describe each circRNA by the two points involved in the circular junction (the genomic boundaries of the circularized transcript: two genomic coordinates) and the strand. The second step of the characterization of circRNAs is their annotation.

Standard CircRNA Detection and Quantification

The first approach for detecting circRNAs used the combination of circRNAs detected by CIRCexplorer2 (CE2, Zhang et al., 2016) and CIRI2 (Gao et al., 2015, 2018) which have become reference tools for the identification of exonic circRNAs (Zeng et al., 2017; Gao and Zhao, 2018; Hansen, 2018; Dodbele et al., 2021). CE2 is able to use several aligners and our choice was to use CE2 associated to STAR-PE (Dobin et al., 2013) alignment mode (Zhang et al., 2016). It is important to note that we have chosen more stringent parameters for the alignment performed with STAR-PE than those suggested by Zhang et al. (2016) (see above) for the detection of chimeric reads. As CE2 identifies reads containing a circular junction within those reads that STAR calls "chimeric reads" (CR), we will call these reads "circular chimeric reads" (hereafter CCRs). CIRI2 (Gao et al., 2018) is based on the bwa-mem aligner (Au et al., 2016) together with a dedicated approach to align unmapped segments. CIRI2 was used to identify the reads containing circular junctions with default parameters. Reads containing a circular junction are called BSJ ("back-spliced junction") reads by CIRI2.

All circular RNAs identified by CE2 as generated from backsplicing of two described exons were considered as exonic circRNAs. Those annotated as "ciRNA" correspond to circRNAs localized entirely in intronic sequences and with the circRNA 5′ junction site corresponding to the intron donor site. Although the term "ciRNA" is the one proposed for intronic circular RNA by Zhang et al. (2013), the location of the 3′ junction of these circRNAs must be analyzed before to consider them as intronic circRNA (Robic and Kühn, 2020).

Only circRNAs detected by both CE2 and CIRI2 were considered for quantification as suggested previously by Hansen (2018). A CIRI2 formatted list of circRNAs was provided to CIRIquant (Zhang et al., 2020) to obtain an accurate quantification of circRNAs. The quantitative measure is the number of BSJ reads provided by CIRIquant. The expression measure for each parent gene is simply the sum of expression measure of the circRNAs it produces. In order to obtain, for each circRNA of each parent gene, an average expression for the tissue, the average expression over all animals was computed after normalization by animal (TMM normalization provided by edgeR, Robinson et al., 2010).

Detection of Orthologous circRNAs

Orthologous circular RNAs were identified based on nucleotide sequence alignments. Each circRNA is represented by the nucleotide sequence crossing the circular junction point (200bp, 100 bp on each side). Given the close evolutionary between bovine and ovine we have limited the detection of orthologous circRNAs to those two species. All circRNA bovine sequences were aligned to ovine circRNA sequences and reciprocal best hits were considered as orthologous circRNAs. For the parent genes, the orthologs were defined as the Ensembl one-to-one orthologs.

Computational Approach for Exhaustive Detection and Characterization of circRNAs Complementary to Standard Tools

In this manuscript, we use CD as an abbreviation for our dedicated approach to detect circRNAs. Our approach to identify reads containing a circular junction is based on split alignment as defined by Gao et al. (2018) and was originally proposed by Memczak et al. (2013). This approach has previously been described (Robic et al., 2020), and we underline only some essential features. The focus of this alternative framework method is limited to selecting reads that are mapped by STAR-SE as CR with only two segments, and where both segments are mapped to the same strand in inverted order. To select CCRs, we extracted information from the tabular file (chimeric.out.junction) provided by STAR, which contains the mapping coordinates of each segment and mapping data (CIGAR). An output file (BED format) containing a list of circRNAs is obtained by clustering of CCR on genomic coordinates. The second part of our approach consists in proposing an annotation for the circRNAs detected. The annotation was performed using the species corresponding gene annotation from Ensembl and in particular, the list of exons and the corresponding list of introns.

For the annotation purpose, we define the following classes: (i) exonic, when both junctions correspond exactly to exon boundaries, where both exons belong to the same gene. (ii) sub-exonic, when both junctions fall strictly within a single exon.

(iii) intronic, when both junctions fall within a single intron, the 5′ junction corresponding to the intron donor site and the 3′ junction located not further than 60 bp away from the intron acceptor site (Robic and Kühn, 2020). CircRNA from ribosomal RNA genes were excluded from the list of sub-exonic circRNAs. All circRNAs with a too small genomic size (when the genomic size < (1/2 length of read + 5 nt)) were excluded from the annotation process.

RESULTS

Circular RNAs: Detection and Annotation

Initial circRNAs Landscape Established by CIRI2 and CIRCexplorer2

The detection pipelines detected on average 8,300 and 16,300 circRNAs per sample for CE2 and CIRI2 respectively. A first analysis on six datasets showed that a very large part of exonic circRNA (>90%) detected by CE2 was also detected by CIRI2, while the fraction of ciRNAs (putative intronic circRNAs, see Materials and Methods) proposed by CE2 and detected also by CIRI2 was less than 2%. These observations underline the fact that the detection of non-exonic circRNAs remains a difficult task and at least subject of debate. As it is common practice (Gong et al., 2020), initially recommended by Hansen (2018), we considered only the circRNAs detected by both tools (Dodbele et al., 2021). Moreover, we retained only circRNAs characterized by at least four reads containing the circular junction, and this threshold was applied after the intersection of CE2 and CIRI2 data (BSJs > = 4). On average and in each of the 66 datasets, 1,957 circRNAs were characterized by CE2 + CIRI2.

Using this strategy, we were able to characterize 12,589 exonic circRNAs and 6 ciRNA in the bovine datasets (**Supplementary File 1**). For pigs, the statistics were 14,137 and 1, for exonic circRNAs and ciRNAs respectively. For sheep, we found 5,556 exonic circRNAs and 3 ciRNAs. A large variability between datasets in this raw number of circRNAs detected was noted (see **Supplementary File 2**). Before all further analyses, the number of circRNAs identified in each of the 66 datasets was corrected by the number of uniquely mapped reads by STAR (**Supplementary File 2**). We compared this normalized number of circRNA in each of the 66 datasets (**Figures 1A-F**). Since some datasets differ by read length, in order to analyze the impact of read length on the circular detection, we produced six new datasets of 2 × 100 bp PE sequencing from the six bta_testis_1-6 datasets, which had been previously sequenced for 2 × 150 bp PE. The detection of exonic circRNAs was performed by CE2, and we observed a 10 to 20% loss of initial exonic circRNAs with shortened reads (data not shown). This experiment shows that even if the length of the reads has an impact on the detection of exonic circRNAs, this impact is moderate. Therefore, the difference of reads lengths from PE sequencing do not explain the large differences observed between circRNA content of the two batches generated from porcine liver at EMBL in two different SRA projects (ssc_liver_8-10 (**Figure 1C**) and ssc_liver_5-7 (**Figure 1D**)). In bovine liver, the number of circRNAs also appeared variable between SRA projects. We observed 13.94 to 17.98 circRNAs per million uniquely mapped reads (per million reads for short) for the 12 first datasets (bta_liver_1-12), and 3.96 to 8.29 for the three others (bta_liver_13-15) (**Figures 1A,C**), although all samples were sequenced in PE mode with 2 × 100 bp. For bta_liver_1_12, the circRNAs per million reads did not differ much between samples, although the dataset included physiologically very divergent animals, i.e., bulls at the end of fattening and cows at the beginning of lactation. Also in porcine testis, the number of circRNAs seemed very different in datasets produced at EMBL (ssc_testis_8-10) to those produced at INRAE (ssc_testis_2-7). However, in this comparison, the age of the considered animals was different: datasets ssc_testis_8-10 were obtained from adult boars (two years old), while ssc_testis_2-7 originated from pubertal animals (<6 months old).

The number of circRNAs detected in testis of very young bulls seemed higher than in testis from pubertal animals (**Figure 1F**). As these two datasets were included in the same SRA project and absence of technical bias could be assumed, a statistical analysis was performed (**Supplementary File 3**), which revealed that the difference in the number of circRNAs was significant (p-Value = 0.016). The number of circRNA in testis of pubertal animals appeared similar in bovine and in pigs, and also the number of circRNAs in testis of adult animals displayed a similar level in pigs and sheep (**Figure 1G**), although these datasets were not generated by the same sequencing lab. This analysis underlines the importance to consider the age (or sexual maturity) of animals for testicular datasets. Since the difference between males and females was not statistically significant for bovine liver and muscle from the same animals (bta_liver_1-6 and _7-12; bta_muscle_1-6 and 7-12), we will no longer differentiate between male and female datasets of this species-tissue combination.

The number of circular junction reads associated with the detection of a circRNA is commonly used to quantify the expression of this circRNA. We chose to perform this quantification by CIRIquant (Zhang et al., 2020), and each circRNA characterized by the CE2 + CIRI2 approach was associated with an expression level measured by the number of BSJs. We considered the sum of the BSJs (corrected by the number of reads uniquely mapped in the dataset) across datasets grouped in the same batch as the expression of this circRNA in the considered batch. We did a ranking of circRNA expressions within each batch; this should enable comparisons of ranking between batches. When we performed a pairwise comparison of the Top-100 most highly expressed circRNAs (**Supplementary Table 1**) between batches, we found different degrees of overlaps between pairs (**Figure 2A** in blue). All comparisons were performed within species, even though we also looked at differences/similarities of the statistics between species. Before comparing batches, we compared six pairs of two randomly selected datasets from the batch bta_liver_1-12, and on average 71% of overlaps were observed (six comparisons performed: 63 to 79%). Similar levels of overlap were noted when comparing the two different batches from bovine liver (72%) and between the two batches from porcine liver (72%). These scores showed that the identification of the most highly expressed circRNAs (at least in liver) is not very sensitive to the

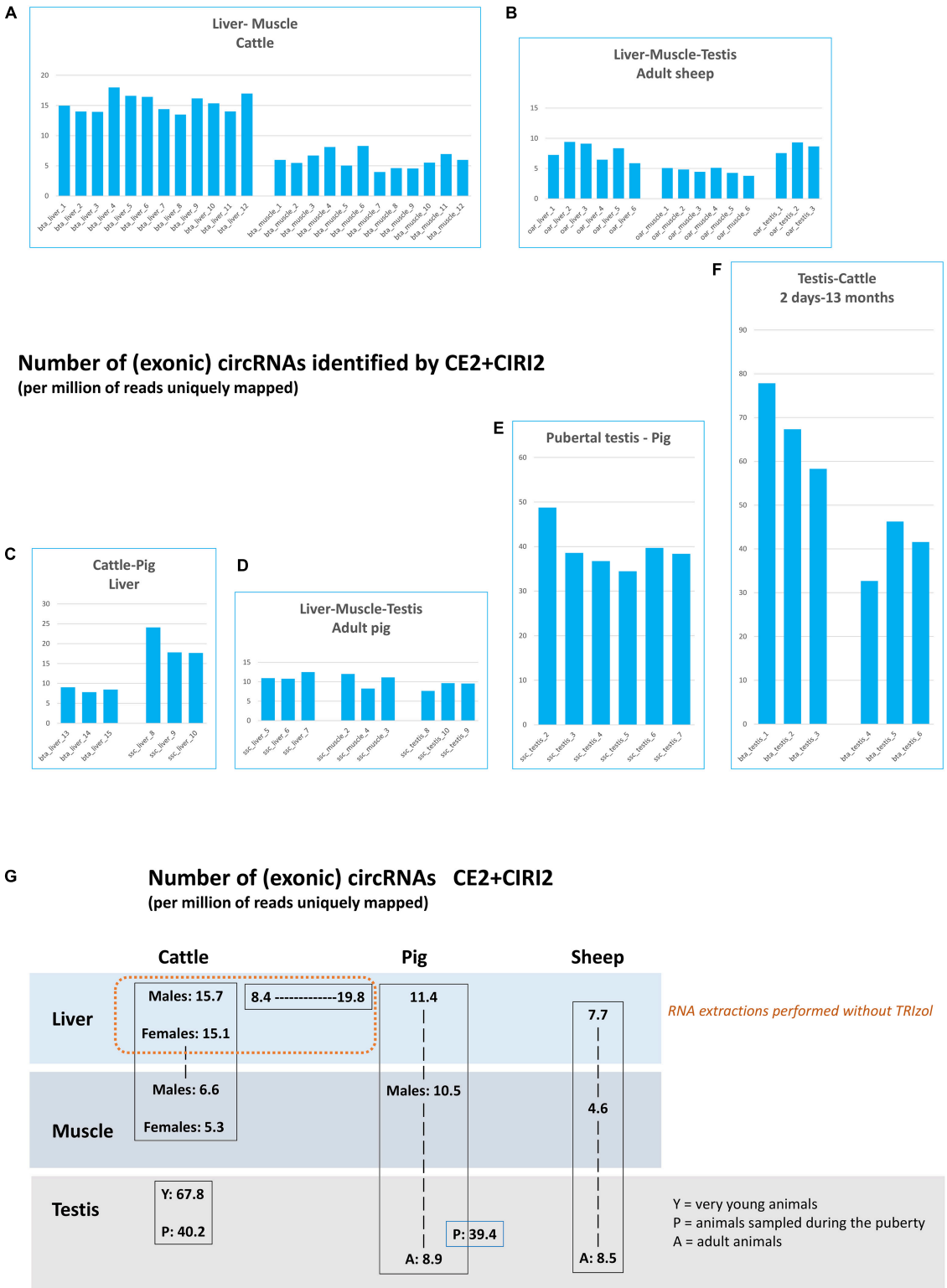
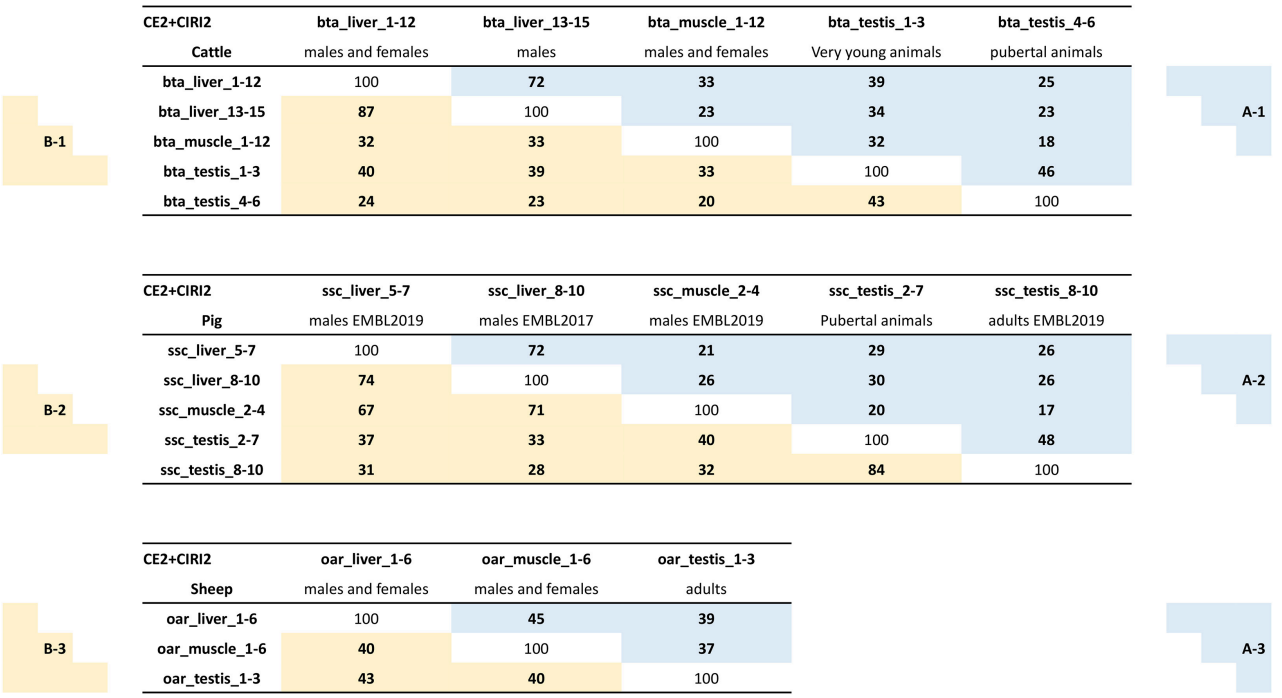


FIGURE 1 | Number of circRNA characterized by CE2 + CIRI2. **(A–G)** These histograms represent the number of circRNA (per million of reads uniquely mapped) characterized jointly by CE2 and CIRI2 and which are detected by at least 4 BSJs (CIRIquant). Histograms are regrouped by SRA projects. **(A)** The two bovine batches produced at FBN. **(B)** the three ovine batches produced at Roslin Institute **(C)** The two batches produced by EMBL in 2017. **(D)** The two batches produced by EMBL in 2019. **(E)** The batch of 6 datasets produced from porcine pubertal testes at INRAE. **(F)** The two batches of bovine testes produced by Yangling University. **(G)** Comparison per tissue and species of the number of exonic circRNAs detected by CE2 + CIRI2 per million of reads uniquely mapped.



CE2+CIRI2 Cattle	bta_liver_1-12 males and females	bta_liver_13-15 males	bta_muscle_1-12 males and females	bta_testis_1-3 Very young animals	bta_testis_4-6 pubertal animals
bta_liver_1-12	100	72	33	39	25
bta_liver_13-15	87	100	23	34	23
bta_muscle_1-12	32	33	100	32	18
bta_testis_1-3	40	39	33	100	46
bta_testis_4-6	24	23	20	43	100

CE2+CIRI2 Pig	ssc_liver_5-7 males EMBL2019	ssc_liver_8-10 males EMBL2017	ssc_muscle_2-4 males EMBL2019	ssc_testis_2-7 Pubertal animals	ssc_testis_8-10 adults EMBL2019
ssc_liver_5-7	100	72	21	29	26
ssc_liver_8-10	74	100	26	30	26
ssc_muscle_2-4	67	71	100	20	17
ssc_testis_2-7	37	33	40	100	48
ssc_testis_8-10	31	28	32	84	100

CE2+CIRI2 Sheep	oar_liver_1-6 males and females	oar_muscle_1-6 males and females	oar_testis_1-3 adults
oar_liver_1-6	100	45	39
oar_muscle_1-6	40	100	37
oar_testis_1-3	43	40	100

FIGURE 2 | Comparisons circRNAs and parent genes between datasets. The three diagrams depict analyses of exonic circRNAs characterized by CE2 + CIRI2. The number of similarities (same circRNA or same parent gene) found for a comparison between two datasets was reported in a box. **A**-(boxes with a blue background): The expression of a circRNA in a batch has been defined as the sum of the BSJs (normalized counts) observed in the different datasets of this batch. The circRNAs were ranked on their expression to establish the Top-100 of circRNAs expressed in this batch (Lists of Top-100/circRNAs relative to these analyses were reported in **Supplementary Table 1**). **B**-(boxes with a yellow background): The circRNA expression of genes in a batch has been defined as the sum of the BSJs (normalized counts) from each circRNA produced by this gene observed in the different datasets of this batch. The parent genes were ranked on their circRNA expression to establish the Top-100 of parent genes expressed in this batch (Lists of Top-100/genes relative to these analyses were reported in **Supplementary Table 2**).

source of data analyzed for circRNAs characterization. Between muscle and liver, similar levels of overlap were noted in bovine (23-33%) and in pigs (21-26%) (**Figures 2A-1,2**). In testis, we noted a similar level of overlap between testis from pubertal animals and muscle from other animals (cattle:18% and pigs: 20%). However, the level of overlap between testis and adult muscle seemed to decrease with the age of testis, because we observed a 32% overlap for testis from young animals (cattle), and 17% for testis from adult pigs. These analyses demonstrate the differences in the circRNA expression in testes in relation to the age of the animals. Curiously, the levels of overlap between testis/liver/muscle appeared higher in sheep (**Figure 2A-3**) than in bovine (**Figure 2A-1**) or pigs (**Figure 2A-2**), probably because of differences in genome annotation. These analyses underline once again the importance to consider the age of animals for testicular datasets, but attenuates the importance of the source of datasets with respect to most highly expressed circRNAs. However, it has to be considered that these analyses are restricted almost exclusively to exonic circRNA.

When we examined the circRNAs detected jointly by CE2 and CIRI2 and retained in our analysis (**Supplementary List 1**), we noted that almost all are exonic circRNAs. As our purpose is to study all types of circRNAs in three tissues of three species, we included a dataset (ssc_testis_1) with a particular circRNA content already explored in previous studies (Robic et al., 2019, 2020). This porcine testicular dataset is known to contain more than 100 intron-derived circRNAs. The major intronic circRNA described in this dataset was a lariat-derived circRNA from the *ATX2NL* gene. CE2 was able to detect circRNAs from the respective *ATX2NL* intron, but with six times less CCRs than previously observed (Robic et al., 2020). CIRI2 did not detect this intronic circRNA, probably because of its small size. CE2 was also able to detect the six intronic circRNAs from the *DNAH17* previously reported (Robic et al., 2019) but again with a lower number of CCRs than expected. These observations confirmed that CE2 is able to detect intronic circRNAs (Das et al., 2020), but as suggested previously (Robic and Kühn, 2020), the strong requests on the two paired-end reads as included in the CE2 algorithm could impair the characterization of intronic circRNAs. The dataset ssc_testis_1 had also been used to describe the first sub-exonic circRNAs (Robic et al., 2020). These sub-exonic circRNAs were characterized by the observation of reads containing a circular junction and spanning parts of the single exon from mono-exonic genes classified as small non-coding RNA. The genes involved were able to produce several, potentially overlapping circRNAs from a single exon. The production of a set of circRNAs by the mono-exonic gene *RMRP* (orthologous gene of porcine/bovine/sheep *RNase_MRP*)

was already highlighted in two species (Liu et al., 2020; Robic et al., 2020). This gene was reported as able to produce several dozens of sets of sub-exonic circRNAs in the dataset ssc_testis_1, but only two sub-exonic circRNAs were found in the list of circRNAs provided by CIRI2. From these data, we could conclude that the lists of circRNAs obtained by the conservative approach of CE2 + CIRI2 output fell short with respect to an exhaustive circRNA detection in the three tissues under investigation.

Exhaustive Detection and Annotation of circRNAs

As our purpose was to study all types of circRNAs, we used an alternative approach (CD) for the exhaustive detection of circRNA (see section "Materials and Methods"). The next objective was to further annotate the detected circRNAs as either exonic, intronic and sub-exonic, and the remaining as undefined or unannotated. Our objective was not to provide an alternative list of exonic circRNAs to the one established by CE2 + CIRI2, but only to identify a maximum number of exonic circRNA as a prerequisite for an improved, subsequent analysis of the other circRNAs. The criteria to annotate exonic circRNAs were identical to those used by CE2, but for intronic circRNAs we were more stringent (see section "Materials and Methods"). As Liu et al. (2020) suggested that the production of sub-exonic circRNAs was not limited to exons from mono-exonic and non-coding RNA-genes, we integrated in our alternative approach also the detection of sub-exonic circRNAs from all exons: from coding and non-coding genes, from mono-exonic and multi-exonic genes. To avoid including false positives in our analysis, we disregarded very rare circularization events: circRNAs were retained when they were characterized by at least 5 CCRs. This choice was motivated by previous studies using a similar approach for the detection of circRNAs (Robic et al., 2019, 2020). On average and in each of the 66 datasets, CD detected 65,500 circRNAs, and after the application of this threshold, 2,644 circRNAs were retained.

Results of the exhaustive detection of circRNA were shown on **Figure 3**. The number of putative circRNAs detected by CD appeared higher than circRNA detected by CE2 + CIRI2. This difference was particularly marked on the datasets from bovine and porcine muscle (**Figures 1A,D**, **3A,D**). The next step was the characterization of exonic circRNAs and they were indicated in blue on the histograms presented on the **Figure 3** (and were listed in **Supplementary List 2**). For example, 10,351 exonic circRNAs were characterized by CD in 18 porcine datasets (by at least five CCRs in one dataset). Only 50 were never detected by CE2 or CIRI2 (10,358 and 9,940 were detected by CIRI2 and CE2 respectively). The number of exonic circRNAs detected by CD (**Figure 4A**) appeared consistent with the number of circRNAs (mainly exonic circRNAs) jointly detected by CE2 and CIRI2 (**Figure 1G**), even though on average, CE2 + CIRI2 detected more circRNA than exonic circRNAs detected by CD.

Next, we proceeded to the identification, in the 67 (66 + 1) datasets, of intronic circular RNA and sub-exonic circRNAs. We detected only a very low number of introns associated with intronic circRNA in several datasets (for example, intronic circRNA were detected for zero to four introns in bovine muscle datasets) (**Supplementary List 3**), and the dataset ssc_testis_1 turned out again as an outlier (132 introns concerned). In contrast to previous studies, in which only mono-exonic non-coding genes were considered, all sub-exonic circRNA, covered by at least 5 CCRs, for all types of genes (mono- and multi-exonic) were listed. The two ribozyme genes, *RNase_MRP* and *RNaseP_nuc*, are the major small non-coding RNA gene able to produce sub-exonic circRNA. For sub-exonic circRNA assigned to multi-exonic genes, each exon involved produced several, possibly overlapping, sub-exonic circRNAs. We noted that several exons within a particular gene could contribute to the production of sub-exonic circRNAs. For example, we observed sub-exonic circRNAs from the nine exons of *FGB* and from 11 out of the 15 exons of the *ALB* gene in bovine liver (**Supplementary List 4**). Therefore, not only the snc genes contribute to sub-exonic circRNA production but also protein-coding genes, lncRNAs and pseudogenes (see below).

No antisense sub-exonic circRNA was detected in porcine and ovine datasets, but we observed that four misc-RNA and two ribozyme genes produced sense and antisense sub-exonic circRNAs in bovine liver and/or testis. Antisense sub-exonic circRNAs were never seen without the corresponding (from the same exon) sense sub-exonic circRNAs.

Tissue Complexity

Analysis of the Number of circRNAs Characterized

We now turn to the comparison of the number of circRNAs observed in liver, muscle and testis in the three species. The circRNAs detected for each dataset by CD, number and associated annotations, exhibit a much higher homogeneity within batches than between batches, indicating that technical bias (library preparation for example) might drive, in part, the observed difference between tissues and species (**Figure 3**). We noted for example large differences for sub-exonic, intronic and exonic circRNAs number between two batches from porcine liver (ssc_liver_8-10 and ssc_liver_5-7) and two batches from bovine liver (bta_liver_1-12 and bta_liver_13-15) (**Figure 4**). We noted also large differences of patterns between the two batches from porcine testis (ssc_testis_2-7 and ssc_testis_8-10), but in this case the different age of the animals might drive the differences. Once again, in the CD analysis the number of exonic circRNAs in testis of pubertal animals appeared similar in bovine (bta_testis_4-6) and in pigs (ssc_testis_2-7) (**Figures 1G**, **4C**), and in testis of adult animals in ovine (oar_testis_1-3) and in pigs (ssc_testis_8-10). Nevertheless, we noted that these datasets were generated by different sequencing labs.

In order to avoid the potential technical bias mentioned above, we compared batches originating from the same SRA project. A significant difference was detected in the number of exonic circRNA and in the number of CCRs associated with exonic circRNAs between young and adult animal testes ($p = 0.016$ and $p = 0.011$). These results confirmed data already observed with circRNAs detected by CE2 + CIRI2 ($p = 0.016$). In cattle, (2 × 12 datasets produced at FBN) and in sheep (2 × 6 datasets) we were able to show that the liver transcriptome contains more exonic circRNAs than muscle. The three comparisons (number of exonic circRNAs detected by CD, number of CCRs associated

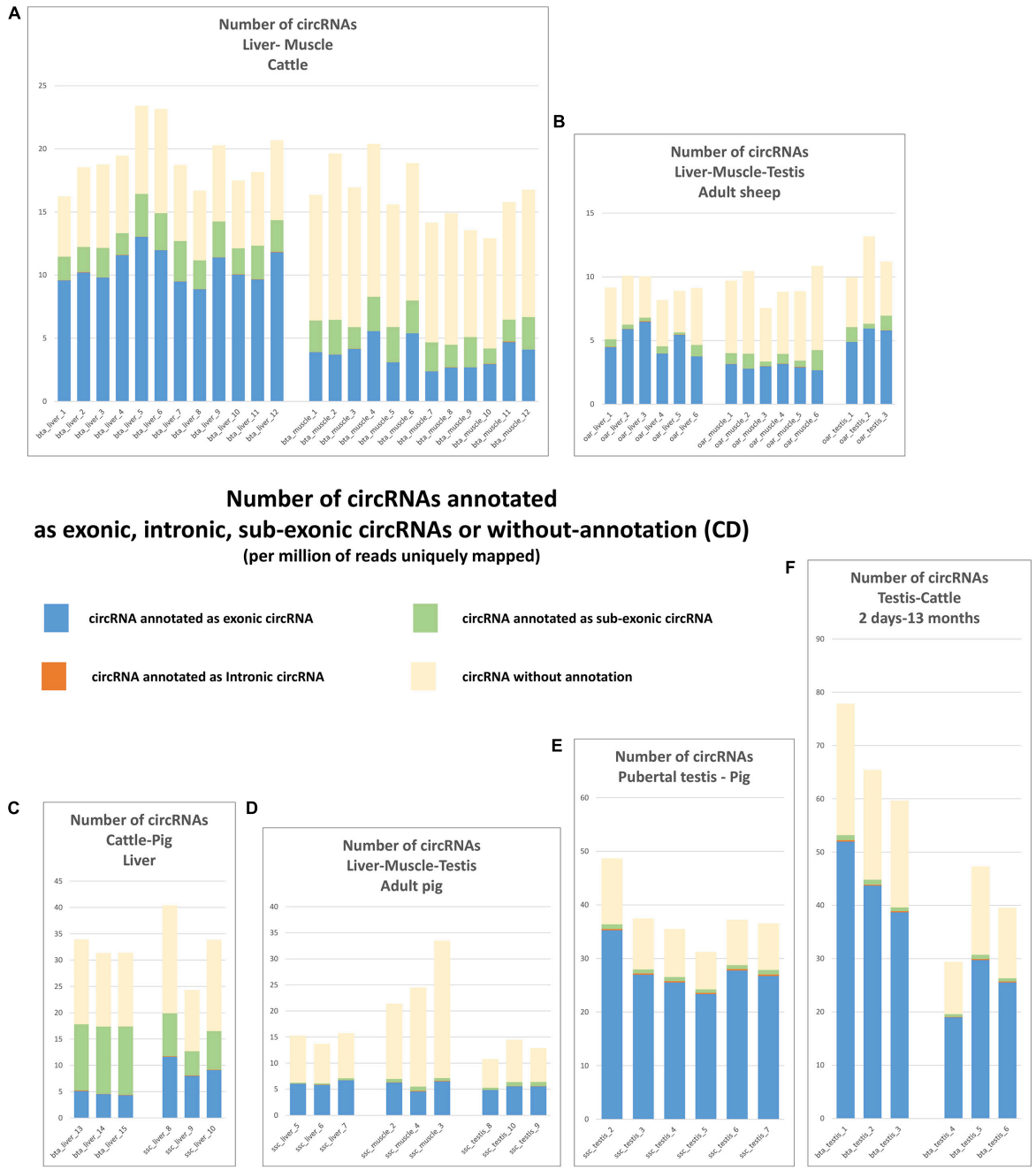


FIGURE 3 | Number of circRNA characterized by CD. These histograms represent the number of circRNA (per million of reads uniquely mapped) which are detected by at least 5 CCRs and annotated as exonic circRNA, intronic circRNAs, sub-exonic or unannotated. Histograms are regrouped by SRA projects. **(A)** The two bovine batches produced at FBN. **(B)** the three ovine batches produced at Roslin Institute **(C)** The two batches produced by EMBL in 2017. **(D)** The two batches produced by EMBL in 2019. **(E)** The batch of 6 datasets produced from porcine pubertal testes at INRAE. **(F)** The two batches of bovine testes produced by Yangling University.

with exonic circRNAs detected by CD and number of exonic circRNAs detected by CE2 + CIRI2) were statistically significant in cattle (p = 3E-9, p = 3E-9, and p = 3E-9, respectively) and in sheep (p = 0.0062, p = 0.019, and p = 0.0031). In contrast, we were not able to confirm this difference in pig. All statistical analyses were reported in **Supplementary File 3**.

Analysis of circRNAs Remaining Without Annotation

Among unannotated circRNAs, a large fraction of very small circRNAs was detected especially in some datasets. For example, in bta_muscle_1-12 we noted 19 to 32% of circRNAs with a genomic size less than 55 bp. It would be necessary to examine the underlying reads to understand this small size, and it is

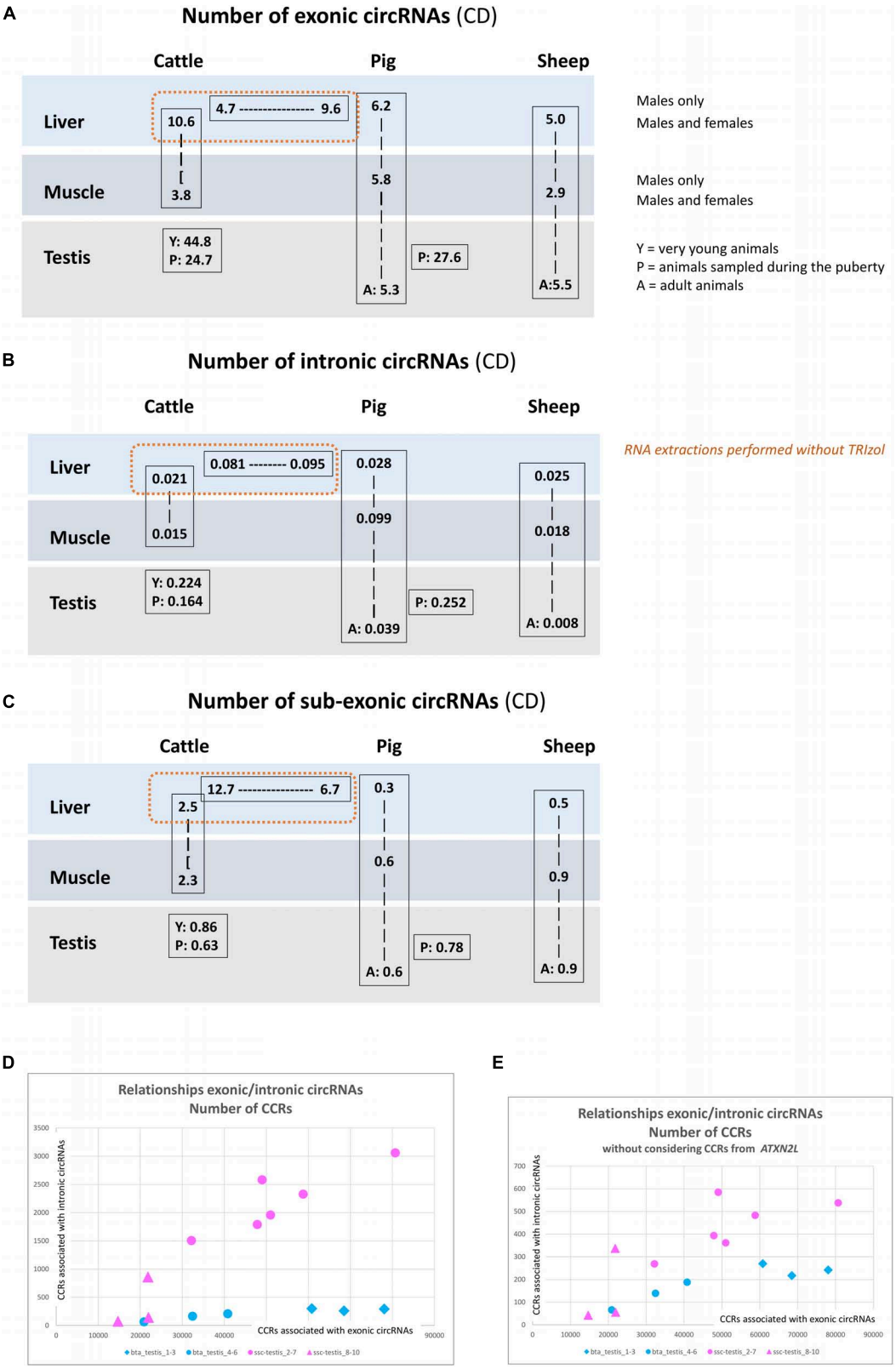


FIGURE 4 | circRNAs detected by CD. **(A)** Number of exonic circRNA detected by CD per million of reads uniquely mapped. **(B)** Number of intronic circRNA detected by CD per million of reads uniquely mapped. **(C)** Number of sub-exonic circRNA detected by CD per million of reads uniquely mapped. **(D,E)** Relationship between the detection of intronic and exonic circRNAs.

possible that they are false positives. Thus, they were kept as non-annotated circRNAs.

We noticed in porcine muscle that a large proportion of circRNAs was not assigned to a specific chromosome, but was localized on unassigned scaffolds. In ssc_muscle_3 and _4, more than 60% of the circRNAs were localized on these unassigned scaffolds, while only 4.5% of all annotated porcine genes are localized there. For example, in ssc_muscle_3 we counted 3,447 circRNAs localized on these unassigned sequences among the total of 5,540 circRNAs characterized. Even more specifically, 2,494 of them were localized on AEMK02000489 and 930 on AEMK02000695. Except a few sub-exonic circRNAs from a gene, which is probably not fully included in the AEMK02000489 scaffold, these circRNAs remained without annotation. These two scaffolds include only 51 and 16 kb of sequence, respectively, and several RNA genes were suspected to be localized there. These observations about circRNAs assigned to these two scaffolds confirm that they included sequences that are transcribed in muscle. Nevertheless, it cannot be excluded that these are only linear transcripts associated with a bad assembly of the sequences in these scaffolds.

More generally, we searched for circRNA clusters without annotation along the chromosomes, because we suspected that the analysis of clusters of unannotated circRNAs would allow us to highlight regions with sequence/assembly/annotation problems. The first example were 450 circRNAs in bta_muscle_6 (40% of unannotated circRNAs characterized in this dataset) detected in the region BTA-2:18,1-18,3Mb. In ssc_muscle_2, we noted 350 circRNAs without annotation in the region SSC-15:84,23-84,49 Mb. The respective genomic sections contain the *Titin* gene (*TTN*), which is an exceptional gene with probably more than 350 exons spread over 300 kb (data from the Ensembl annotation of the human genome), while only 7 or 13 exons were identified in the TTN annotation of the pig and cattle genomes (Ensembl v-102), respectively. In the sheep genome, the *TTN* gene is not annotated in Ensembl (v-102), but in the dataset oar_muscle_2, 200 circRNA without annotation were detected (30% of unannotated circRNAs characterized in this dataset) in the region suspected to contain this gene. The second example was initiated by the characterization of a cluster of circRNAs without annotation in several bovine liver and muscle datasets. This region on BTA-27: 6.21-6.23Mb is known to contain the *Defensin* gene cluster, which is extremely expanded in ruminants. The assembly in this genomic region is difficult due to a substantial number of copies of the same or very similar sequences. In addition, it is assumed that bovine individuals differ in the number of *Defensin* gene copies. In sheep, this region is apparently not included in the reference genome considered.

There are also regions, where not the genome assembly, but only the annotation is deficient and the impact on the number of detected circRNAs remaining without annotation is very limited. For example in a region of BTA-7 (2,73-2,74 Mb) seven circRNAs without an annotation were detected in bta_muscle_6. NCBI and Ensembl do not annotate a gene there. However, RNA-seq data displayed at NCBI would strongly support a gene annotation and a new gene with a very large number of exons had been annotated (Nolte et al., 2020).

To finish, a last example, one circRNA was detected only in the dataset ssc_muscle_4 but with a very high number of CCRs by CD (and with a very high number of BSJs by CIRI2). This circRNAs could be explained by a fusion between one exon from ENSSSCG00000029441 (probably *MYH2*) and one from ENSSSCG00000018005 (*MYH8*). A potential fusion of exons from two annotated genes would explain why the annotation-dependent CE2 pipeline did not retain this circRNA. Even though we cannot discard the hypothesis of a deficiency in genome annotation, we believe rather this to be a structural alteration in the respective genomic region restricted to this particular animal.

The above examples underline that the accuracy of the reference assembly and of the annotation has a major impact on the number of unannotated circRNAs preventing from drawing any conclusion from the observed difference between tissues and samples.

Intronic circRNAs

In bovine and porcine datasets, respectively 53 introns (from 53 genes) and 80 introns (from 79 genes) were able to produce intronic circRNAs. A large part of intronic circRNAs characterized in bovine and porcine datasets were mainly detected in testicular datasets. In porcine datasets, the top-ranked expressed intronic circRNA, from an intron of *ATXN2L,* has ten times more CCRs than the second ranked one (*PEX10*). This intronic circRNA was detected in each of the 18 porcine datasets and is always among the strongest contributors of intronic circRNAs. Even though *ATXN2L* was ranked at the second position in terms of CCRs associated with intronic cirRNAs in bovine testis, the landscape of the production of intronic circRNAs in bovine testis is not dominated by the production of a particular intron. The *ATXN2L* intron concerned is in an orthologous position in pig and cattle. The orthologous ovine gene is not annotated in the reference genome used (Oar_rambouillet_v1.0).

The number of intronic circRNA (**Figure 4B**) seems to be related to the number of exonic circRNA (**Figure 4A**) but not to the number of sub-exonic circRNAs (**Figure 4C**). Specifically, the number of CCRs associated to intronic circRNAs appears to be correlated to the number of CCRs associated to exonic circRNA in porcine and in bovine testis (**Figures 4D,E**). If we consider the six bovine testicular datasets (from young and pubertal animals), the correlation coefficient is 0.92. The correlation is similar between the six porcine testicular datasets from pubertal animals regardless of whether all introns with intronic circRNAs are considered (**Figure 4D**) or if the *ATXN2L* intron with the highest intronic circRNA expression is excluded ($r = 0.88$) (**Figure 4E**). Excluding CCRs from *ATXN2L* enables a comparison of intronic circRNA contents for bovine and porcine testis at the same scale.

So far, intronic circRNAs have been identified only from coding genes (reviewed by Robic and Kühn, 2020), but the current study shows that lncRNAs can also be involved in the production of intronic circRNAs. In pigs, this study highlighted that the lncRNAs ENSSSCG00000048463 and ENSSSCG00000041596 generated intronic circRNAs,

however at a low expression level, with 6 and 11 associated CCRs, respectively.

Analysis of the Production of Sub-Exonic circRNAs

To analyze the catalog of genes capable of producing sub-exonic and/or exonic circular RNAs, we examined the 2 × 12 datasets produced at FBN from bovine liver and muscle. We observe that 1,914 and 839 genes are able to produce exonic circRNAs in liver and muscle, respectively, while only 472 and 228 are able to produce sub-exonic circRNAs in the respective tissues. Only 124 genes produce both exonic and sub-exonic circRNAs in liver, while in muscle we find only 37. The ability to produce sub-exonic circRNA is therefore not related to the ability to produce exonic circRNA.

The top-3 ranked genes producing sub-exonic circRNAs in bovine liver are *ALB, COX1 and FGB*. In bovine muscle, we could identify *COX1, MYH1, MYH2*, and *ACTA1* among the top-5 ranked genes producing sub-exonic circRNAs. In ovine muscle, *XIRP2*, *MYH1*, *ACTA1*, and *MYH7* are among the top-6 ranked genes. Two myosin genes are found in the top ranking list of the strongest contributors to sub-exonic circRNAs in porcine muscle. In ovine liver, nearly half of the CCRs are assigned to sub-exonic circRNAs produced by *ALB*. In the porcine liver, *ALB, FGB,* and *FGA* are the top-3 genes producing sub-exonic circRNAs. In porcine and ovine testis, the strongest contributor of sub-exonic circRNAs is *HSPCA*. All these coding parent genes producing a large number of sub-exonic circRNA are also known to be among the top-ranked contributors of linear transcripts in the respective tissue. The contribution of protein-coding genes to the production of sub-exonic circRNAs represents a large fraction of CCRs characterizing sub-exonic circRNAs (especially in bovine liver and muscle). Moreover, the list of protein coding genes providing the strongest contribution of sub-exonic circRNA seems to be a direct reflection of the respective list for linear transcript contribution. This would be a feature of sub-exonic circRNAs that is not shared with exonic circRNA.

In previous studies, sub-exonic circRNAs had been searched in mono-exonic nc genes (Robic et al., 2020). However, also coding mono-exonic genes contribute to sub-exonic circRNA production. Specifically, the gene *COX1* is a mono-exonic gene localized on the mitochondrial genome. It is among genes able to produce a high number of sub-exonic circRNAs in bovine and ovine liver and in bovine and ovine muscle, while in pigs its contribution is insignificant (in liver and muscle). In cattle, where 13 protein-coding genes are described on the mitochondrial genome, 12 were identified as able to produce sub-exonic circRNAs. All these 13 mitochondrial protein-coding genes are mono-exonic genes (Taanman, 1999).

In spite of our new data on the contribution of coding genes to circRNA production, non-coding genes were also important contributors of sub-exonic circRNA in some datasets. *RNase_MRP* is the strongest contributor of sub-exonic circRNA in each dataset of ssc_testis_2-7. This observation confirms data obtained previously on ssc_testis_1 (Robic et al., 2020). Among the other 60 datasets, there is an important contribution of *RNase_MRP* to sub-exonic circRNAs in oar_liver_5. We also noted a significant contribution of *RNaseP_nuc* to subexonic circular RNAs in bta_muscle_9 while no sub-exonic circular RNA of this gene was detected in the other bovine muscle datasets. Nevertheless, apart from the sub-exonic circRNAs produced by ribozymes, we have to be careful with respect to the possible production of sub-exonic circRNAs by sncRNAs. Some batches seem to be very rich in some sncRNAs while others display not a single read aligned on the respective reference genome (data not shown). With the currently available data and metadata descriptions, it is difficult to differentiate a tissue/age specificity from a difference resulting from a technical bias e.g., the RNA extraction methodology.

Our data show that genes able to produce sub-exonic circRNA can be separated into two sub-groups. Mono-exonic and nc genes were already described as being able to produce sub-exonic circRNAs, and this study shows that mitochondrial genes, which are protein-coding and mono-exonic genes, are also concerned. Furthermore, this study demonstrates that multi-exonic genes, in particular protein-coding genes, can also produce sub-exonic circRNAs. In addition, the coding-genes that are major contributors of sub-exonic circRNAs strongly contribute also to the production of linear transcripts. All data were reported in **Supplementary List 4**.

circRNAs and Non-coding Genes

The current knowledge about nc genes is still poor in the livestock species investigated here, which has an impact on evaluating their contribution to circRNA production. While there is a similar number of protein-coding genes annotated in livestock and human genomes (cow-21,861, pig-21,280, sheep-20,477, human-20,448, Ensembl v-101), comparing the number of annotated pseudogenes (cow-492, pig-1,626, sheep-830, human-15,217) and of lncRNAs (cow-1,480, pig-6,790, sheep-2,229, human-16,909) demonstrates that non-coding genes are still poorly described in livestock species. This study shows that nc genes can also contribute to the production of intronic, sub-exonic and exonic circRNAs. For the production of intronic circRNAs, the current study highlights only two lncRNAs, but for the sub-exonic circRNAs, the contribution of non-coding genes is unquestionable (**Table 2**). In cattle and pigs, 945 and 998 genes were characterized as able to produce sub-exonic circRNAs (**Table 2**), respectively. Among them, 4 and 18 were lncRNAs. Surprisingly in sheep, where only 462 genes are characterized as able to produce sub-exonic circRNAs, we found a relatively higher number of lncRNAs (15). There is a much higher number of lncRNAs involved in exonic circRNA production in sheep compared to pigs considering the number of overall annotated lncRNAs (CE2 + CIRI2, **Table 3**). For example, in ovine testis, we observed that the strongest contributor of exonic circRNAs in terms of BSJ (**Supplementary Table 2**) is a lnc. This lncRNA was able to produce eleven exonic circRNAs of which the three most expressed were ranked at #2, #38, and #50 among the top-ranked exonic circRNAs in ovine testis (**Supplementary Table 1**). This might be due to fewer, but more precisely annotated lncRNA genes in sheep including a higher number of described exons. All non-coding genes, which were confirmed by both approaches (CE2 + CIRI2 and CD) as able to produce exonic circRNAs, are reported in

TABLE 2 | Coding and non-coding genes are to produce sub-exonic circRNAs.

Genes concerned by sub-exonic circRNAs		Protein-coding genes		Long non-coding genes		Pseudogenes		Small non-coding genes
		N. of genes concerned	N. of exons concerned	N. of genes concerned	N. of exons concerned	N. of genes concerned	N. of exons concerned	N. of genes concerned
Cattle	945	908	1,451	4	4	3	3	30 11 misc_RNA, 2 ribozyme, 2 Snc, 2 Sca_RNA, 13 Sno
Pig	998	954	1,228	19	22	6	7	19 2 misc_RNA, 2 ribozyme, 2 Sca_RNA, 12 Sno, 1 Y_RNA
Sheep	465	434	569	15	20	7	9	9 1 misc_RNA, 2 ribozyme, 1 Sca_RNA, 5 Sno

Genes able to produce sub-exonic circRNAs were characterized using CD approach. Lists were available in Supplementary List 4.

TABLE 3 | Non-coding genes able to produce exonic circRNAs.

	lncRNAs No. of genes concerned	Pseudogenes No. of genes concerned	sncRNAs No. of genes concerned
Cattle	6	0	1 Sno
Pig	32	7	1 Sno
Sheep	103	2	0

Only non-coding genes highlighted by both approaches used for circRNA detection (CE2 + CIRI2 and CD) were considered as able to produce exonic circRNAs. Their respective Ensembl_Id were reported in Supplementary File 4.

Supplementary File 4. Non-coding genes are also involved in the production of exonic circRNA, not only lncRNAs but also pseudogenes and snoRNAs (**Table 3**).

Comparison of Circular Transcriptomes (exonic circRNAs)

Comparison of Circular Transcriptomes Between Tissues

When we examined the number of genes producing the 100 strongest expressed exonic circRNAs (CE2 + CIRI2) of a given batch (hereafter Top-100/circRNAs), we counted 87 to 89 distinct genes for the five porcine batches, and 92 to 93 for the three ovine batches. For four bovine batches 90 to 97 genes produced the 100 strongest expressed exonic circRNAs, the exception was in muscle where only 83 distinct parent genes were identified (**Supplementary Table 1**). As we previously proposed a comparison based on the most highly expressed exonic circRNA (**Figure 2A**), we propose now a comparison based on genes with the highest expression in terms of BSJ. The Top-100 list of parent genes most strongly producing exonic circRNAs was established for each batch from the three species (hereafter Top-100/genes, **Supplementary Table 2**), and analyzed. The levels of overlap between testis/liver/muscle appeared higher in sheep than in cattle or pigs (**Figure 2B-3** in yellow). In cattle, the levels of overlap observed for pairwise comparisons of Top-100/genes (**Figure 2B-1**) were similar to those noted for Top-100/circRNAs (**Figure 2A-1** in blue). The main exception is the overlaps between the two batches from bovine liver (87% for Top-100/genes and 72% for Top-100/circRNAs). In pigs, the level of overlap observed for pairwise comparisons of Top-100/genes (**Figure 2B-2**) were almost systematically higher than those noted for Top-100/circRNAs (**Figure 2A-2**). On Top-100/genes, overlaps between the pubertal and adult batches from porcine testis (84%) were better than the overlaps observed between the two batches from porcine liver (74% for Top-100/genes, **Figure 2B-2**). Exonic circRNA production with high BSJ seemed focused on a group of genes that produce several distinct exonic circRNAs with a balance of circular isoforms which would be dependent on considered tissue (**Supplementary Table 1** and **Supplementary File 5**). One example to illustrate this: The gene *NUP210L* was ranked at the second position in terms of BSJs counts in porcine pubertal testis, and 30 distinct exonic circRNAs were characterized with a dominant form. In adult porcine testis this gene was ranked at the third position with 23 circular isoforms characterized but without dominant form. When we examined the Top-100/genes established for the porcine batches (**Supplementary Table 2**), we noticed that *KANSL1L* was among the strongest producers of exonic circRNAs (CE2 + CIRI2) in the five batches (#1 to #7). The bovine orthologous gene produced also many exonic circRNAs, with the highest ranking noted for testis from young bulls (#17) and the lowest ranking for pubertal testis (#62).

Impact of Genes Able to Produce Multiple Exonic circRNAs

We further examined genes with multiple exonic circRNAs characterized. The protein-coding gene *SMARCC1* is able to produce 41 distinct exonic circRNAs in the porcine testis (CE2 + CIRI2) (**Supplementary File 5**). Less than 30 exons were reported for this porcine gene (as in humans), which are spread across 190 kb (Ensembl, v-102). In contrast, the bovine *SMARCC1* is able to produce only five distinct exonic circRNAs in testis and no circRNAs from the ovine *SMARCC1* gene were identified. This result was confirmed among the exonic cirRNAs characterized by CD. The strongest producer of distinct circRNAs (**Supplementary File 5**) in bovine testis is *DNAH14,* but this gene does not appear in the ovine and porcine lists, because of poor knowledge (sequences/annotation) about this gene in these species. In muscle, the gene producing the largest diversity of circRNAs is the same in the three species (**Supplementary File 5**). The *Nebulin* gene (*NEB*) is a gene with a large number of exons. In humans, more than 180 exons were characterized in a region of 200 kb (Ensembl v-102). Among the three species investigated in our study, the highest number of distinct exonic circRNAs from this gene is noted for cattle. In the three species, *NEB* produced this diversity of exonic circRNAs quasi-exclusively in muscle (29/30 cattle, 15/16 pigs, 17/19 sheep). When we examined lists of genes present in Top-100/genes (**Supplementary Table 2**), we noticed that these lists contained mainly genes with multiple exonic circRNAs characterized. *NEB* was a good example to illustrate this. In porcine and ovine muscle, *NEB* was ranked at the third position of genes expressing a high quantity of circular transcripts where 16 and 19 circular isoforms were respectively characterized (**Supplementary Table 2**). In pigs and in sheep, a circular form among all exonic circRNAs from *NEB* appeared dominant and this dominant form was ranked at #2 and #4 on the ovine and porcine Top-100/circRNAs in muscle (**Supplementary Table 1**). These dominant alternative circular transcripts are not pig-sheep orthologous circRNAs. In bovine muscle, 30 distinct *NEB* circRNAs were characterized and this parent gene was ranked at the second position in terms of BSJs counts (**Supplementary Table 2**). Even though the expression of the strongest expressed alternative form of *NEB* circRNA was 20 times higher than the lowest, there was not really a dominant form in bovine muscle.

Comparison of the Circular Transcriptome Between Species

The comparison of the expression profiles between species was performed using two different approaches. First, a comparison between the expression profiles of orthologous parent genes was performed (see Methods). We observe here a clear correlation between the expression profile between species for the same tissue as exemplified by the comparison of sheep and bovine tissues expression profiles (**Figure 5A**). Ranking the expression profiles from highest to lowest for each species and each tissue however underlines that this ranking is not strictly conserved between the three species (**Table 4**). From the examination of this table, it would be tempting to deduce (**Table 4**) that for example (1) the circular transcripts expression of ovine *SUGT1* would be a differential characteristic with respect to pig muscle (2) the circular transcripts expression of ovine *TRDN* would be a differential characteristic with respect to bovine muscle. When we examined respective annotations available, this suggestion appeared possible for *SUGT1*. In contrast, this suggestion did not stand up to the examination of the annotation of the bovine *TRDN* gene. Moreover, when we examined these data, we found no clear overlap as demonstrated by two examples: (1) The *NEB* gene was ranked high across all three species: at #3 on the ovine list and also #3 on the porcine list and #23 on the bovine list. (2) In contrast, the #8 of the ovine list (*SLC9A2*) was found at #60 and #2,854 on the bovine and porcine lists, respectively. The genes that are the strongest producers of circRNA in sheep muscle are not necessarily genes that produce a lot of circRNA in pig or cattle muscle.

In order to bypass the limitation of comparing the expression of parent genes, we identified directly orthologous circRNAs based on sequence similarities (see Methods). From 3,899 ovine circRNAs and 8,723 bovine circRNAs, we were able to identify 1,832 orthologous circRNAs (see Suppl_Lists-1). For comparison, the Ensembl one-to-one ortholog set contains 16,110 orthologs, among 21,861 bovine and 20,477 ovine coding genes. Again, we observe a correlation between the expressions of circRNAs between species for the same tissue (**Figure 5B**) suggesting that, just as the linear transcriptome (Kryuchkova-Mostacci and Robinson-Rechavi, 2016), the circRNA transcription profile is, at least in part, conserved across species.

DISCUSSION

The imbalance between batches observed during the circRNA characterization phase and other issues (e.g., incomplete genome assemblies/annotations) precluded us to perform a global comparison between tissues and between species. Instead, we highlight a few examples to show different and sometimes seemingly contradictory results to demonstrate the complex issue of these comparisons.

Even though each of the datasets considered individually was of correct quality, the agglomeration of datasets from different origins proved to be difficult. This was a surprising outcome, because many studies used a similar approach with mRNA-seq datasets from different sources (Soumillon et al., 2013; Fang et al., 2020). However, while it is difficult to compare the number of circRNAs when datasets come from different sources, the contents in circRNAs (the most expressed circRNAs or the genes with the highest expressing of circRNA in terms of BSJ) are quite comparable. We can put up the hypothesis that protocols for RNA preparation and sequencing have a significant impact on circRNA recovery. We assume that a subset of circular RNAs may be present in tissues in a complex form and that RNA purification methods may differ with respect to their recovery (Pamudurti et al., 2017; Ragan et al., 2019). The differences between RNA preparation protocols are not always well explained but, for example, we are sure that there are differences in the use or non-use of TRIzol, which exist between

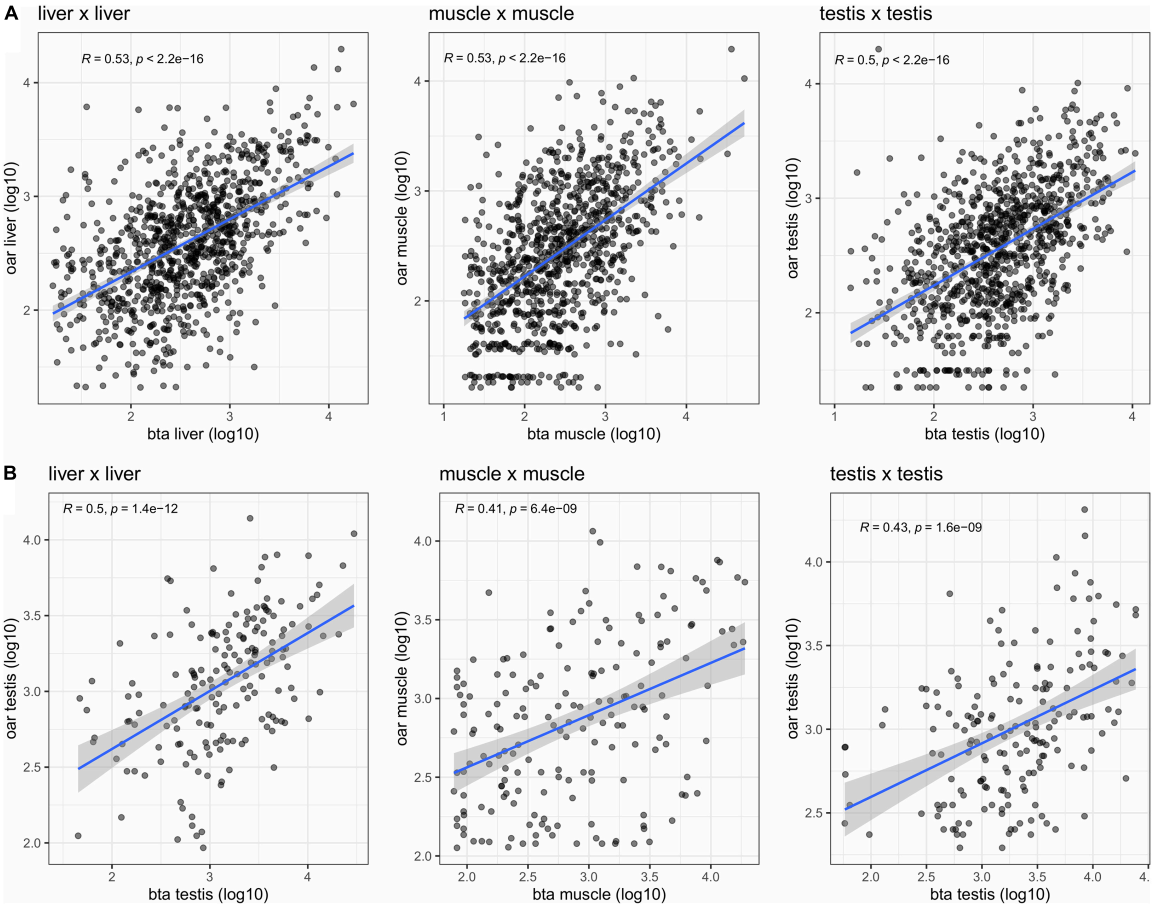


FIGURE 5 | Comparative analysis of the bovine and ovine transcriptomes. **(A)** Circular expression profiles of orthologous genes (measured as number of BSJs) were compared in liver, and in muscle. **(B)** Expression profiles of orthologous circRNAs (measured as number of BSJs) were compared in liver, and in muscle. Only genes/circRNAs with a substantial expression were kept for these correlation analyses (log10(BSJs) > 1.2 for each tissue and each species).

TABLE 4 | Strongest parent genes for cirRNA production identified in muscle.

Ovine gene	Rank in oar_muscle	Bovine orthologous gene	Rank in bta_muscle	Porcine orthologous gene	Rank in ssc_muscle
ENSOARG00020005060 *SUGT1*	1	ENSBTAG00000002137	382	ENSSSCG00000039338	–
ENSOARG00020003405 *UBE3A*	2	ENSBTAG00000002487	1,005	ENSSSCG00000004832	4
ENSOARG00020024247 *NEB*	3	ENSBTAG00000006907	23	ENSSSCG00000016397	3
ENSOARG00020019783 *ANO5*	4	ENSBTAG00000019394	392	ENSSSCG00000013344	1,673
ENSOARG00020010566 *LMO7*	5	ENSBTAG00000010693	2	ENSSSCG00000040184	386
ENSOARG00020016807 *PPP2R3A*	6	ENSBTAG00000023416	233	ENSSSCG00000033185	1,189
ENSOARG00020020875 *TRDN*	7	ENSBTAG00000038849	–	ENSSSCG00000027613	263
ENSOARG00020002529 *SLC9A2*	8	ENSBTAG00000001706	60	ENSSSCG00000008153	2,854
ENSOARG00020012532 *SENP6*	9	ENSBTAG00000005869	1	ENSSSCG00000020702	733
ENSOARG00020003993 *ANO6*	10	ENSBTAG00000002902	130	ENSSSCG00000000804	–
ENSOARG00020021163 *SLTM*	11	ENSBTAG00000011319	853	ENSSSCG00000004592	20
ENSOARG00020002353 *SNX13*	12	ENSBTAG00000014074	207	ENSSSCG00000024761	273
ENSOARG00020006228 *MYBPC1*	13	ENSBTAG00000011392	3	ENSSSCG00005042452	–
ENSOARG00020025783 *GOLGA4*	14	ENSBTAG00000016563	377	ENSSSCG00000011243	–
ENSOARG00020017442 *ZEB1*	15	ENSBTAG00000020053	704	ENSSSCG00000011025	630

At left, the list is provided encompassing the top-15 parent genes for exonic circular RNAs with respect to the number of BSJs identified in sheep muscle. In the center and on the right, the respective ranking is given for orthologous genes in the list of parent genes for exonic circular RNA in bovine and porcine muscle.

batches for this study. For example, the RNA produced for the ssc_testis_1-7 datasets was obtained from a dry-powdered tissue sample before being treated with TRIzol (Robic et al., 2016). In contrast, datasets generated at FBN or at Roslin Institute were produced from a tissue sample homogenized directly in TRIzol (Clark et al., 2017; Nolte et al., 2019). Moreover, an additional on-column-purification step was performed for RNAs extracted at FBN (bta_muscle_1-12) (Nolte et al., 2019) or at INRAE (ssc_testis_1-7) (Robic et al., 2019). Some protocols are described too succinctly to be sure that this type of additional step did not performed. These observations demonstrate the need for harmonized or at least fully described laboratory methods attached as metadata to enable samples to be fully useful in functional annotation of genomes as agreed upon in the global FAANG initiative.

Although the diversity regarding the source of datasets had somewhat limited our analyses, we were able to show that the ruminant liver contains more exonic circRNA than muscle. In testis, the number of exonic circRNAs seemed associated with the age of the animals. In bulls, the testis contained more circRNAs at birth than at puberty. An inverse dynamic was observed in rat (Zhou et al., 2018; Gong et al., 2020). Nevertheless, at birth, a rodent's testis presents large differences with a bull testis (Fujihara et al., 2011; McGowan et al., 2018; Picut et al., 2018). When we compared the circRNA expression of two datasets from the same tissue of the same species, we observed differences, but much less than those between two tissues of the same species. However, the similarities between species are more difficult to quantify, because annotations relative to parent genes are often deficient in at least one of the species considered. The overlap between pubertal testis and testis at other stages led to an intermediate value and showed that the testis is a tissue with a maturation in progress. We showed that there are points of agreement in the circular transcriptome of the same tissue in two species, but also many divergences. Some of the strongest parental genes for exonic circular RNAs may be also among those genes, which produce a large quantity of circular transcripts in several tissues. Nevertheless, this characteristic of high circRNA expression across tissues may be limited to one species (*KANSL1L* in pigs). Moreover, the parent genes of exonic circRNAs are often capable of producing several distinct circRNAs. This has an impact on the composition of the circular transcriptome and the balance between the different circular isoforms contributed also to the composition of circular transcriptome. It seems that among these circular isoforms, there may be a dominant form, but this is not a rule. The balance between the different isoforms will have to be studied in the future, as this seems to be a very specific question for the circular transcriptome. Analyses presented here showed that it is not enough to have a set of orthologous genes capable of producing circRNAs to obtain a similar circRNA landscape in the same tissue from both species. The fact that exonic circRNAs can be produced from the same exons (orthologous circRNA as found by Suenkel et al., 2020) or not appears at this stage as a detail. For example *SMARCA5* is known to produce exonic circRNAs in connection with biological data in humans (Kong et al., 2017) and pigs (Robic et al., 2019), but the exons involved are different. One future direction might be to take into account the relative "weight" of exonic circRNAs/parent genes for comparative analyses of several circular transcriptomes. Among lists of genes able to produce exonic circRNAs in one species, we can find a proportion of genes, which are able to produce exonic circRNAs also in a second species and in the same tissue. Nevertheless, the relative contribution of those genes to the second circular transcriptome is not predictable. There may be points of concordance in the circular transcriptome of the same tissue in two species, but it will be difficult to conclude from one species to another demonstrating the need to conduct a comprehensive characterization within each species. Besides, we would like to emphasize that in this study we considered circRNAs distinguishable by their circular junction. We know nothing about the internal structure of circRNA and multiple distinct circular RNAs can share the same circular junction (Dodbele et al., 2021).

This study shows that multi-exonic genes can also generate sub-exonic circRNAs. These multi-exonic genes were most often protein-coding genes, but some lncRNA and pseudogenes were also highlighted. Because the list of genes, which are the strongest contributors of sub-exonic circRNAs (especially in bovine liver and muscle), seems to be a direct reflection of the respective list of contributors of linear transcripts, we suspect that these circRNAs from multi-exonic genes are mostly the result of splicing machinery errors or a destruction process of linear transcripts. For sub-exonic circRNA from mono-exonic genes, this current study confirmed the previous study (Robic et al., 2020, which was based only on the ssc_testis_1 dataset) that ribozymes and other snc-RNA genes are able to produce sub-exonic circRNAs. We provide data supporting the production of sub-exonic circRNAs by mitochondrial protein-coding genes, which is new, but not a surprise. The transcription of these mono-exonic genes does not require the splicing step (no intronic sequence to be removed), and frequently sub-exonic circRNAs include a notable part of potentially transcribed sequences (unique exons). These genes have probably kept some features of ancestral genes (prokaryotic genes) including the production of circRNAs (Danan et al., 2012). It was already described for ribozymes (Cervera and de la Pena, 2020). We believe that it is a constitutive phenomenon relative to these mono-exonic genes, where a part of transcripts is circular. We cannot rule out the hypothesis that the transcripts of these genes adopt the circular form for a better life span or a best biological efficiency.

When we wanted to compare results between species, we were confronted with problems related to the reference genome, because some of the genes were not annotated with the same quality in all species. We would like to emphasize that the assignment of a given circRNA to a parent gene is dependent on the knowledge of the genome and available annotation. We began this study with strict constraints on the annotation of the parent circRNA gene (see Materials and Methods). Although excellent analyses to compare circular transcriptomes have recently been published using a different approach (Ji et al., 2019), we still believe that, especially in animal species, it is important to perform comparative studies with only circRNAs with a clearly

identified parent gene (Dong et al., 2017). Moreover, this approach avoid also a large number of false positive circRNA annotations (Kaur et al., 2018; Ragan et al., 2019). This study shows that nc genes can give rise to intronic, sub-exonic and exonic circRNAs. For exonic circRNAs, we were surprised to observe the highest similarities between tissues and the smallest number of distinct exonic circRNA in sheep. Our observations on circRNAs showed that the ovine reference genome might suffer from a deficit of described exons in protein-coding genes. The ovine lncRNAs included in the genome annotation, however, seemed to be better described than lncRNAs from pigs. We would underline that the current knowledge about nc genes is still poor in the livestock species investigated here (Gao et al., 2019; Nolte et al., 2020). When we started this study, we assumed that unannotated circRNAs would reveal the existence of new exons/transcripts/genes, and we thought that the list contained many circRNAs that could be annotated with a single effort on genome annotation as suggested in a previous study (Robic et al., 2020). Nevertheless, the current study revealed that the vast majority of unannotated circRNAs were grouped in clusters along the genome (especially in muscle). We showed that these clusters pointed to genomic regions with problems regarding gene annotation/assembly/sequences. In these genomic regions, the problems are often multiple, and the identification of new linear or circular transcripts seems to be a dangerous process, if it is not associated with a real parallel effort on linear transcriptome annotation and even an improved genome assembly.

This study highlights the importance of improving genome annotation to better annotate circRNAs observed. To our disappointment, not all detected circRNAs can directly contribute to the annotation of new genes. Nevertheless, we believe that a "wide-angle" approach to study circular RNAs can help locate genomic regions with multiple problems. This study highlights the importance of improving genome annotation to better understand the circRNA production.

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/s.

AUTHOR CONTRIBUTIONS

AR, TF, and CK conceived the overall idea of the study and analyzed the results. CC built the bioinformatics tool to characterize circRNAs with the assistance of TF. CC and TF performed all bioinformatics analyses. TF coordinates the study and analyses. AR performed the formal analysis, validation, and data curation, and prepared and wrote the original draft. All authors reviewed the manuscript, read and agreed to the published version of the manuscript, and contributed substantially to the work reported.

FUNDING

INRAE (GenPhySE and Animal Genetics division) and the Institute of Genome Biology of FBN supported studies around circular RNAs. These studies, which are fully associated with the FAANG initiative, were also supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement 815668 (BovReg).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.665153/full#supplementary-material

Supplementary File 1 | Some features relative to circRNAs characterization (CE2 + CIRI2).

Supplementary File 2 | Characterization of exonic circRNAs performed by CE2 + CIRI2.

Supplementary File 3 | All statistics relative to comparisons of the number of circRNAs.

Supplementary File 4 | List of non-coding genes considered as able to produce exonic circRNAs (CE2 + CIRI2 and CD).

Supplementary File 5 | Genes able to produce multiple exonic circRNAs (CE2 + CIRI2).

Supplementary Table 1 | Top-100/circRNAs: Lists of 100 strongest expressed circRNAs in 5 porcine, 5 bovine and 3 ovine batches.

Supplementary Table 2 | Top100/parent genes: Lists of 100 genes expressed the most circRNAs (in term of BSJ) in 5 porcine, 5 bovine and 3 ovine batches.

Supplementary List 1 | This file consists of 6 large sheets listing the exonic circRNAs detected by CE2 + CIRI2 and also the bovine-ovine orthologous circRNAs.

Supplementary List 2 | This file consists of 6 large sheets relative to exonic circRNAs detection by CD and their parent genes.

Supplementary List 3 | This file consists of 6 sheets relative to intronic circRNAs detection by CD and their parent genes.

Supplementary List 4 | This file consists of 9 large sheets relative to sub-exonic circRNAs detection by CD and their parent genes.

REFERENCES

Au, C. H., Wa, A., Ho, D. N., Chan, T. L., and Ma, E. S. (2016). Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms. *Diagn. Pathol.* 11:11.

Barrett, S. P., Parker, K. R., Horn, C., Mata, M., and Salzman, J. (2017). ciRS-7 exonic sequence is embedded in a long non-coding RNA locus. *PLoS Genet.* 13:e1007114. doi: 10.1371/journal.pgen.1007114

Cattle (2021). *Cattle*. Cambridge: Ensembl. http://www.ensembl.org/Bos_taurus/Info/Index

Cervera, A., and de la Pena, M. (2020). Small circRNAs with self-cleaving ribozymes are highly expressed in diverse metazoan transcriptomes. *Nucleic Acids Res.* 48, 5054–5064. doi: 10.1093/nar/gkaa187

Chen, L. L. (2020). The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nat. Rev. Mol. Cell Biol.* 21, 475–490. doi: 10.1038/s41580-020-0243-y

Chen, L., Yang, R., Kwan, T., Tang, C., Watt, S., Zhang, Y., et al. (2020). Paired rRNA-depleted and polyA-selected RNA sequencing data and supporting multi-omics data from human T cells. *Sci. Data* 7:376.

Cheng, J., Metge, F., and Dieterich, C. (2016). Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* 32, 1094–1096. doi: 10.1093/bioinformatics/btv656

Clark, E. L., Bush, S. J., McCulloch, M. E. B., Farquhar, I. L., Young, R., Lefevre, L., et al. (2017). A high resolution atlas of gene expression in the domestic sheep (Ovis aries). *PLoS Genet.* 13:e1006997. doi: 10.1371/journal.pgen.1006997

Danan, M., Schwartz, S., Edelheit, S., and Sorek, R. (2012). Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res.* 40, 3131–3142. doi: 10.1093/nar/gkr1009

Das, D., Das, A., Sahu, M., Mishra, S. S., Khan, S., Bejugam, P. R., et al. (2020). Identification and Characterization of Circular Intronic RNAs Derived from Insulin Gene. *Int. J. Mol. Sci.* 21:4302. doi: 10.3390/ijms21124302

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dodbele, S., Mutlu, N., and Wilusz, J. E. (2021). Best practices to ensure robust investigation of circular RNAs: pitfalls and tips. *EMBO Rep.* 22:e52072.

Dong, R., Ma, X. K., Chen, L. L., and Yang, L. (2017). Increased complexity of circRNA expression during species evolution. *RNA Biol.* 14, 1064–1074. doi: 10.1080/15476286.2016.1269999

Fang, L., Cai, W., Liu, S., Canela-Xandri, O., Gao, Y., Jiang, J., et al. (2020). Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Res.* 30, 790–801. doi: 10.1101/gr.250704.119

Fujihara, M., Kim, S. M., Minami, N., Yamada, M., and Imai, H. (2011). Characterization and in vitro culture of male germ cells from developing bovine testis. *J. Reprod. Dev.* 57, 355–364. doi: 10.1262/jrd.10-185m

Gao, Y., and Zhao, F. (2018). Computational Strategies for Exploring Circular RNAs. *Trends Genet.* 34, 389–400. doi: 10.1016/j.tig.2017.12.016

Gao, Y., Li, S., Lai, Z., Zhou, Z., Wu, F., Huang, Y., et al. (2019). Analysis of Long Non-Coding RNA and mRNA Expression Profiling in Immature and Mature Bovine (*Bos taurus*) Testes. *Front. Genet.* 10:646. doi: 10.3389/fgene.2019.00646

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* 16:4.

Gao, Y., Zhang, J., and Zhao, F. (2018). Circular RNA identification based on multiple seed matching. *Brief Bioinform.* 19, 803–810. doi: 10.1093/bib/bbx014

Gong, B., Xu, J., and Tong, W. (2020). Landscape of circRNAs Across 11 Organs and 4 Ages in Fischer 344 Rats. *Chem. Res. Toxicol.* 34, 240–246. doi: 10.1021/acs.chemrestox.0c00144

Hansen, T. B. (2018). Improved circRNA Identification by Combining Prediction Algorithms. *Front. Cell Dev. Biol.* 6:20. doi: 10.3389/fcell.2018.00020

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19, 141–157. doi: 10.1261/rna.035667.112

Ji, P., Wu, W., Chen, S., Zheng, Y., Zhou, L., Zhang, J., et al. (2019). Expanded Expression Landscape and Prioritization of Circular RNAs in Mammals. *Cell Rep.* 26, 3444–3460e3445.

Kaur, S., Mirza, A. H., and Pociot, F. (2018). Cell Type-Selective Expression of Circular RNAs in Human Pancreatic Islets. *Noncoding RNA* 4:38. doi: 10.3390/ncrna4040038

Kong, Z., Wan, X., Zhang, Y., Zhang, P., Zhang, Y., Zhang, X., et al. (2017). Androgen-responsive circular RNA circSMARCA5 is up-regulated and promotes cell proliferation in prostate cancer. *Biochem. Biophys. Res. Commun.* 493, 1217–1223. doi: 10.1016/j.bbrc.2017.07.162

Kristensen, L. S., Andersen, M. S., Stagsted, L. V. W., Ebbesen, K. K., Hansen, T. B., and Kjems, J. (2019). The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.* 20, 675–691.

Kryuchkova-Mostacci, N., and Robinson-Rechavi, M. (2016). Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. *PLoS Comput. Biol.* 12:e1005274. doi: 10.1371/journal.pcbi.1005274

Li, H. M., Ma, X. L., and Li, H. G. (2019). Intriguing circles: Conflicts and controversies in circular RNA research. *RNA* 10:e1538. doi: 10.1002/wrna.1538

Liu, X., Hu, Z., Zhou, J., Tian, C., Tian, G., He, M., et al. (2020). Interior circular RNA. *RNA Biol.* 17, 87–97. doi: 10.1080/15476286.2019.1669391

McGowan, M., Holland, M. K., and Boe-Hansen, G. (2018). Review: Ontology and endocrinology of the reproductive system of bulls from fetus to maturity. *Animal* 12, s19–s26.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338.

Nolte, W., Weikard, R., Brunner, R. M., Albrecht, E., Hammon, H. M., Reverter, A., et al. (2020). Identification and Annotation of Potential Function of Regulatory Antisense Long Non-Coding RNAs Related to Feed Efficiency in *Bos taurus* Bulls. *Int. J. Mol. Sci.* 21:3292.

Nolte, W., Weikard, R., Brunner, R. M., Albrecht, E., Hammon, H. M., Reverter, A., et al. (2019). Biological Network Approach for the Identification of Regulatory Long Non-Coding RNAs Associated With Metabolic Efficiency in Cattle. *Front. Genet.* 10:1130. doi: 10.3389/fgene.2019.01130

Pamudurti, N. R., Bartok, O., Jens, M., Ashwal-Fluss, R., Stottmeister, C., Ruhe, L., et al. (2017). Translation of CircRNAs. *Mol. Cell* 66, 9–21e27.

Picut, C. A., Ziejewski, M. K., and Stanislaus, D. (2018). Comparative Aspects of Pre- and Postnatal Development of the Male Reproductive System. *Birth Defects Res.* 110, 190–227.

Pig (2021). *Pig*. Cambridge: Ensembl. http://www.ensembl.org/Sus_scrofa/Info/Index

Ragan, C., Goodall, G. J., Shirokikh, N. E., and Preiss, T. (2019). Insights into the biogenesis and potential functions of exonic circular RNA. *Sci. Rep.* 9:2048.

Rawlings, N., Evans, A. C., Chandolia, R. K., and Bagu, E. T. (2008). Sexual maturation in the bull. *Reprod. Domest. Anim.* 43(Suppl. 2), 295–301.

Robic, A., and Kühn, C. (2020). Beyond Back Splicing, a Still Poorly Explored World: Non-Canonical Circular RNAs. *Genes* 11:1111.

Robic, A., Demars, J., and Kühn, C. (2020). In-Depth Analysis Reveals Production of Circular RNAs from Non-Coding Sequences. *Cells* 9:1806.

Robic, A., Faraut, T., Djebali, S., Weikard, R., Feve, K., Maman, S., et al. (2019). Analysis of pig transcriptomes suggests a global regulation mechanism enabling temporary bursts of circular RNAs. *RNA Biol.* 16, 1190–1204.

Robic, A., Feve, K., Riquet, J., and Prunier, A. (2016). Transcript levels of genes implicated in steroidogenesis in the testes and fat tissue in relation to androstenone accumulation in fat of pubertal pigs. *Domest. Anim. Endocrinol.* 57, 1–9.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140.

Salzman, J., Gawad, C., Wang, P. L., Lacayo, N., and Brown, P. O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* 7:e30733. doi: 10.1371/journal.pone.0030733

Sheep (2021). *Sheep*. Cambridge: Ensembl. http://www.ensembl.org/Ovis_aries_rambouillet/Info/Index

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., et al. (2013). Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* 3, 2179–2190.

Stoll, L., Rodriguez-Trejo, A., Guay, C., Brozzi, F., Bayazit, M. B., Gattesco, S., et al. (2020). A circular RNA generated from an intron of the insulin gene controls insulin secretion. *Nat. Commun.* 11:5611.

Suenkel, C., Cavalli, D., Massalini, S., Calegari, F., and Rajewsky, N. (2020). A Highly Conserved Circular RNA Is Required to Keep Neural Cells in a Progenitor State in the Mammalian Brain. *Cell Rep.* 30, 2170–2179e2175.

Taanman, J. W. (1999). The mitochondrial genome: structure, transcription, translation and replication. *Biochim. Biophys. Acta* 1410, 103–123.

Taggart, A. J., Lin, C. L., Shrestha, B., Heintzelman, C., Kim, S., and Fairbrother, W. G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. *Genome Res.* 27, 639–649.

Wang, P. L., Bao, Y., Yee, M. C., Barrett, S. P., Hogan, G. J., Olsen, M. N., et al. (2014). Circular RNA is expressed across the eukaryotic tree of life. *PLoS One* 9:e90859. doi: 10.1371/journal.pone.0090859

Xiao, M. S., Ai, Y., and Wilusz, J. E. (2020). Biogenesis and Functions of Circular RNAs Come into Focus. *Trends Cell Biol.* 30, 226–240.

Yang, W., Zhao, F., Chen, M., Li, Y., Lan, X., Yang, R., et al. (2020). Identification and characterization of male reproduction-related genes in pig (*Sus scrofa*)

using transcriptome analysis. *BMC Genomics* 21, 381. doi: 10.1186/s12864-020-06790-w

Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* 13:e1005420. doi: 10.1371/journal.pcbi.1006158

Zhang, J., Chen, S., Yang, J., and Zhao, F. (2020). Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat. Commun.* 11:90.

Zhang, X. O., Dong, R., Zhang, Y., Zhang, J. L., Luo, Z., Zhang, J., et al. (2016). Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.* 26, 1277–1287.

Zhang, Y., Zhang, X. O., Chen, T., Xiang, J. F., Yin, Q. F., Xing, Y. H., et al. (2013). Circular intronic long noncoding RNAs. *Mol. Cell* 51, 792–806.

Zhao, Q., Liu, J., Deng, H., Ma, R., Liao, J. Y., Liang, H., et al. (2020). Targeting Mitochondria-Located circRNA SCAR Alleviates NASH via Reducing mROS Output. *Cell* 183, 76–93e22.

Zhou, T., Xie, X., Li, M., Shi, J., Zhou, J., Knox, K., et al. (2018). Rat BodyMap transcriptomes reveal unique circular RNA features across tissue types and developmental stages. *RNA* 24, 1443–1456.

Zucko, D., and Boris-Lawrie, K. (2020). Circular RNAs Are Regulators of Diverse Animal Transcriptomes: One Health Perspective. *Front. Genet.* 11:999. doi: 10.3389/fgene.2020.00999

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.

Copyright © 2021 Robic, Cerutti, Kühn and Faraut. 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 May 2021
doi: 10.3389/fgene.2021.664260

Large-Scale Multiplexing Permits Full-Length Transcriptome Annotation of 32 Bovine Tissues From a Single Nanopore Flow Cell

*Michelle M. Halstead, Alma Islas-Trejo, Daniel E. Goszczynski, Juan F. Medrano, Huaijun Zhou and Pablo J. Ross**

Department of Animal Science, University of California, Davis, Davis, CA, United States

OPEN ACCESS

Edited by:
Francisco Peñagaricano, University of Wisconsin–Madison, United States

Reviewed by:
James E. Koltes, Iowa State University, United States
Cong-jun Li, Animal Genomics and Improvement Laboratory, Agricultural Research Service, United States Department of Agriculture (USDA), United States

***Correspondence:**
Pablo J. Ross
pross@ucdavis.edu

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

Received: *04 February 2021*
Accepted: *06 April 2021*
Published: *20 May 2021*

Citation:
Halstead MM, Islas-Trejo A, Goszczynski DE, Medrano JF, Zhou H and Ross PJ (2021) Large-Scale Multiplexing Permits Full-Length Transcriptome Annotation of 32 Bovine Tissues From a Single Nanopore Flow Cell. Front. Genet. 12:664260. doi: 10.3389/fgene.2021.664260

A comprehensive annotation of transcript isoforms in domesticated species is lacking. Especially considering that transcriptome complexity and splicing patterns are not well-conserved between species, this presents a substantial obstacle to genomic selection programs that seek to improve production, disease resistance, and reproduction. Recent advances in long-read sequencing technology have made it possible to directly extrapolate the structure of full-length transcripts without the need for transcript reconstruction. In this study, we demonstrate the power of long-read sequencing for transcriptome annotation by coupling Oxford Nanopore Technology (ONT) with large-scale multiplexing of 93 samples, comprising 32 tissues collected from adult male and female Hereford cattle. More than 30 million uniquely mapping full-length reads were obtained from a single ONT flow cell, and used to identify and characterize the expression dynamics of 99,044 transcript isoforms at 31,824 loci. Of these predicted transcripts, 21% exactly matched a reference transcript, and 61% were novel isoforms of reference genes, substantially increasing the ratio of transcript variants per gene, and suggesting that the complexity of the bovine transcriptome is comparable to that in humans. Over 7,000 transcript isoforms were extremely tissue-specific, and 61% of these were attributed to testis, which exhibited the most complex transcriptome of all interrogated tissues. Despite profiling over 30 tissues, transcription was only detected at about 60% of reference loci. Consequently, additional studies will be necessary to continue characterizing the bovine transcriptome in additional cell types, developmental stages, and physiological conditions. However, by here demonstrating the power of ONT sequencing coupled with large-scale multiplexing, the task of exhaustively annotating the bovine transcriptome – or any mammalian transcriptome – appears significantly more feasible.

Keywords: transcriptome, annotation, nanopore, cattle, tissue-specific, alternative splicyng, long-read sequencing, full-length transcript

INTRODUCTION

The proteome diversity observed in eukaryotes is largely attributed to alternative transcript isoforms, which result from use of alternate transcription start sites, polyadenylation sites, and splice sites. In particular, the complexity of alternative splicing seems to have increased during the course of evolution (Keren et al., 2010), such that transcript isoforms exist for the majority of genes in higher order eukaryotes (Pan et al., 2008; Wang et al., 2008; Mercer et al., 2012). This diversification of the transcriptome and proteome not only drives adaptation and speciation (Harr and Turner, 2010; Mudge et al., 2011), but also facilitates cellular diversity and the development of complex organisms with tissues and organs (Graveley, 2001; Linker et al., 2019). Indeed, transcript isoforms and splicing patterns vary between cell types, tissues, developmental stages, and environmental conditions (Kalsotra et al., 2008; Wang et al., 2008; Vaquero-Garcia et al., 2016; Zhang et al., 2016). Moreover, because alternative splicing can fundamentally alter protein structure and function, aberrant isoforms have been linked to various diseases, including cancer (Paronetto et al., 2016; Zhang et al., 2019).

More than 90% of human genes are subject to alternative splicing (Pan et al., 2008; Workman et al., 2019); as such, considerable efforts have been made by consortia such as GENCODE to exhaustively annotate transcript isoforms in humans and mice. However, projects seeking to annotate the genomes of non-model organisms generally lack the necessary resources for manual curation. Consequently, transcriptome annotations for non-model organisms, including species of high economic significance like livestock, are often incomplete or inaccurate (Andersson et al., 2015; Ungaro et al., 2017). Moreover, transcriptome complexity and splicing patterns are not well-conserved between species (Barbosa-Morais et al., 2012). Transcript structures inferred from related species are therefore likely to be insufficient or inaccurate.

Worldwide, over a billion cattle (*Bos taurus*) are raised for meat and dairy production (Robinson et al., 2014), and although selection programs have significantly benefited from genomics tools in the past decade (Meredith et al., 2012; Saatchi et al., 2012; Thompson-Crispi et al., 2014; García-Ruiz et al., 2016), a comprehensive characterization of the bovine transcriptome is essential to improve our understanding of the biological processes that underpin complex traits like productivity, efficiency, and disease resistance (Georges et al., 2019).

Until recently, transcriptome annotations – including that of the bovine genome – were primarily based on short-read RNA-seq data from next-generation sequencing (NGS) platforms. The high throughput of these sequencers was optimal for quantifying gene expression, but because of sequencing length limitations, it is necessary to fragment RNA or cDNA during library preparation. The resulting reads are generally shorter (<200 bases) than most full-length transcripts, and although several computational approaches have been developed to reconstruct transcript structures from short-read RNA-seq data, they do not always infer the correct structures (Grabherr et al., 2011; Trapnell et al., 2012; Pertea et al., 2015; Conesa et al., 2016).

Alternatively, long-read sequencing technologies, such as Pacific Biosciences (PacBio) (McCarthy, 2010; Rhoads and Au, 2015) and Oxford Nanopore Technologies (ONT) (Bayega et al., 2018), have made it possible to sequence reads up to 50 kb in length, allowing for the sequencing of full-length transcripts without the need for reconstruction. In recent years, PacBio single-molecule real-time (SMRT) isoform sequencing (Iso-seq) has been implemented to improve transcriptome annotations in humans (Sharon et al., 2013; Tilgner et al., 2014), rabbits (Chen et al., 2017), chickens (Thomas et al., 2014; Kuo et al., 2017), pigs (Li et al., 2018; Beiki et al., 2019), and cattle (Rosen et al., 2020). Indeed, the transcriptome accompanying the most recent bovine genome assembly was curated from both short-read RNA-seq and Iso-seq data (Rosen et al., 2020); however, the Iso-seq dataset was limited, as it included fewer tissue transcriptomes than the short-read RNA-seq data, and was of considerably lower sequencing depth, producing only about a half a million consensus reads.

An alternative long-read sequencing technology, ONT sequencing, measures changes in ionic current as fragments move through protein nanopores, and does not depend on enzyme-based nucleotide incorporation or detection of fluorescence (Ip et al., 2015). Due to its affordability and higher throughput – the ONT PromethION generates 20 times more reads per flow cell than the PacBio Sequel II (Garalde et al., 2018) – ONT has been widely used for transcriptome annotation in organisms ranging from yeast to humans (Sharon et al., 2013; Tilgner et al., 2014; Oikonomopoulos et al., 2016; Byrne et al., 2017; Jenjaroenpun et al., 2018; Kadobianskyi et al., 2019; Seki et al., 2019; Sessegolo et al., 2019; Workman et al., 2019; Müller et al., 2020; Sahoo et al., 2020), permitting the discovery of isoforms that were difficult to observe from short-read sequencing alone (Steijger et al., 2013; Venturini et al., 2018).

Despite the incorporation of Iso-seq data (Rosen et al., 2020), the bovine transcriptome still only includes 1.59 transcripts per gene on average, whereas the human genome annotation accounts for an average of 3.78 transcript isoforms per gene (Ensembl v101 annotations). This discrepancy suggests that the transcriptomic complexity of the bovine genome has yet to be fully characterized, and that current annotations are likely missing information on rare and tissue-specific isoforms. In this study, we coupled ONT sequencing with large-scale multiplexing to identify and characterize the expression of transcript isoforms in cattle. From a single ONT flow cell, we obtained over 25 million full-length uniquely mapped reads, allowing us to characterize the transcriptomes of 32 adult bovine tissues across four individuals. This powerful approach paves the way for future transcriptomic studies, facilitating research on a wider variety of cell types, physiological conditions, and developmental stages. Moreover, the resulting transcript predictions will help to inform selection programs seeking to improve production traits, fertility, and environmental adaptation – factors which are of considerable scientific and economic interest.

MATERIALS AND METHODS

Sample Collection

Tissue samples were collected from two male and two female Line 1 Hereford cattle, aged 14 months old, which were provided by the Fort Keogh Livestock and Range Research lab. Animals were euthanized by captive bolt under USDA inspection at the University of California, Davis, with all permissions obtained and in concordance with Protocol for Animal Care and Use no. 18464 (approved by Institutional Animal Care and Use Committee at the University of California, Davis). Samples were collected within 1–2 h of euthanasia, flash frozen in liquid nitrogen, and stored at –80°C until processing.

RNA Extraction and Library Construction

Frozen tissues kept at –80°C were homogenized with a mortar and pestle in liquid nitrogen. Total RNA was extracted using Trizol (Invitrogen, Carlsbad, CA, United States) followed by a column clean-up using the Direct-zol RNA Mini Prep Plus kit (Zymo Research, Irvine, CA, United States) and performing an in-column DNA digestion. Integrity of the DNase-treated RNA was verified on the Experion electrophoresis system (Bio-Rad, Hercules, CA, United States). For each sample, 50 ng total RNA was transferred to 0.2 ml PCR tubes and adjusted to a final volume of 9 μl with nuclease free water. Reactions were prepared (9 μl total RNA, 1 μl 10 μM VNP primer, 1 μl 10 mM dNTPs) and incubated for 5 min at 65°C, then snap cooled on a pre-chilled freezer block. Strand-switching buffer (4 μl 5x RT buffer, 1 μl RNaseOUT, 1 μl nuclease-free water, and 2 μl 10 μM strand-switching primer) was then added to the snap-cooled, annealed mRNA, and incubated at 42°C for 2 min. One μl of Maxima H Minus Reverse Transcriptase was added, and reactions were incubated at 42°C for 90 min, 85°C for 5 min, then held at 4°C. A round of PCR was used to introduce barcodes to the cDNA using the Oxford Nanopore PCR barcoding expansion 1-96 kit (Cat. No. EXP-PBC096). Barcoding PCR reactions were set up for each cDNA (1 μl PCR barcode, 19 μl first-strand cDNA, 20 μl LongAmp Taq 2x master mix), and cycled for [3 min at 95°C] x1 cycle, [15 s at 95°C, 15 s at 62°C, 7 min at 65°C] x13 cycles, [15 min at 65°C] x1 cycle, then held at 4°C. Each barcoded cDNA was purified in 1x Ampure XP Beads, eluted in 20 μl of nuclease free water and quantified using Qubit. Barcoded cDNAs were pooled in a final volume of 47 μl. The DNA Technologies Core and Expression Analysis Laboratory at the University of California Davis performed adapter ligation on the cDNA pool with the SQK-DCS109 kit following manufacturer's guidelines. Finally, 50 fmol of adapter ligated library was loaded onto a PromethION flow cell (vR9.4.1).

Pre-processing of ONT Sequencing Data

The quality of raw sequencing data, including read length and average quality, was checked using Nanoplot (v1.0.0). Base calling and demultiplexing (**Supplementary Table 1**) were performed using ont-guppy-for-minknow (v3.0.5) and reads with a quality score below 7 were discarded. Data were then processed with Pychopper (v2.4.0) to identify and orient full-length reads; these were then mapped to the ARS-UCD1.2 genome assembly using minimap2 (v2.16r922) (Li, 2018) with options "-ax splice -uf -k14 -G 1000000." The maximum allowable intron size was increased to 1 Mb, based on the longest intron observed in the Ensembl (v101) annotation. Uniquely mapped reads with a minimum quality score of 10 were extracted with Samtools (v1.7).

Preliminary Analysis of Gene Expression

Uniquely mapped reads were used to obtain raw gene expression counts, based on the Ensembl v101 annotations for each species, using HTSeq (v0.11.2) (Anders et al., 2015) with options "-i gene_id –type = exon –stranded = yes –mode = intersection-non-empty." Raw gene counts were subjected to variance stabilizing transformation (VST) with DESeq2 (v1.26.0) (Love et al., 2014) for principal components analysis, conducted with the prcomp function from the R package Stats (v3.6.3). Expression profiles of the top 5,000 genes with the most variance in VST counts were visualized with pheatmap (v1.0.12).

Predicting Transcript Isoforms

Uniquely mapped reads from all samples were pooled to predict transcripts using the Pinfish pipeline (v0.1.0)[1]. Briefly, reads with similar structure were grouped into clusters of three or more alignments, with an exon boundary tolerance of 20 bp and terminal exon boundary tolerance of 60 bp. These transcript clusters were then polished and mapped back to the genome. Polished transcripts were then grouped into "loci" based on 3′ ends and collapsed to remove likely products of RNA degradation, using an internal exon boundary tolerance of 5 bp, a 3′-exon boundary tolerance of 100 bp, and a 5′-exon boundary tolerance of 5,000 bp. Because of the high prevalence of predicted single-exon transcripts, predicted transcripts were then compared to the Ensembl (v101) and NCBI RefSeq (release 106) annotations using gffcompare (v0.12.1), and only single-exon transcripts that demonstrated same-strand overlap with reference exons of protein-coding genes, or which were strongly supported (cluster size $\geq$ 100 alignments), were retained in the final transcript set. The set of predicted transcripts was converted to GTF format using gffread (v0.12.2) and visualized in the Integrated Genomics Viewer (v2.8.9). To visualize repetitive elements, the RepeatMasker track was downloaded from the UCSC genome annotation database for the April 2018 ARS-UCD1.2/bosTau9 assembly.

Comparing Predicted Transcripts to Reference Annotations

Based on gffcompare class codes, predicted transcripts were classified as known isoforms of a reference gene (class code " = " when comparing to either annotation), novel isoforms of a reference gene (class codes 'c,' 'k,' 'j,' 'm,' 'n,' or 'o' when comparing to either annotation, never ' = '), novel loci (class codes 'i,' 'u,' 'y,' or 'x' when comparing to either annotation, never ' = ,' 'c,' 'k,' 'j,' 'm,' 'n,' or 'o'), or potential artifacts (class codes 'e,' 's,' or 'p' when comparing to either annotation, but never any other class codes).

[1]github.com/nanoporetech/pinfish

Characterization of Predicted Transcripts

To determine the novelty of start and end sites of predicted novel isoforms, the TSS and TES of predicted novel isoforms were compared to the TSS and TES of the closest matching reference transcripts (based on gffcompare output). The usage of alternative polyadenylation sites for reference Ensembl transcripts was determined using TAPAS (Arefeen et al., 2018) with read length set to 750 bp, which was the mean read length according to the Nanoplot report. As input for TAPAS, genome-wide read depth was determined with Samtools (v1.7). The prevalence of different alternative splicing events in the final set of predicted transcripts was determined with SUPPA (v2.3), using the function generateEvents to identify local events, including skipped exons, mutually exclusive exons, retained introns, alternative 5′ or 3′ splice sites, and alternative first and last exons. Finally, the coding potential of predicted transcripts was calculated with CPPred (Tong and Liu, 2019) using the built-in human model with default parameters. To determine if predicted intergenic transcripts (gffcompare class code 'u') preferentially occurred near annotated genes, distance from each predicted intergenic transcript to the nearest reference gene was calculated using Bedtools closest (v2.26.0) with option "-d." For comparison, the genomic coordinates of predicted intergenic transcripts were randomized with Bedtools shuffle (excluding regions that were already annotated as genes by Ensembl or NCBI), and these coordinates were also compared to reference genes using Bedtools closest. The distance between predicted intergenic transcripts and the closest reference genes was compared to the distance between randomized coordinates and the closest reference genes with an independent 2-group Mann–Whitney *U*-test.

Inferring Biological Functions of Predicted Transcripts at Novel Loci

To interpret the function of predicted transcripts at novel loci, their sequences were compared against several databases. First, sequences were compared against the NT (NCBI non-redundant nucleotide, v5) database with BLASTN (v2.6.0), requiring a minimum e-value of 1e-10 for matches. Then, sequences were compared against the NR (NCBI non-redundant protein, v5) and SwissProt (downloaded from NCBI, v5) databases with Diamond BLASTX (v2.0.5.143), again setting the minimum *e*-value to 1e-10. For transcripts with SwissProt matches, the corresponding UniProt identifiers were associated with functional terms using DAVID (v6.8), including KEGG terms, GO "DIRECT" terms, and Clusters of Orthologous Groups of proteins (COG) ontology terms.

Predicted Transcript Expression Quantification

To determine the expression of predicted transcripts, reads were directly mapped to the predicted transcriptome. Predicted transcripts were converted from GTF to FASTA format with the gffread utility (v0.12.2). Strand-corrected full-length ONT reads (output of Pychopper) were then directly mapped to the predicted transcriptome using minimap2 (v2.16r922) with options "-t 10 -ax map-ont -p 0." Alignments with a minimum quality score of 10 were extracted with Samtools (v1.7). From these alignments, expression of predicted transcripts in transcripts per million (TPM) was determined with Nanocount (v2.3.0). For the identification of tissue-specific transcripts, samples with unclear identity were excluded. These samples included those that did not cluster with biological replicates (abomasum-F1, colon-F1, and lung-M1), tissues with unclear identity because samples did not cluster together (esophagus, skin and thyroid), and tissues with only a single replicate (duodenum-M1, hypothalamus-M1, and uterine endometrium-F1).

Identification and Characterization of Tissue-Specific Transcripts

The tissue specificity index (TSI) (Julien et al., 2012) for each transcript was calculated as follows, such that x_i was the average expression (TPM) in a given tissue, and n was the number of tissues:

$$TSI = \frac{\max_{1 \le i \le n} (x_i)}{\sum_{i=1}^{n} x_i}$$

Transcripts were then categorized as tissue-specific (TSI $\geq$ 0.8), broadly expressed (TSI $<$ 0.5), or biased toward a group of tissues (0.5 $\leq$ TSI $<$ 0.8). To interpret the biological significance of tissue-specific transcripts, those with corresponding Ensembl IDs were submitted to DAVID (v6.8) for functional enrichment analysis, considering only GO "DIRECT" terms. In each case, the top five most significant GO terms were reported (Benjamini-corrected p-value $<$ 0.05). Finally, to determine whether the TSS used by tissue-specific transcripts were uniquely active in that tissue, the coordinates of TSS ($\pm$50 bp) for tissue-specific transcripts for a given tissue were extracted and compared to the TSS ($\pm$50 bp) of every other predicted transcript using Bedtools (v2.26.0) intersect, with option "-s" to only consider same-strand overlap. The TSS from tissue-specific transcripts that did not overlap any other TSS from the remaining set of predicted transcripts were considered uniquely active in that tissue.

RESULTS

Total RNA was extracted from 93 biological samples and used to generate cDNA libraries, which were multiplexed and sequenced on a single PromethION flow cell. Samples consisted of 32 tissues collected from two male (M1, M2) and two female (F1, F2) adult Line 1 Hereford cattle. These animals were specifically chosen for their relation to Dominette, the individual sequenced for the original cattle reference genome. Sequencing yielded 53.7 million reads, with a read length N50 of 893, average read length of 759 bases, and average quality of 8.8 (**Supplementary Figure 1**). After demultiplexing, 35.3 million reads passed quality thresholds (greater than Q7), and further processing yielded 30.3 million full-length strand-oriented reads which were aligned to the ARS-UCD1.2 assembly, resulting in 25.5 million unique alignments that could be used for transcript prediction (**Supplementary Table 2**). On average, about 270,000 reads were obtained per

sample (**Supplementary Table 2**), and about 800,000 reads were obtained per tissue (**Supplementary Table 3**).

A preliminary evaluation of gene expression was conducted by counting alignments attributed to genes in the Ensembl (v101) annotation (**Supplementary Data 1**). Principal components analysis and hierarchical clustering of normalized gene expression generally clustered samples by tissue and organ system (**Figure 1**), with the exception of lung-M1, which was attributed extremely few reads, abomasum-M1 and colon-F1, which did not cluster with biological replicates, and esophagus, skin, and thyroid samples, which clustered ambiguously. In particular, male esophagus samples clustered with muscle, whereas female esophagus clustered with skin and stomach samples, suggesting potential sampling error during collection of male esophagus. Samples of questionable origin, based on aberrant clustering patterns, were excluded from tissue-specific analyses, but retained in the complete dataset for predicting transcript models. Brain and testis were among the most informative tissues, based on transcriptomic complexity and number of expressed loci (**Supplementary Figure 2A**).

Mapped reads from all samples were pooled to predict transcript models using the Pinfish pipeline. Briefly, transcripts were predicted from clusters of three or more alignments. Predicted transcripts were then polished and collapsed to filter out likely degradation products. In total, 244,945 transcript models were predicted, consisting of 76,110 multi-exon and 168,835 single-exon transcripts. Multi-exon transcripts localized to 23,694 loci, of which 13,053 (55%) corresponded to multiple transcripts. Comparing the predicted multi-exon transcripts to Ensembl and NCBI gene annotations revealed high precision,

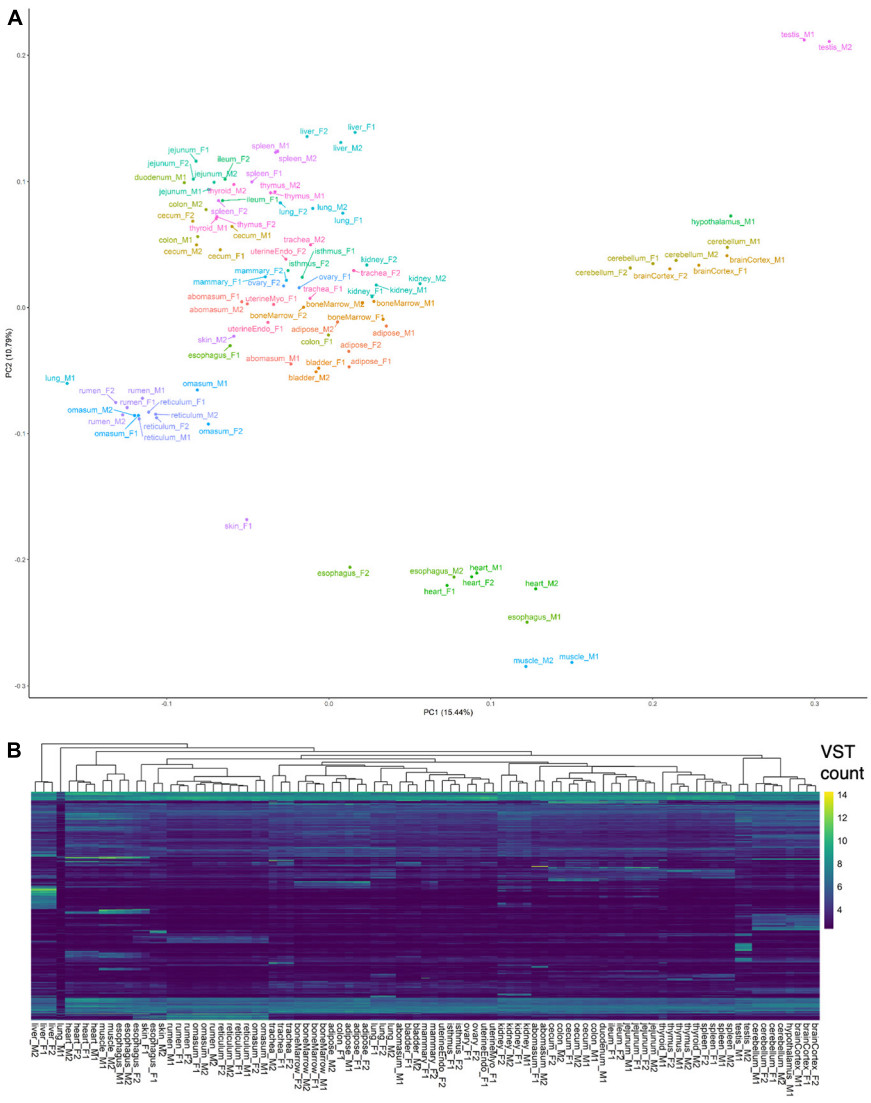


FIGURE 1 | Preliminary analysis of transcriptomes. **(A)** Principal components analysis of VST-normalized gene counts. **(B)** Hierarchical clustering of samples based on top 5,000 genes with highest variance in VST counts.

particularly at the base and intron levels, with most reference exons and introns captured by the predicted multi-exon transcripts (**Table 1**).

Compared to multi-exon transcripts, single-exon transcripts were supported by fewer reads ($p < 2.2e\text{-}16$; one-sided Z-test) (**Supplementary Figure 3A**), and tended to not directly overlap annotated exons, instead occurring predominantly within reference introns (**Supplementary Figure 3B**). Consequently, only single-exon transcripts that corresponded to annotated protein-coding genes, or those which were supported by more than 100 alignments (i.e., the top 1% most strongly supported single-exon transcripts) (**Supplementary Figure 4**), were retained in the final transcript set, which comprised 99,044 predicted transcripts (22,934 single-exon and 76,110 multi-exon transcripts) belonging to 31,824 genomic loci. Although only a small percentage of the retained single-exon transcripts were predicted to be coding (5%), the expression patterns of single-exon transcripts clearly distinguished brain tissues from the others (**Supplementary Figure 5**), suggesting these transcripts are biologically relevant. Expression of non-coding transcripts also distinguished brain, as well as testis, from other tissues (**Supplementary Figure 6**). Overall, transcript predictions accounted for 72% (15,716/21,861) of protein-coding genes in the Ensembl annotation and 78% (16,487/21,039) of protein-coding genes in the NCBI annotation.

Comparing the predicted transcript set to either the Ensembl or the NCBI annotations (**Supplementary Data 2, 3**) revealed that most predicted transcripts either exactly matched a reference transcript exon-by-exon, or demonstrated some same strand overlap with reference exons (**Figure 2A**). In all, 21% of predicted transcripts exactly matched a reference transcript from either Ensembl or NCBI, 61% were considered novel isoforms of reference genes based on same strand overlap of reference exon(s), 6% did not correspond to a reference gene and were considered novel loci, and 12% were classified as potential artifacts, possibly due to mapping error, pre-mRNA fragments, or polymerase run-on.

Considering the largest class of predicted transcripts were novel isoforms of known genes, we then sought to quantify the extent to which variation in transcription start sites, end sites, alternative splicing, and alternative polyadenylation sites contributed to transcriptome complexity. Transcript degradation, especially at the 5′ end, is certainly a concern in long-read transcriptomics, although the 3′ ends are considered to be more reliable. The Pinfish pipeline used to predict transcripts tries to take this limitation into account by collapsing transcripts with similar exon structure and variable 5′ ends, within a 5,000 bp 5′ exon boundary tolerance. Considering all 5′ ends of predicted transcripts (±100 bp), we found that 28% overlapped 5′ ends of Ensembl or RefSeq transcripts (±100 bp), and 45% overlapped TSS (±100 bp) identified by the 5′-complete sequencing technique RAMPAGE (Goszczynski et al., 2020). Even when predicted 5′ ends did not directly coincide with Ensembl, RefSeq or RAMPAGE annotations (**Supplementary Data 4**), they still preferentially occurred in the vicinity of RAMPAGE TSS (39% of these 5′ ends occurred within 1kb of RAMPAGE TSS) and were not biased downstream of RAMPAGE TSS (**Supplementary Figure 7**), which would have been characteristic of degradation. Most novel isoforms began within 2 kb of the reference transcription start site (51%, 28,289 transcripts) and terminated within 2 kb of the reference transcription end site (58%, 31,913 transcripts) (**Supplementary Figure 8**). Additional variation was present at TES, as alternative polyadenylation sites were detected for 30% of reference Ensembl transcripts (5,821/19,613 transcripts) (**Supplementary Figure 9**).

The main source of transcriptional variation resulted from alternative splicing (**Figure 2B**). Alternative first exons were common in predicted multi-exon transcripts, reflecting the use of alternative promoters in different regulatory contexts. This phenomenon was clearly reflected at the *RSPH9* locus, which encodes a component of motile flagella and is associated with multiple transcript variants from alternative splicing in humans, although only a single isoform had been annotated in cattle (**Figure 2C**). Besides the alternative splicing evident at this locus, three different transcription start sites were utilized, resulting in ten isoforms, several of which demonstrated tissue-specific expression patterns (**Supplementary Figure 10**). In a given tissue sample, 10,844 ± 2,010 (S.D.) loci were expressed

TABLE 1 | Sensitivity and precision estimates of predicted multi-exon transcripts compared to reference multi-exon transcripts from the Ensembl (v101) and NCBI (release 106) annotations.

	Predicted vs. Ensembl		Predicted vs. NCBI		NCBI vs. Ensembl	
Level	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
Base	58.5	72.5	49.6	81.2	88.7	66.3
Exon	54.8	58.3	55.7	66.5	81.3	73.2
Intron	60.1	80.7	57.3	86.5	90.0	79.4
Transcript	29.3	12.9	24.6	20.2	48.8	26.7
Locus	52.6	47.9	62.7	56.8	75.6	76.9
Missed exons	53,069/171,341 (31.0%)		65,592/207,468 (31.6%)		8,891/222,022 (4.0%)	
Novel exons	15,130/193,597 (7.8%)		8,366/203,236 (4.1%)		27,725/257,826 (10.8%)	
Missed introns	48,381/151,779 (31.9%)		55,075/177,905 (31.0%)		3,643/195,870 (1.9%)	
Novel introns	4,528/112,911 (4.0%)		2,447/117,961 (2.1%)		12,338/222,064 (5.6%)	

Comparison excludes reference loci without predicted transcripts and predicted transcripts at novel loci.

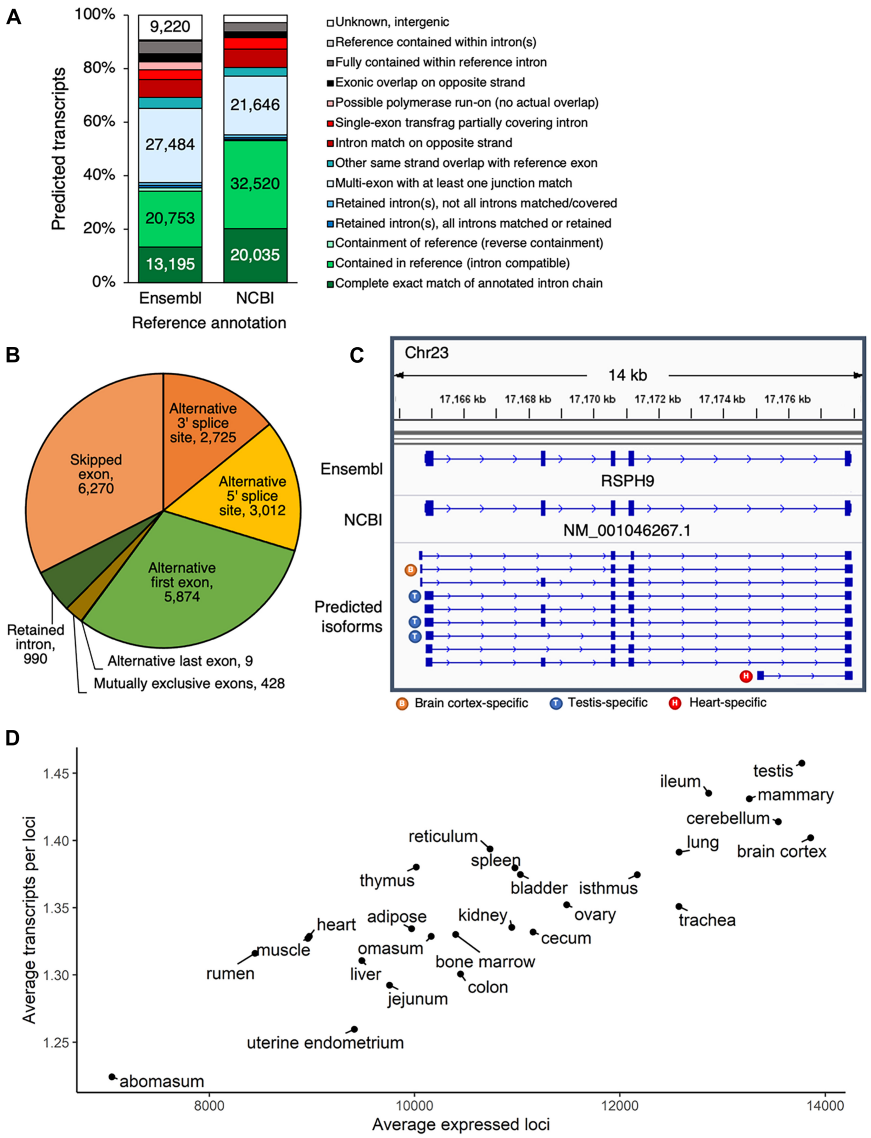


FIGURE 2 | Predicted transcripts capture transcriptome complexity. **(A)** Comparison of predicted isoforms to Ensembl and NCBI gene annotations. **(B)** Frequency of alternative splicing events in predicted multi-exon transcript isoforms. **(C)** Predicted isoforms at the RSPH9 locus, which is thought to code for a component of motile cilia and flagella. In humans, multiple splicing is known to produce transcript variants, but only one transcript had been annotated in cattle, according to both the Ensembl and NCBI annotations. **(D)** Based on the predicted transcript set, number of expressed loci and ratio of expressed transcripts per loci, averaged per tissue.

with 1.35 ± 0.06 (S.D.) predicted isoforms expressed per locus. Testis was the most informative tissue, with the most expressed loci and highest ratio of expressed transcripts per gene, whereas abomasum demonstrated the lowest transcriptomic complexity (**Figure 2D**).

Given the large number of sampled tissues, tissue-specific isoforms could be identified from this dataset with high resolution. Tissue-specific transcripts are fundamental to understanding the basis of biological differences between tissues, and can serve as useful biomarkers (Stutterheim et al., 2008; Prensner et al., 2013), as they are often implicated in tissue-specific functions, development, and disease (Leucci et al., 2016). To identify tissue-specific isoforms, the tissue-specificity index (TSI) was calculated from the average expression of predicted transcripts (transcripts per million; TPM) in each tissue with at least two high-confidence biological replicates (adipose, bladder, bone marrow, brain cortex, cecum, cerebellum, colon, heart, ileum, isthmus, jejunum, kidney, liver, lung, mammary gland, muscle, omasum, ovary, reticulum, rumen, spleen, testis, thymus, trachea, and uterine endometrium) (**Supplementary Data 5**). For a given transcript, the TSI varies between 0 (uniformly expressed across all tissues) and 1 (uniquely expressed in a single tissue). Transcripts that were only expressed in a single sample were excluded from the tissue-specificity analysis.

Overall, the TSI demonstrated a bimodal distribution, with most transcripts either broadly (TSI closer to zero)

or specifically (TSI closer to 1) expressed (**Figure 3A**). This pattern was observed for both single- and multi-exon transcripts (**Supplementary Figure 11A**). The TSI was closely linked to the average expression across samples, with highly expressed transcripts (average TPM $\geq$ 10) more often generally expressed across many tissues, whereas moderately- (1 $\leq$ average TPM $<$ 10) and lowly expressed transcripts (average TPM $<$ 1) tended to be more tissue-specific (**Figure 3B**). Overall, 48,867 transcripts (74%) were widely expressed (TSI $<$ 0.5), 7,066 transcripts (11%) were highly tissue-specific (TSI $\geq$ 0.8), and 10,203 transcripts (15%) demonstrated expression in a small subset of tissues (0.5 $\leq$ TSI $<$ 0.8). Interestingly, compared with multi-exon transcripts, single-exon transcripts were more likely to be brain-specific (**Supplementary Figure 11B**), and were generally predicted to be non-coding (95%), which is consistent with the central role of non-coding RNA in the brain (Guennewig and Cooper, 2014). Transcripts with intermediate TSI scores likely includes isoforms specific to higher-order structures from which multiple tissues were sampled (e.g., brain, pre-stomach, gastrointestinal tract), or tissues of similar embryonic origin (e.g., ectodermal, mesodermal, endodermal) as has been observed by previous transcriptomic studies in the pig (Perez-Montarelo et al., 2012).

An overwhelming proportion of tissue-specific transcripts (61%) were attributed to testis, and most of these were either novel isoforms (49%) or novel loci (20%) (**Figure 3C**). More than

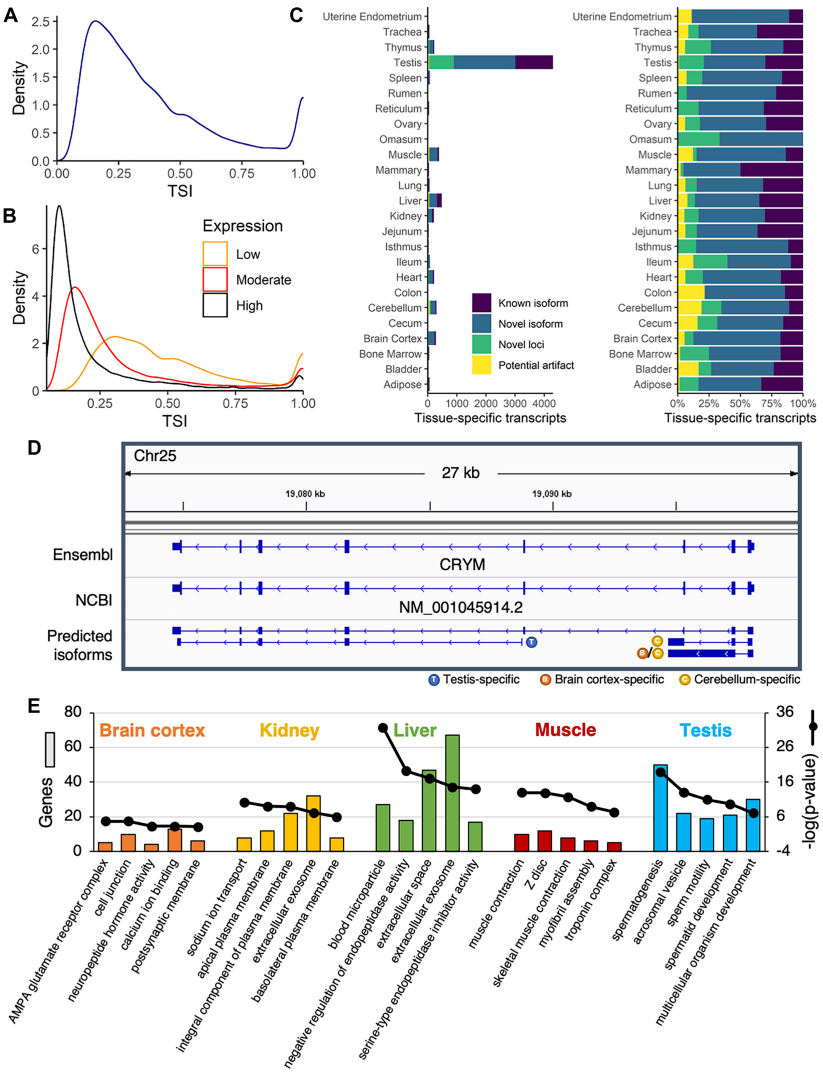


FIGURE 3 | Identification of tissue-specific isoforms. **(A)** Density plot of the tissue-specificity index (TSI) identified for each predicted transcript, based on average transcripts per million (TPM) in each tissue. **(B)** Density plot of TSI for predicted transcripts with low (average TPM $<$ 1), moderate (1 $\leq$ average TPM $<$ 10), or high expression (average TPM $\geq$ 10). **(C)** Number of tissue-specific transcripts (TSI $\geq$ 0.8) attributed to each tissue, categorized as known or novel isoforms, novel loci, or potential artifacts. **(D)** The annotated transcript at the CRYM locus was expressed across a range of tissues, whereas novel isoforms were either testis- or brain-specific. **(E)** Functional enrichment of genes corresponding to tissue-specific isoforms in brain cortex, kidney, liver, muscle, and testis. Top five most significant gene ontology terms reported (Benjamini corrected *p*-value $<$ 0.05).

80% of the transcription start sites used by testis-specific isoforms were only active in testis (**Supplementary Figure 12**), suggesting pervasive use of alternative promoters in this tissue. This alternative promoter usage was evident at the *CRYM* locus, with a novel testis-specific isoform beginning at the third annotated exon (**Figure 3D**). The remaining novel *CRYM* isoforms were brain-specific, whereas the sole annotated transcript variant was broadly expressed across tissues. This locus illustrated a broader pattern: novel isoforms of annotated genes were expressed in fewer tissues and at lower levels than previously annotated isoforms ($p < 2.2e\text{-}16$; Welch two sample *t*-test) (**Supplementary Figure 13**), suggesting that the reference genome annotations failed to capture rare isoforms with potentially significant biological functions. Indeed, genes with tissue-specific isoforms were strongly biased toward tissue-specific functions (**Figure 3E**).

To gain some insight into the potential biological functions of isoforms at novel loci, transcript sequences were compared against several BLAST databases (**Supplementary Data 6**). Strong matches (*E*-value < 1e-10) were identified for 93% (5,944/6,370) of transcripts at novel loci when comparing against the NT database (NCBI non-redundant nucleotide sequences), 42% (2,678/6,370) against the NR database (NCBI non-redundant protein sequences), and 12% (794/6,370) against the SwissProt database (curated protein sequences). Based on gene ontology (GO) terms and KEGG pathways associated with SwissProt identifiers, transcripts at novel loci are involved in a variety of biological functions, such as lysine degradation, cAMP signaling, and phosphodiester bond hydrolysis (**Figure 4A**). Of note, two of the top ten most common biological process GO terms were related to RNA-mediated transposition, indicating that some novel transcripts could correspond to transposons that have not been completely silenced.

The genomic distribution of novel loci was biased toward contigs; whereas only 0.4% of all predicted transcripts (342/99,044) localized to contigs, 7.7% of transcripts at novel intergenic sites (126/1,628) were on contigs. Nevertheless, novel intergenic transcripts preferentially occurred closer to annotated genes (on average 60 kb away from an Ensembl transcript) than would be expected by random chance (on average 140kb away from an Ensembl transcript) ($p < 2.2e\text{-}16$; Independent two-group Mann–Whitney *U*-test) (**Supplementary Figure 14**). Transcripts at novel loci tended to be shorter than those of annotated genes with fewer exons (**Supplementary Figure 15**), despite the exclusion of most intergenic single-exon predicted transcripts.

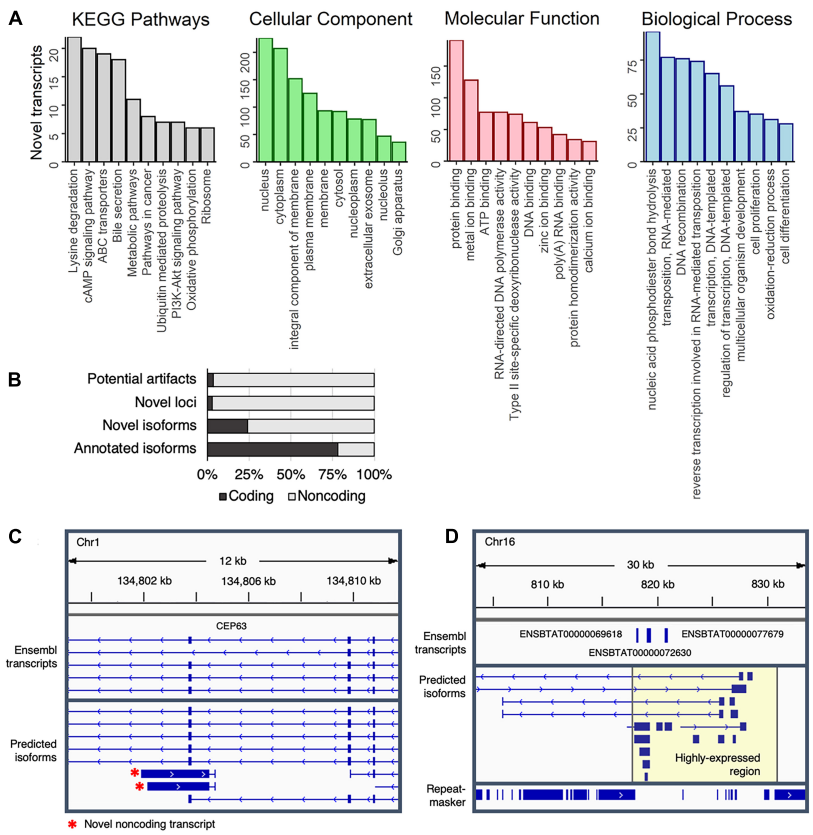


FIGURE 4 | Characterization of predicted transcripts at novel loci. **(A)** The top ten represented KEGG pathways and GO terms (separated into Cellular Component, Molecular Function, and Biological Process terms) represented in transcripts at novel loci that corresponded to a UniProt identifier. **(B)** Coding potential of predicted transcripts. **(C)** Novel non-coding antisense transcript at the CEP63 locus. **(D)** Highly expressed section of chromosome 16. RepeatMasker track shows repetitive elements, which were depleted in the highly expressed region (highlighted in yellow).

Nearly all predicted transcripts at novel loci appeared to be non-coding (**Figure 4B**), which could partially explain the lower number of matches in protein-based databases (NR and SwissProt) as compared to the nucleotide-based database (NT). For instance, transcription of the anti-sense strand at the *CEP63* locus – a centrosomal protein crucial for division of brain cells – produces short (∼2 kb long) non-coding transcripts (**Figure 4C**) that are expressed in a mutually exclusive pattern with the main *CEP63* isoform (**Supplementary Figure 16**), potentially suggesting that *CEP63* expression is regulated by a previously unannotated antisense non-coding RNA.

Of note, more than 1.5 million reads (6.5% of the entire dataset), were aligned to a single 15 kb region on chromosome 16 (**Figure 4D**). Surprisingly, this region contained no RefSeq transcripts, although the Ensembl annotation included three single-exon transcripts that were predicted to code for NADH hydrogenase and ATP synthase subunits. Considering this region was strongly expressed across all samples (**Supplementary Figure 17**), these transcripts likely serve fundamental biological roles that remain to be established. Additionally, because gene expression is generally normalized based only on reads that align to the exome, the inclusion of these loci in future annotations could improve estimates of gene expression in transcriptomic-based studies.

DISCUSSION

Although long-read sequencing has been extensively implemented for the study of transcription dynamics, resulting datasets have generally either been limited by sample size or sequencing depth. To address this limitation, here we demonstrate that by coupling ONT sequencing with large-scale multiplexing, we were able to profile the full-length transcriptomes of 32 adult bovine tissues from a single ONT flow cell. Of the nearly 100,000 predicted transcripts, over 60% were novel isoforms of reference genes, indicating that the complexity of the bovine transcriptome is comparable to what has been described in humans. Moreover, this high percentage of novel isoforms is consistent with other studies that have used long-read sequencing to improve annotations in pigs (80% of identified transcripts were novel), rabbits (66%), and cattle (60%) (Chen et al., 2017; Beiki et al., 2019; Rosen et al., 2020). Compared to previous efforts to annotate full-length bovine transcripts (Rosen et al., 2020), this study leveraged a single ONT flow cell to interrogate more tissues (32 versus 23) from multiple individuals (four replicates versus one) at a greater sequencing depth (25 million versus 553,798 reads). In terms of cost, speed, and throughput, these comparisons highlight the power of this method for transcriptome annotation.

Overall, our transcript predictions substantially increased the ratio of isoform variants per reference bovine gene from 1.59 to 3.57 (74,312 transcripts at 20,811 reference Ensembl loci), which is consistent with the ratio observed in humans (3.78 transcripts per reference Ensembl locus) (**Supplementary Figure 18**). Although not all of the 5′ ends of predicted transcripts directly overlapped Ensembl, RefSeq, or RAMPAGE TSS, the corresponding transcripts (**Supplementary Data 4**) were not disregarded. Just as this study cannot provide a comprehensive catalog of full-length bovine transcripts, analysis of RAMPAGE data may have missed credible TSS. Further efforts to annotate regulatory elements in bovine tissues (i.e., by profiling chromatin accessibility and histone modifications) should help to further refine the 5′ ends of transcript models; however, these data are not yet available for all tissues.

Notably, this study only profiled samples from a single breed – Hereford – which was specifically chosen because it is also the basis for the current bovine genome assembly. Consequently, these data cannot account for the substantial phenotypic and genetic variation observed between different breeds and subspecies of cattle (Weigel et al., 2017). For instance, taurine breeds are known to have higher fertility than indicine breeds, whereas indicine breeds demonstrate higher resistance to disease and parasites and thrive in hotter climates. Although a recent study reported identification of haplotype-specific transcripts by PacBio sequencing, the dataset was limited to seven tissues from a *Bos Taurus* hybrid fetus (Low et al., 2020). Moving forward, it will be of considerable scientific and economic interest to continue investigating breed-specific transcriptomes, with the goal of better understanding the biological mechanisms that underpin phenotypic differences between animals.

Although this study interrogated over 30 adult tissues, the resulting annotation is still far from exhaustive. Transcription was only detected at about 60% of reference loci; the remainder may not have been expressed in the sampled tissues, or may have been expressed at such a low level that expression was not detected due to lower sequencing depth per sample. On average, we found each tissue expressed about 10,000 loci, although some tissues – specifically brain and testis – demonstrated substantially more complex transcriptomes. On the other hand, nearly 30% of all reads attributed to abomasum samples originated from *LYZ2* (**Supplementary Data 7**), part of the lysozyme *c* family of digestive proteins that play an important role in ruminant digestion (Irwin, 2015). Such highly abundant transcripts can be problematic for transcriptomic studies, as they make it harder to detect rare transcripts with potential biological significance. This problem is intensified for long-read sequencing methods, which generally have lower throughput; however, it is possible to specifically target such transcripts by hybridization, for example by the CRISPR-Cas9 based method DASH (depletion of abundant sequences by hybridization) (Gu et al., 2016) which was recently employed to deplete hemoglobin transcripts prior to ONT sequencing of polar bear blood (Byrne et al., 2019). To gain a more complete picture of transcription in bovine abomasum or blood – the latter of which was not profiled in this study – it will likely be necessary to deplete abundant transcripts, such as lysozyme and hemoglobin, in order to detect rarer isoforms.

Another potential limitation of our approach is that it was based on cDNA, the generation of which is inherently limited by the capacity of reverse transcriptase to amplify long transcripts. As a result, it was difficult to capture full-length transcripts for some of the longest genes, such as titin (*TTN*), which also tended to produce fragmented Iso-seq reads (Rosen et al., 2020).

In addition to fragment length limitations, PCR amplification can also introduce substantial GC-content bias into libraries (Mamanova et al., 2010), altering transcript abundance and library complexity. Furthermore, by using oligo-dT primers for cDNA generation, as opposed to random primers, our transcript predictions are likely biased against RNAs that are generally not polyadenylated (e.g., non-coding RNAs). Single-molecule sequencing platforms, such as the MinION (Garalde et al., 2018), avoid these PCR biases altogether by reading native RNA nucleotides directly as they pass through a nanoscale sensor. Moreover, direct RNA sequencing can identify post-transcriptional events like ribonucleotide modifications, which are increasingly recognized as key regulators of several biological processes (Jantsch et al., 2018).

Nevertheless, native RNA long-read sequencing is somewhat limited by throughput and transcript truncation. A single MinION flow cell produces only about half a million aligned reads (Soneson et al., 2019), as compared to the 30 million aligned reads generated by this study from a single PromethION flow cell. In addition, a significant portion of native RNA reads are truncated during Nanopore direct RNA sequencing, especially the last 10–15 nucleotides at the 5′ end (Soneson et al., 2019; Workman et al., 2019). In theory, this issue could be resolved by filtering out ONT reads that do not begin within defined promoters, which were recently experimentally determined in cattle (Goszczynski et al., 2020), but this approach would undoubtedly reduce the quantitative nature of the data.

These limitations notwithstanding, as long-read sequencing technologies continue to improve, both native RNA and single-cell ONT strategies are likely to become increasingly accurate, informative and practical, providing unprecedented insight into transcriptome complexity and cell-to-cell heterogeneity (Lebrigand et al., 2020). In fact, recent efforts to computationally correct sequencing errors in ONT data are capable of reducing the error rate from 14% (Workman et al., 2019) to about 1% (Sahlin et al., 2020), such that it should be possible for future studies to use ONT sequencing for reference-free *de novo* transcriptome analysis.

As it stands, we have demonstrated the potential for multiplexing paired with ONT sequencing as a powerful and accessible technique for isoform identification and expression profiling. Nevertheless, to comprehensively capture the transcriptomic complexity of the bovine genome, future studies will need to continue to characterize transcript isoforms in a broader range of tissues and cell types, representing different developmental stages, disease states, and physiological conditions. The ability to identify full-length transcripts from nearly one hundred samples using a single ONT flow cell makes the task of exhaustively annotating a mammalian transcriptome significantly more feasible.

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: https://www.ncbi.nlm.nih.gov/geo/, GSE160028.

ETHICS STATEMENT

The animal study was reviewed and approved by the University of California, Davis Institutional Animal Care and Use Committee.

AUTHOR CONTRIBUTIONS

PR designed the study. JM and HZ contributed to the experimental design. AI-T performed the RNA extractions and generated libraries for sequencing. MH and DG performed the bioinformatics analyses. MH and PR wrote the manuscript. All the authors have read and approved the manuscript.

FUNDING

Funding for sample collection was provided by the United States Department of Agriculture, National Institute of Food and Agriculture, Agriculture and Food Research Initiative grant no. 2015-43567015-22940 awarded to HZ and PR. Funding for library generation, sequencing, and bioinformatics analyses was provided by the United States Department of Agriculture, National Institute of Food and Agriculture, Agriculture and Food Research Initiative grant no. 2017-67015-26297 awarded to PR.

ACKNOWLEDGMENTS

We thank Dr. Ying Wang, Dr. Perot Saelao, and Ganrea Chanthavixay for their assistance in sample collection, organization, and storage. We also thank the staff at the University of California DNA Technologies Core for their guidance and assistance with Nanopore library construction and sequencing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.664260/full#supplementary-material

REFERENCES

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project. *Genome Biol.* 16:57.

Arefeen, A., Liu, J., Xiao, X., and Jiang, T. (2018). TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics* 34, 2521–2529. doi: 10.1093/bioinformatics/bty110

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., et al. (2012). The evolutionary landscape of alternative splicing

in vertebrate species. *Science* 21, 1587–1593. doi: 10.1126/science.1230612

Bayega, A., Fahiminiya, S., Oikonomopoulos, S., and Ragoussis, J. (2018). Current and future methods for mRNA analysis: a drive toward single molecule sequencing. *Methods Mol Biol.* 1783, 209–241. doi: 10.1007/978-1-4939-7834-2_11

Beiki, H., Liu, H., Huang, J., Manchanda, N., Nonneman, D., Smith, T. P. L., et al. (2019). Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. *BMC Genomics* 20:344. doi: 10.1186/s12864-019-5709-y

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., et al. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8:16027.

Byrne, A., Supple, M. A., Volden, R., Laidre, K. L., Shapiro, B., and Vollmers, C. (2019). Depletion of hemoglobin transcripts and long-read sequencing improves the transcriptome annotation of the polar bear (*Ursus maritimus*). *Front. Genet.* 10:643. doi: 10.3389/fgene.2019.00643

Chen, S.-Y., Deng, F., Jia, X., Li, C., and Lai, S.-J. (2017). A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Sci. Rep.* 7:7648.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13.

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15:201. doi: 10.1038/nmeth.4577

García-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-López, F. J., and Van Tassell, C. P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc. Natl. Acad. Sci. U S A.* 113, E3995–E4004.

Georges, M., Charlier, C., and Hayes, B. (2019). Harnessing genomic information for livestock improvement. *Nat. Rev. Genet.* 20, 135–156. doi: 10.1038/s41576-018-0082-2

Goszczynski, D. E., Halstead, M. M., Islas-Trejo, A. D., Zhou, H., and Ross, P. J. (2020). Transcription initiation mapping in 31 bovine tissues reveals complex promoter activity, pervasive transcription, and tissue-specific promoter usage. *bioRxiv [preprint]* doi: 10.1101/2020.09.05.284547

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29:644. doi: 10.1038/nbt.1883

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107. doi: 10.1016/s0168-9525(00)02176-4

Gu, W., Crawford, O'Donovan, B. D., Wilson, M. R., Chow, E. D., Retallack, H., et al. (2016). Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* 17:41.

Guennewig, B., and Cooper, A. A. (2014). The central role of noncoding RNA in the brain. *Int. Rev. Neurobiol.* 116, 153–194. doi: 10.1016/b978-0-12-801105-8.00007-2

Harr, B., and Turner, L. M. (2010). Genome-wide analysis of alternative splicing evolution among Mus subspecies. *Mol. Ecol.* 19, 228–239. doi: 10.1111/j.1365-294x.2009.04490.x

Ip, C. L. C., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., et al. (2015). MinION analysis and reference consortium: phase 1 data release and analysis. *F1000Research* 4:1075. doi: 10.12688/f1000research.7201.1

Irwin, D. M. (2015). Genomic organization and evolution of ruminant lysozyme c genes. *Zool. Res.* 36:1.

Jantsch, M. F., Quattrone, A., O'Connell, M., Helm, M., Frye, M., Macias-Gonzales, M., et al. (2018). Positioning Europe for the EPITRANSCRIPTOMICS challenge. *RNA Biol.* 15, 829–831.

Jenjaroenpun, P., Wongsurawat, T., Pereira, R., Patumcharoenpol, P., Ussery, D. W., Nielsen, J., et al. (2018). Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of Saccharomyces cerevisiae CEN. PK113-7D. *Nucleic Acids Res.* 46:e38. doi: 10.1093/nar/gky014

Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schütz, F., et al. (2012). Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol.* 10:e1001328. doi: 10.1371/journal.pbio.1001328

Kadobianskyi, M., Schulze, L., Schuelke, M., and Judkewitz, B. (2019). Hybrid genome assembly and annotation of *Danionella trans*lucida. *Sci Data* 6:156.

Kalsotra, A., Xiao, X., Ward, A. J., Castle, J. C., Johnson, J. M., Burge, C. B., et al. (2008). A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proc Natl Acad Sci. U S A.* 105, 20333–20338. doi: 10.1073/pnas.0809045105

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* 11, 345–355. doi: 10.1038/nrg2776

Kuo, R. I., Tseng, E., Eory, L., Paton, I. R., Archibald, A. L., and Burt, D. W. (2017). Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* 18:323. doi: 10.1186/s12864-017-3691-9

Lebrigand, K., Magnone, V., Barbry, P., and Waldmann, R. (2020). High throughput error corrected nanopore single cell transcriptome sequencing. *Nat. Commun.* 11:4025.

Leucci, E., Vendramin, R., Spinazzi, M., Laurette, P., Fiers, M., Wouters, J., et al. (2016). Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* 531, 518–522.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, Y., Fang, C., Fu, Y., Hu, A., Li, C., Zou, C., et al. (2018). A survey of transcriptome complexity in *Sus scrofa* using single-molecule long-read sequencing. *DNA Res.* 25, 421–437. doi: 10.1093/dnares/dsy014

Linker, S. M., Urban, L., Clark, S. J., Chhatriwala, M., Amatya, S., McCarthy, D. J., et al. (2019). Combined single-cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity. *Genome Biol.* 20:30.

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, 1–21.

Low, W. Y., Tearle, R., Liu, R., Koren, S., Rhie, A., Bickhart, D. M., et al. (2020). Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and *Brahman cattle*. *Nat. Commun.* 11:2071.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., et al. (2010). Target-enrichment strategies for next-generation sequencing. *Nat. Methods* 7, 111–118.

McCarthy, A. (2010). Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chem. Biol.* 17, 675–676. doi: 10.1016/j.chembiol.2010.07.004

Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., et al. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* 30, 99–104. doi: 10.1038/nbt.2024

Meredith, B. K., Kearney, F. J., Finlay, E. K., Bradley, D. G., Fahey, A. G., Berry, D. P., et al. (2012). Genome-wide associations for milk production and somatic cell score in holstein-friesian cattle in Ireland. *BMC Genet.* 13:21. doi: 10.1186/1471-2156-13-21

Mudge, J. M., Frankish, A., Fernandez-Banet, J., Alioto, T., Derrien, T., Howald, C., et al. (2011). The origins, evolution, and functional potential of alternative splicing in vertebrates. *Mol. Biol. Evol.* 28, 2949–2959. doi: 10.1093/molbev/msr127

Müller, T., Boileau, E., Talyan, S., Kehr, D., Varadi, K., Busch, M., et al. (2020). Updated and enhanced pig cardiac transcriptome based on long-read RNA sequencing and proteomics. *J. Mol. Cell Cardiol.* 150, 23–31. doi: 10.1016/j.yjmcc.2020.10.005

Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D., and Ragoussis, J. (2016). Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* 6:31602.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259

Paronetto, M. P., Passacantilli, I., and Sette, C. (2016). Alternative splicing and cell survival: from tissue homeostasis to disease. *Cell Death Differ.* 23, 1919–1929. doi: 10.1038/cdd.2016.91

Perez-Montarelo, D., Hudson, N. J., Fernandez, A. I., Ramayo-Caldas, Y., Dalrymple, B. P., and Reverter, A. (2012). Porcine tissue-specific regulatory networks derived from meta-analysis of the transcriptome. *PLoS One* 7:e46159. doi: 10.1371/journal.pone.0046159

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a

transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Prensner, J. R., Iyer, M. K., Sahu, A., Asangani, I. A., Cao, Q., Patel, L., et al. (2013). The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genet.* 45, 1392–1398. doi: 10.1038/ng.2771

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genom. Proteom. Bioinform.* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Robinson, T. P., Wint, G. R. W., Conchedda, G., Van Boeckel, T. P., Ercoli, V., Palamara, E., et al. (2014). Mapping the global distribution of livestock. *PLoS One* 9:e96084. doi: 10.1371/journal.pone.0096084

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., et al. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* 9:giaa021.

Saatchi, M., Schnabel, R. D., Rolf, M. M., Taylor, J. F., and Garrick, D. J. (2012). Accuracy of direct genomic breeding values for nationally evaluated traits in US Limousin and Simmental beef cattle. *Genet. Sel Evol.* 44:38. doi: 10.1186/1297-9686-44-38

Sahlin, K., Sipos, B., James, P. L., and Medvedev, P. (2020). Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nat. Commun.* 12:2.

Sahoo, B., Guttula, P. K., and Gupta, M. K. (2020). Comparison of spermatozoal RNA extraction methods in goats. *Anal. Biochem.* 614:114059. doi: 10.1016/j.ab.2020.114059

Seki, M., Katsumata, E., Suzuki, A., Sereewattanawoot, S., Sakamoto, Y., Mizushima-Sugano, J., et al. (2019). Evaluation and application of RNA-Seq by MinION. *DNA Res.* 26, 55–65. doi: 10.1093/dnares/dsy038

Sessegolo, C., Cruaud, C., Da Silva, C., Cologne, A., Dubarry, M., Derrien, T., et al. (2019). Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Sci. Rep.* 9:14908.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–1014. doi: 10.1038/nbt.2705

Soneson, C., Yao, Y., Bratus-Neuenschwander, A., Patrignani, A., Robinson, M. D., and Hussain, S. (2019). A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* 10:3359.

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Akerman, M., Alioto, T., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184. doi: 10.1038/nmeth.2714

Stutterheim, J., Gerritsen, A., Zappeij-Kannegieter, L., Kleijn, I., Dee, R., Hooft, L., et al. (2008). PHOX2B is a novel and specific marker for minimal residual disease testing in neuroblastoma. *J. Clin. Oncol.* 26, 5443–5449. doi: 10.1200/jco.2007.13.6531

Thomas, S., Underwood, J. G., Tseng, E., and Holloway, A. K. (2014). Long-read sequencing of chicken transcripts and identification of new transcript isoforms. *PLoS One* 9:e94650. doi: 10.1371/journal.pone.0094650

Thompson-Crispi, K., Atalla, H., Miglior, F., and Mallard, B. A. (2014). Bovine mastitis: frontiers in immunogenetics. *Front. Immunol.* 5:493. doi: 10.3389/fimmu.2014.00493

Tilgner, H., Grubert, F., Sharon, D., and Snyder, M. P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. U S A.* 111, 9869–9874. doi: 10.1073/pnas.1400447111

Tong, X., and Liu, S. (2019). CPPred: coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Res.* 47:e43. doi: 10.1093/nar/gkz087

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Ungaro, A., Pech, N., Martin, J.-F., McCairns, R. J. S., Mévy, J.-P., Chappaz, R., et al. (2017). Challenges and advances for transcriptome assembly in non-model species. *PLoS One* 12:e0185020. doi: 10.1371/journal.pone.0185020

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., Gonzalez-Vallinas, J., Lahens, N. F., Hogenesch, J. B., et al. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* 5:e11752.

Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L., and Swarbreck, D. (2018). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* 7:giy093.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi: 10.1038/nature07509

Weigel, K. A., VanRaden, P. M., Norman, H. D., and Grosu, H. (2017). A 100-Year Review: methods and impact of genetic selection in dairy cattle—from daughter–dam comparisons to deep learning algorithms. *J. Dairy Sci.* 100, 10234–10250. doi: 10.3168/jds.2017-12954

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., et al. (2019). Nanopore native RNA sequencing of a human poly (A) transcriptome. *Nat. Methods* 16, 1297–1305.

Zhang, H., Brown, R. L., Wei, Y., Zhao, P., Liu, S., Liu, X., et al. (2019). CD44 splice isoform switching determines breast cancer stem cell state. *Genes Dev.* 33, 166–179. doi: 10.1101/gad.319889.118

Zhang, X., Chen, M. H., Wu, X., Kodani, A., Fan, J., Doan, R., et al. (2016). Cell-type-specific alternative splicing governs cell fate in the developing cerebral cortex. *Cell* 166, 1147–1162. doi: 10.1016/j.cell.2016.07.025

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.

Copyright © 2021 Halstead, Islas-Trejo, Goszczynski, Medrano, Zhou and Ross. 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 May 2021
doi: 10.3389/fgene.2021.628849

Characterizing Genetic Regulatory Elements in Ovine Tissues

Kimberly M. Davenport[1], Alisha T. Massa[2], Suraj Bhattarai[3], Stephanie D. McKay[3], Michelle R. Mousel[4,5], Maria K. Herndon[2], Stephen N. White[2,4,6], Noelle E. Cockett[7], Timothy P. L. Smith[8] and Brenda M. Murdoch[1,6]* on behalf of The Ovine FAANG Project Consortium*

[1] Department of Animal, Veterinary, and Food Science, University of Idaho, Moscow, ID, United States, [2] Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, WA, United States, [3] University of Vermont, Burlington, VT, United States, [4] USDA, ARS, Animal Disease Research Unit, Pullman, WA, United States, [5] Paul G. Allen School for Global Animal Health, Washington State University, Pullman, WA, United States, [6] Center for Reproductive Biology, Washington State University, Pullman, WA, United States, [7] Utah State University, Logan, UT, United States, [8] USDA, ARS, U.S. Meat Animal Research Center (USMARC), Clay Center, NE, United States

OPEN ACCESS

Edited by:
Hans Cheng,
U.S. National Poultry Research Center, Agricultural Research Service (USDA), United States

Reviewed by:
Huaijun Zhou,
University of California, Davis, United States
Cong-jun Li,
Animal Genomics and Improvement Laboratory, Agricultural Research Service (USDA), United States

***Correspondence:**
Timothy P. L. Smith
tim.smith2@usda.gov
Brenda M. Murdoch
bmurdoch@uidaho.edu

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

Received: 13 November 2020
Accepted: 25 March 2021
Published: 20 May 2021

Citation:
Davenport KM, Massa AT, Bhattarai S, McKay SD, Mousel MR, Herndon MK, White SN, Cockett NE, Smith TPL and Murdoch BM (2021) Characterizing Genetic Regulatory Elements in Ovine Tissues. Front. Genet. 12:628849. doi: 10.3389/fgene.2021.628849

The Ovine Functional Annotation of Animal Genomes (FAANG) project, part of the broader livestock species FAANG initiative, aims to identify and characterize gene regulatory elements in domestic sheep. Regulatory element annotation is essential for identifying genetic variants that affect health and production traits in this important agricultural species, as greater than 90% of variants underlying genetic effects are estimated to lie outside of transcribed regions. Histone modifications that distinguish active or repressed chromatin states, CTCF binding, and DNA methylation were used to characterize regulatory elements in liver, spleen, and cerebellum tissues from four yearling sheep. Chromatin immunoprecipitation with sequencing (ChIP-seq) was performed for H3K4me3, H3K27ac, H3K4me1, H3K27me3, and CTCF. Nine chromatin states including active promoters, active enhancers, poised enhancers, repressed enhancers, and insulators were characterized in each tissue using ChromHMM. Whole-genome bisulfite sequencing (WGBS) was performed to determine the complement of whole-genome DNA methylation with the ChIP-seq data. Hypermethylated and hypomethylated regions were identified across tissues, and these locations were compared with chromatin states to better distinguish and validate regulatory elements in these tissues. Interestingly, chromatin states with the poised enhancer mark H3K4me1 in the spleen and cerebellum and CTCF in the liver displayed the greatest number of hypermethylated sites. Not surprisingly, active enhancers in the liver and spleen, and promoters in the cerebellum, displayed the greatest number of hypomethylated sites. Overall, chromatin states defined by histone marks and CTCF occupied approximately 22% of the genome in all three tissues. Furthermore, the liver and spleen displayed in common the greatest percent of active promoter (65%) and active enhancer (81%) states, and the liver and cerebellum displayed in common the greatest percent of

poised enhancer (53%), repressed enhancer (68%), hypermethylated sites (75%), and hypomethylated sites (73%). In addition, both known and *de novo* CTCF-binding motifs were identified in all three tissues, with the highest number of unique motifs identified in the cerebellum. In summary, this study has identified the regulatory regions of genes in three tissues that play key roles in defining health and economically important traits and has set the precedent for the characterization of regulatory elements in ovine tissues using the Rambouillet reference genome.

Keywords: FAANG, epigenetics, ChIP-seq, WGBS, methylation, sheep, functional genomics, histone modifications

INTRODUCTION

Regulatory element characterization and chromatin state determination in relevant tissues was identified as a critical need for implementing precision breeding within the livestock industry by the Agricultural Animal Genomics Community (Rexroad et al., 2019). To this end, the Functional Annotation of Animal Genomes (FAANG) consortium and the Ovine FAANG project members seek to molecularly define the epigenome in food animals, including sheep (Andersson et al., 2015; Tuggle et al., 2016; Giuffra et al., 2019). Modeled upon the ENCODE project (The ENCODE Project Consortium, 2012), FAANG aims to characterize the epigenome including chromatin histone modifications and DNA methylation (Andersson et al., 2015). The core objective of the Ovine FAANG Project Consortium is to develop a deep and robust public database of transcriptional regulatory features in the sheep genome.

Sheep production for meat, milk, and wool is an important agricultural industry across the globe with more than one billion sheep suited to a diverse range of climates (Hegde, 2019). This diversity is reflected in genetic differences between sheep breeds utilized for varied purposes (Meadows et al., 2008; Al-Mamun et al., 2015). Populations bred for different environments and for contrasting production traits provide the opportunity to study a range of phenotypes within the species. Analysis of elements that control gene expression in sheep tissues is needed as many complex traits such as rumen fatty acid metabolism, lanolin and wool production, growth, and carcass traits cannot be explained solely by variation in transcribed regions (Jiang et al., 2014; Villar et al., 2015; Clark et al., 2017; Kingsley et al., 2019). *In vivo* analysis of regulatory elements will allow researchers to test hypotheses of biological function of putative causal mutations in relevant production tissues. Understanding the phenotypic influences of genetic variance that lie in promoter and enhancer regions is important for trait prediction and the improvement of sheep production.

Functional variants that are causally implicated in phenotypic variation are increasingly found to lie outside of transcribed regions within DNA regulatory elements (Albert and Kruglyak, 2015; Xiang et al., 2019). These regulatory elements can be defined by epigenetic analyses that have not been systematically conducted in sheep. A library of putative regulatory elements in the sheep genome was recently predicted using inference from chromatin states defined in humans (Naval-Sanchez et al., 2018). However, direct experimental characterization of regulatory elements in individual ovine tissues is needed.

The work presented here represents the foundation in the preparation for a deep survey using the same methodology across tissues of the index animal from which the new sheep reference genome was developed. Since the larger FAANG effort has $N = 2$ for each tissue by design (i.e., a large array of tissues from the individual from which the genome was derived), the data collected here also provide a resource for evaluating the larger effort by permitting estimation of interindividual variation in the appearance and tissue distribution of regulatory elements. Three tissues were selected for this study based on their prominence in defining production traits and to span tissues of endodermal, mesodermal, and ectodermal origin and because each presents unique procedural challenges for performing chromatin immunoprecipitation with sequencing (ChIP-seq) assays. The liver is an endodermal-derived tissue that is a key metabolic component of the alimentary system (Villar et al., 2015) and contains a variety of complex carbohydrates that can inhibit various enzymatic reactions required in the ChIP-seq protocol. The spleen is a mesodermal-derived parenchymatous organ important for immune cell production and maturation and contains many natural deoxyribonucleases (DNase) which can present challenges to obtaining sufficient yield of high-quality DNA (Young and Sinsheimer, 1965). The cerebellum is an ectodermal-derived tissue representative of brain tissue and contains a high lipid content which can affect the efficiency of DNA extraction. With these three varied tissues, we developed workflows for assessing chromatin-associated histone modifications, CTCF-binding sites, and DNA methylation to define regulatory elements.

The histone modifications characterized in this study include the trimethylation of histone 3 lysine 4 (H3K4me3) which denotes promoters and acetylation of histone 3 lysine 27 (H3K27ac) which denotes active enhancers (Barski et al., 2007; Wang et al., 2008). The monomethylation of histone 3 lysine 4 (H3K4me1) was characterized to explore poised enhancers, and the trimethylation of histone 3 lysine 27 (H3K27me3) was utilized to define repressed enhancers which silences gene expression in broad regions (Barski et al., 2007; Wang et al., 2008; Pauler et al., 2009). The CCTC-binding factor protein (CTCF) is a key component of the anchors at topologically associated domain boundaries (Lee and Iyer, 2012; Ghirlando and Felsenfeld, 2016). Determination of CTCF and multiple histone

modifications, referred to as marks, allowed us to take advantage of the combinatorial nature of chromatin structure and gene expression regulation (Jenuwein and Allis, 2001; Wang et al., 2008) to categorize the sheep genome into chromatin states.

DNA methylation data derived from whole-genome bisulfite sequencing (WGBS) were incorporated to validate regulatory regions and chromatin states. In mammals, several groups have identified CpG islands that lack methylation are located at gene promoters (Deaton and Bird, 2011). Repressed promoters are marked by higher degrees of methylation associated with transcriptionally silenced gene expression (Weber et al., 2007). Histone methylation and DNA methylation are co-dependent epigenetic marks as enzymatic formation of one will guide the formation of the other and H3K4me3 may physically inhibit methylation of DNA during development (Meissner et al., 2008). Histone methylations and DNA methylation serve as templates for rebuilding one another during mitosis and meiosis and further reinforce segmentation of the genome into functional regions of active or repressed chromatin in adult somatic cells (Cedar and Bergman, 2009) justifying the utility of combined analysis in sheep.

Our objective for this study was to identify the locations of gene regulatory elements in sheep by characterizing histone modifications, CTCF binding, and DNA methylation for the cerebellum, liver, and spleen. Defining regulatory elements in the sheep genome will provide the basis for a greater understanding of the mechanisms that underpin phenotypic variation in important health and production traits in sheep.

MATERIALS AND METHODS

Sample Collection

Tissue was collected postmortem from two pairs of healthy half siblings (one ewe and one wether per pair) totaling four yearling crossbred sheep (Columbia, Polypay, Rambouillet, Suffolk, Targhee) as approved by the Washington State University Institutional Animal Care and Use Committee. Small pieces of liver, spleen, and cerebellum tissues were collected within 30 min postmortem, briefly rinsed with ice cold 1 × PBS, and promptly snap frozen in liquid nitrogen. Samples were transferred from liquid nitrogen directly into a −80°C freezer for storage.

Chromatin Immunoprecipitation

Chromatin immunoprecipitation (ChIP) was performed using commercial antibodies for the histone modifications H3K4me3 (Abcam, cat. # ab8580), H3K4me1 (Abcam, cat. # ab8895), H3K27ac (Abcam, cat. # ab4729), H3K27me3 (Abcam, cat. # ab6002), and CTCF (Millipore, cat. # 07-729) with SimpleChIP Plus Enzymatic Chromatin IP Kit according to the manufacturer's instructions (Cell Signaling Technologies cat. # 9005, Danvers, MA, United States) (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007; Park, 2009). Briefly, tissue was cross-linked with 37% formaldehyde and disaggregated with a Dounce homogenizer. After cell membrane lysis, micrococcal nuclease (MNase) was added and incubated at 37°C and 200 rpm for 20 min to shear the chromatin. Next, the nuclear membrane was lysed, and the sheared chromatin isolated by centrifuging at 15,000×*g* for 1 min at 4°C. Chromatin was incubated with 1 μg of antibody overnight at 4°C in a Hula mixer for 16 h. The following morning, protein G-coated magnetic beads were added and incubated 2 h at 4°C in the Hula mixer. The sample was washed twice with a low salt and once with a high salt buffer. Cross-links were reversed by incubating the sample at 65°C for 30 min at 400 rpm in a thermomixer. Purification was performed with the DNA Purification Buffers and Spin Columns Kit following the manufacturer's instructions (Cell Signaling Technologies, cat. # 14209, Danvers, MA, United States).

Chromatin Immunoprecipitation With Sequencing Library Preparation and Sequencing

Purified DNA samples were quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, catalog number Q32854, Waltham, MA, United States). The DNA size and integrity were verified using a Fragment Analyzer (Agilent, Santa Clara, CA, United States). Libraries were prepared with the TruSeq ChIP Library Preparation Kit (Illumina, Inc., catalog number IP-202-1012, San Diego, CA, United States) for 75 base pair paired-end reads following the manufacturer's instructions and sequenced to at least 20 million mapped reads for "narrow" histone marks H3K4me3, H3K27ac, and CTCF libraries and at least 40 million mappable reads each for "broad" histone marks H3K4me1 and H3K27me3 libraries.

Whole-Genome Bisulfite Sequencing Library Preparation and Sequencing

Whole-genome bisulfite sequencing was performed as a service by Novogene (Beijing, China) on the liver, spleen, and cerebellum in all four animals. Briefly, DNA extracted from these tissues was subjected to agarose gel electrophoresis to test for DNA degradation and potential RNA contamination. The DNA was then quantified using a Nanodrop spectrophotometer (NanoDrop Technologies, Rockland, DE, United States) and a Qubit2.0 fluorometer (Life Technologies, Carlsbad, CA, United States). Lambda phage DNA was spiked in as a negative control for DNA methylation. Since lambda phage DNA lacks DNA methylation, all the cytosines in its DNA should be converted to uracil during bisulfite conversion. Any unchanged cytosine in the lambda phage DNA can thus be used to determine the efficiency of bisulfite conversion. For library construction, DNA samples were fragmented into 200–400 bp using sonication (Covaris S220, Woburn, MA, United States). Next, end repair, A-ligation, and methylation sequencing adapter ligation was performed. The adapter sequences were 5′ adapter (5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′) and 3′ adapter (5′-GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG-3′). Following this, the DNA library was subjected to bisulfite treatment (EZ DNA Methylation Gold Kit, Zymo Research, Irvine, CA, United States). Library concentration was first quantified by Qubit2.0, diluted to 1

ng/μl before checking insert size on Agilent 2100 (Agilent Technologies, Santa Clara, CA, United States), and quantified with more accuracy by quantitative PCR (effective concentration of library > 2 nM). Libraries were then pooled per sample and sequenced paired-end.

Chromatin Immunoprecipitation With Sequencing Data Quality Control, Mapping, and Peak Calling

Quality control assessment of ChIP-seq reads was performed with FastQC, and Trim Galore was used to trim adapter sequences and low-quality bases. PCR duplicates were removed with Picard and the remaining read pair sequences were then mapped to the sheep reference genome *Oar_rambouillet_v1.0* with Bowtie2 (Langmead and Salzberg, 2012; Broad Institute, 2019). Cross-correlations were calculated using MACS2 *predicted* in Galaxy Version 2.1.1.20160309.1 (**Supplementary Figure 1**) (Afgan et al., 2018). Peaks for narrow histone marks H3K4me3 and H3K27ac as well as transcription factor CTCF were called using MACS2 with an input control and a false discovery rate of 0.05 (Zhang et al., 2008; Feng et al., 2012; Thomas et al., 2017). For broad peak histone modifications H3K4me1 and H3K27me3, SICER was implemented with the same input control and a false discovery rate of 0.05 to better account for broader sequence pileup distributions (Zang et al., 2009; Micsinai et al., 2012; Siska and Kechris, 2017). The number of uniquely mapped sequences, non-redundant fraction (NRF), and fraction of reads in peaks (FRiP) for each ChIP-seq sample were calculated using Picard (Heinz et al., 2010; Friedman and Alm, 2012; Landt et al., 2012; Siska and Kechris, 2017; Afgan et al., 2018) (**Supplementary Table 1**). Peak numbers were averaged across samples. Peaks common to multiple samples were determined with BEDTools intersect. The peaks common to three samples with the greatest NRF were determined for H3K4me3 (F1, M1, and M2 for liver; F2, M1, and M2 for spleen; and F1, M1, and M2 for cerebellum), H3K27ac (F1, M1, and M2 for liver; F2, M1, and M2 for spleen; and F1, M1, and M2 for cerebellum), H3K4me1 (F1, M1, and M2 for liver; F2, M1, and M2 for spleen; and F1, M1, and M2 for cerebellum), H3K27me3 (F1, M1, and M2 for liver; F2, M1, and M2 for spleen; and F2, M1, and M2 for cerebellum), and CTCF (F2, M1, and M2 for liver; F2, M1, and M2 for spleen; and F2, M1, and M2 for cerebellum). These consensus peaks were compared with transcription start site locations identified with CAGE assays from the ewe used to generate the reference genome using the deepTools *computeMatrix* function, and heatmaps were plotted with the *plotHeatmap* function (Ramírez et al., 2014; Salavati et al., 2020). Furthermore, peaks were annotated with the GTF file from the reference genome *Oar_rambouillet_v1.0*, and peaks were categorized as near a transcription start site (TSS) (+2 to −2 kb), exonic, intronic, near a transcription termination site (TTS) (+1 to −1 kb), and intergenic using the Homer *annotatePeaks.pl* function (Heinz et al., 2010). Furthermore, normalized bigwig files depicting the sequence enrichment for each library were directly visualized with integrative genomics viewer (IGV) for some gene regions which are known to be active and repressed in each tissue (Robinson et al., 2011). Spearman correlations were calculated between sample BAM signal files using deepTools in Galaxy Version 2.1.1.20160309.1 (Friedman and Alm, 2012; Ramírez et al., 2014; Siska and Kechris, 2017; Afgan et al., 2018).

DNA Methylation Data Quality Control, Mapping, and Methylation Level Characterization

The quality of raw sequences from WGBS was assessed using FastQC v0.11.5. Adapters and low-quality bases (phred score < 20) were trimmed using Trimgalore v0.4.5 with default parameters. Cleaned data for each sample was aligned to the sheep reference genome *Oar_rambouillet_v1.0* using Bowtie2 aligner within BSseeker2 v2.1.8 with default parameters (Langmead and Salzberg, 2012; Guo et al., 2013). The X-chromosome was removed from the analysis to make male and female samples comparable. After mapping, BAM files for the same individual sequenced on multiple lanes were merged, fixmated, and sorted and PCR duplicates were removed using Samtools v1.6 (Li et al., 2009). The methylation level in each cytosine was determined using BSseeker2 with default parameters. Basic statistics on methylation were determined using the *mstat* function in CGmaptools v0.0.6 (Guo et al., 2018). Regions of the genome hypomethylated and hypermethylated for each sample were determined with methPipe v3.4.3 following the manual with default parameters (Song et al., 2013).

Chromatin State and CTCF Motif Analysis

Chromatin states were characterized by employing a hidden Markov model in ChromHMM, which assessed signal overlap between histone marks within a tissue and binned the genome into a given number of chromatin states (Ernst and Kellis, 2010, 2012, 2017; Gorkin et al., 2017, 2020). The two male samples (M1 and M2) exhibited the greatest NRF and Spearman correlations and were therefore used in chromatin state analysis. The *LearnModel* function in ChromHMM was implemented with given chromatin states of two through 20 for each animal, and the model with the optimal number of chromatin states was examined using the *CompareModels* function in ChromHMM (Gorkin et al., 2017, 2020). The optimal number of chromatin states was determined as the model where the median Pearson correlation for all states plotted against each chromatin state model plateaued and were tightly correlated with the model with the greatest number of states (**Supplementary Figure 2**) (Gorkin et al., 2017, 2020). The consensus of chromatin states between two animals (M1 and M2) was used to generate the heatmap and for further comparative analyses. Location similarities and differences between chromatin states, hypermethylated regions, and hypomethylated regions were assessed with BEDTools intersect within each tissue, and the consensus within each tissue was used to examine chromatin state and DNA methylation similarities and differences between liver, spleen, and cerebellum tissues (Quinlan, 2014). An Upset R plot was generated to display chromatin state similarities and differences between tissues (Lex et al., 2014; Conway et al., 2017). Significantly enriched

known and *de novo* CTCF motifs were identified and compared with other species by implementing the findMotifs.pl script in HOMER (Heinz et al., 2010). The proximity of annotated TSS generated from CAGE data to promoter chromatin states was examined with deepTools *computeMatrix* and *plotHeatmap* functions (**Supplementary Figure 7**) (Ramírez et al., 2014; Salavati et al., 2020).

RESULTS

Genetic regulatory elements were characterized across the sheep genome in the liver, spleen, and cerebellum using CTCF binding and ChIP-seq of four histone marks, as well as DNA methylation status. Locating regulatory elements within and between tissues will provide the basis for identifying variation in these elements that may influence various phenotypic traits in sheep. Furthermore, these results represent a resource for estimating interindividual variation in the regulatory states of tissues to provide context for the FAANG project that aims to characterize these states in a broad array of tissues in a single individual from which the reference genome was produced.

Mapping Summary and Statistics

Mapping statistics were calculated to assess the assay quality, library preparation, and sequence coverage for each sample. Across animals, ChIP-seq reads had a consistent average mapping rate of 78.23, 78.39, and 76.82% to the *Oar_rambouillet_v1.0* genome for the liver, spleen, and cerebellum, respectively. The number of uniquely mapped paired-end reads averaged 40,757,252 for H3K4me3, 42,306,275 for H3K27ac, 53,171,657 for H3K4me1, 55,901,184 for H3K27me3, and 45,491,017 for CTCF across all three tissues. The number of uniquely mapped reads, NRF, and FRiP for each sample are displayed in **Supplementary Table 1**.

Whole-genome bisulfite sequencing of cerebellum, liver, and spleen samples from the four sheep generated a total of 986, 1,070, and 904 million paired end reads, respectively, with a read length of 2 × 150 bp. The number of reads uniquely mapped to the reference genome was 84.24, 78.86, and 82.48% for the cerebellum, liver and spleen, respectively. The uniquely mapped bases covered the reference genome (*Oar_rambouillet_v1.0*; genome size ∼2.87 Gb) at an average depth of 21 × (range 18× to 26×). Bisulfite conversion rate was ∼99.9% for all the samples. Mapping statistics for each tissue sample per sheep are displayed in **Supplementary Table 2**.

Chromatin Immunoprecipitation With Sequencing Peak Calling

The locations of sequence signal enrichment were identified for all four histone marks and CTCF for each liver, spleen, and cerebellum sample by mapping the reads to the reference genome *Oar_rambouillet_v1.0*. The number of peaks normalized by chromosome length (in Mb; **Figure 1**) and the width of the peaks along the assembly were calculated from the mapped read depth. For each mark, the percent of the total number of peaks observed in the genome that lie on each chromosome is plotted in **Figure 1** which shows an overall even distribution across chromosomes with some exceptions. The lowest number of peaks was called in narrow mark H3K4me3 (means of 10,458 in the liver, 13,389 in the spleen, and 16,911 in the cerebellum), with the lowest number of peaks per Mb on chromosomes 23 (2.77 peaks/Mb), 26 (2.64 peaks/Mb), and 16 (2.47 peaks/Mb) in the liver, spleen, and cerebellum, respectively. The greatest number of H3K4me3 peaks per Mb for the liver, spleen, and cerebellum was on chromosomes 14 (6.16 peaks/Mb), 20 (5.17 peaks/Mb), and 11 (4.61 peaks/Mb), respectively. The average widths of H3K4me3 peaks were 168, 178, and 313 bp for the liver, spleen, and cerebellum. The mean number of peaks called for the H3K27ac mark was 30,553 in the liver, 35,327 in the spleen, and 35,877 in the cerebellum with the lowest number of peaks called on chromosomes 10 (2.54 peaks/Mb), 26 (2.25 peaks/Mb), and 6 (2.72 peaks/Mb) for the respective tissues. The greatest number of H3K27ac peaks was called on chromosome 11 for all three tissues, and peak widths averaged 239, 240, and 238 bp in the liver, spleen, and cerebellum for this narrow mark. The final narrow mark, CTCF, averaged 26,517 peaks in the liver, 28,362 in the spleen, and 26,244 in the cerebellum. The lowest number of CTCF peaks were called on chromosome 24 (1.56 peaks/Mb for the liver, 1.49 peaks/Mb for the spleen, and 2.05 peaks/Mb in the cerebellum), and the greatest number of peaks were called on chromosome 6 (5.50 peaks/Mb in the liver, 5.73 peaks/Mb in the spleen, and 5.07 peaks/Mb in the cerebellum) for all three tissues. The width of CTCF peaks was similar to other narrow marks, with averages of 114 bp in the liver, 265 bp in the spleen, and 144 bp in the cerebellum.

The greatest number of peaks was called in broad mark H3K4me1 (means of 47,828 in the liver, 33,931 in the spleen, and 51,766 in the cerebellum), which is consistent with several tissues in cattle (Fang et al., 2019). Chromosomes with the lowest number of H3K4me1 peaks per Mb included 21 (2.34 peaks/Mb) for the liver, 26 (2.90 peaks/Mb) for the spleen, and 20 (3.12 peaks/Mb) for the cerebellum, and the greatest number of peaks per Mb was on chromosome 7 (4.99 peaks/Mb for the liver, 7.79 peaks/Mb for the spleen, and 4.98 peaks/Mb in the cerebellum) for all three tissues. The average width of broad peak H3K4me1 was greater than for the narrow peaks described above, as expected, at 948 bp for the liver, 2,963 bp for the spleen, and 1,909 bp for the cerebellum. Lastly, the broad mark H3K27me3 had a lower average number of peaks called compared with H3K4me1 (mean of 39,162 in the liver, 29,939 in the spleen, and 26,244 in the cerebellum). The lowest number of H3K27me3 peaks per Mb of chromosome length were on chromosomes 26 (3.04 peaks/Mb), 24 (2.58 peaks/Mb), and 11 (1.84 peaks/Mb) for the liver, spleen, and cerebellum, respectively. The greatest number of peaks was on chromosome 13 (4.86 peaks/Mb) for the liver and chromosome 6 for both the spleen (4.39 peaks/Mb) and cerebellum (4.94 peaks/Mb). The average width of broad H3K27me3 peaks was 440 bp in the liver, 2,143 bp in the spleen, and 653 bp in the cerebellum. Peaks in common across the animals were calculated for all five ChIP-seq experiments and displayed for the liver, spleen, and cerebellum (**Supplementary Figure 2**). Interestingly, half siblings (F1 and M1, F2 and M2) displayed a greater number of peaks in common with each other.

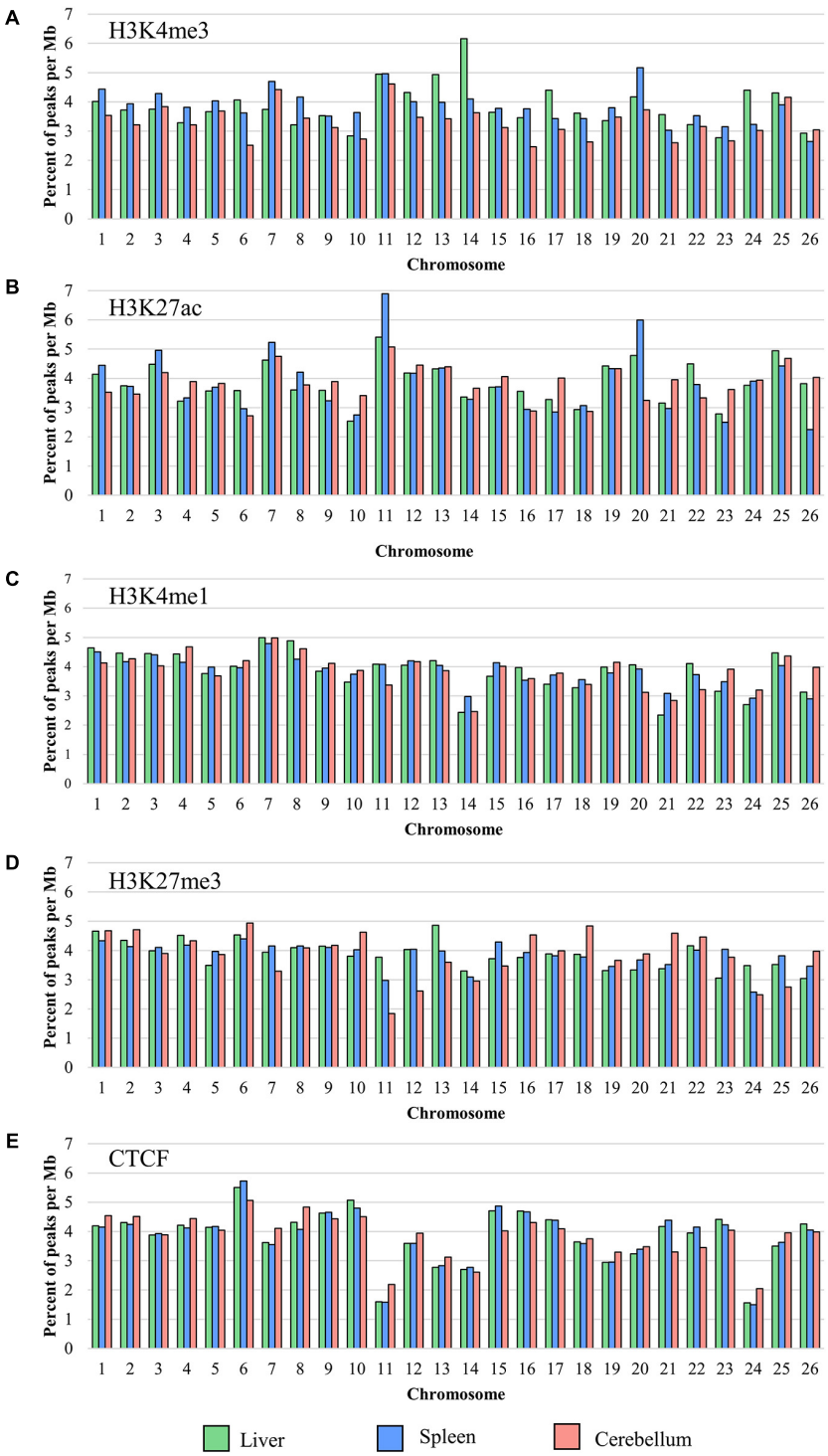


FIGURE 1 | The percent of the total number of peaks normalized per Mb on each chromosome for **(A)** H3K4me3, **(B)** H3K27ac, **(C)** H3K4me1, **(D)** H3K27me3, and **(E)** CTCF averaged from all four animals (F1, F2, M1, and M2).

The proximity of H3K4me3 peaks to TSS was investigated by comparing consensus H3K4me3 peaks and CAGE data generated by Salavati et al. (2020). Not surprisingly, H3K4me3 peaks were detected on both sides of the TSS in the liver, spleen, and cerebellum tissues. The signal distributions and heatmaps from 2 kb upstream and downstream of the TSS locations are displayed in **Figure 2**. In addition, the consensus peaks for H3K4me3, H3K27ac, H3K4me1, H3K27me3, and CTCF were

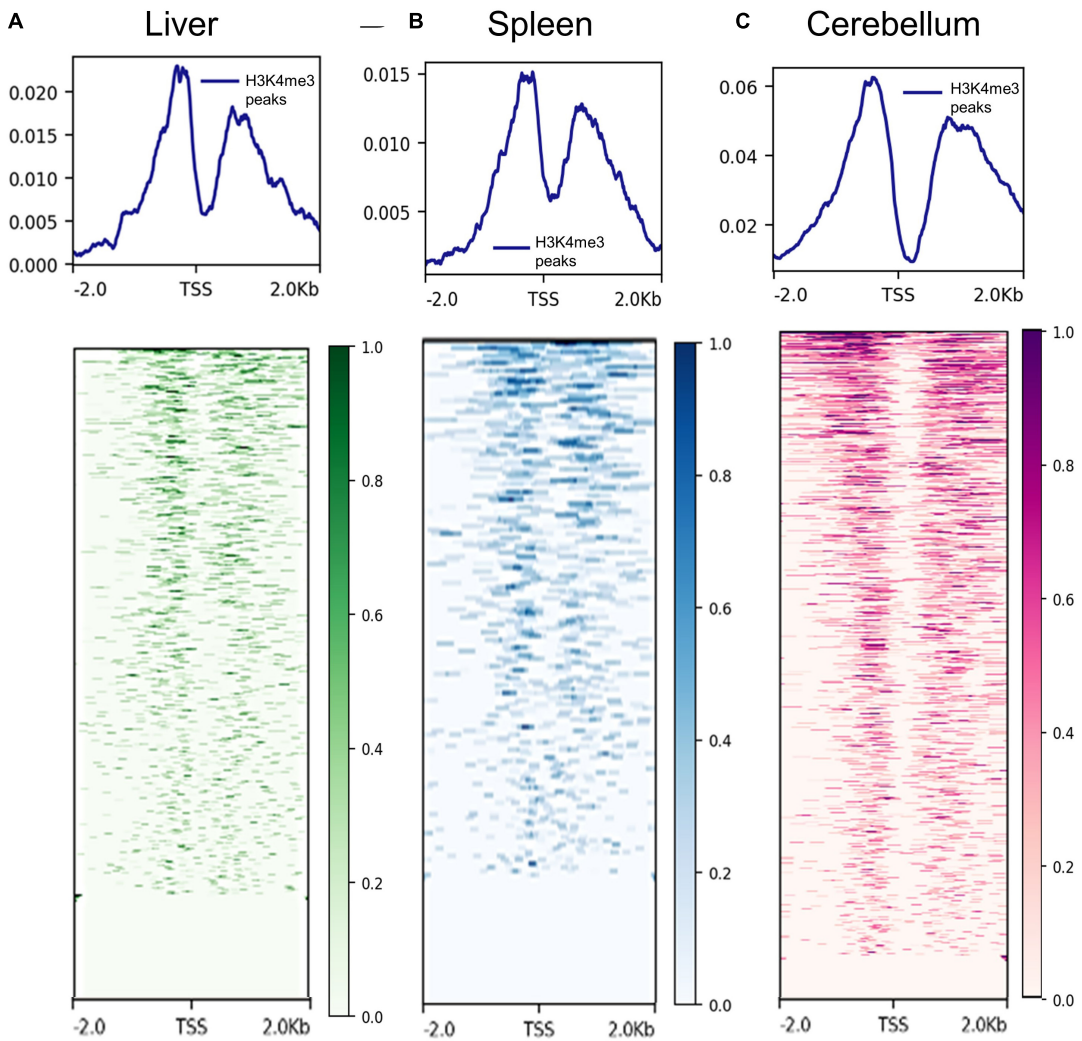


FIGURE 2 | Signal of H3K4me3 ChIP-seq peaks 2 kb upstream and downstream of transcription start sites (TSS) identified by CAGE assays. **(A)** Liver H3K4me3 signal (from F1, M1, and M2 consensus peaks) near TSS annotated in the reference genome, **(B)** spleen H3K4me3 signal (from F2, M1, and M2 consensus peaks) near annotated transcription start sites (TSS), and **(C)** cerebellum H3K4me3 signal (from F1, M1, and M2 consensus peaks) near annotated TSS.

annotated with the *Oar_rambouillet_v1.0* genome annotation file and these classifications are displayed in **Supplementary Figures 3–5**. The histone modification H3K4me3 had the greatest proportion of peaks annotated as near a TSS when compared with other histone modifications in all three tissues. H3K27ac and H3K4me1 histone modifications displayed intronic annotation most commonly, and H3K27me3 and CTCF displayed mostly intergenic peak annotation.

Visual Assessment of Sequence Pileup

The peak predictions were directly examined in the IGV (Robinson et al., 2011) for regions known to be active or repressed in the three tissues, to provide an evaluation of the success of the process in properly classifying chromatin states. One example of an expected active region for each liver, spleen, and cerebellum tissue as well as one region expected to be repressed in all tissues is displayed in **Figure 3**. Albumin (*ALB*), a gene that encodes a plasma protein synthesized in hepatocytes and expected to be active in the liver, has one promoter and two enhancers annotated in humans that are within 2 kb upstream from the start of the gene (Frain et al., 1990; Hayashi et al., 1992; Bernardi et al., 2012; Fagerberg et al., 2014). Sequence pileup for active histone marks in the liver was observed in all four sheep that overlap with approximate locations of regulatory elements of *ALB* in humans, and there were low levels of DNA methylation in these regions (**Figure 3A**). The region upstream of Solute carrier family 11 member 1 (*SLC11A1*), a gene expected to be active in the spleen and encodes a membrane protein involved with macrophage development, displayed sequence pileup for active marks H3K4me3 and H3K27ac and low levels of DNA methylation directly upstream (**Figure 3B**) (Hedges et al., 2013). Paired box 6 (*PAX6*) is known to be involved in the development of neural tissues and maturation of granule neurons in the cerebellum and is known to have a promoter

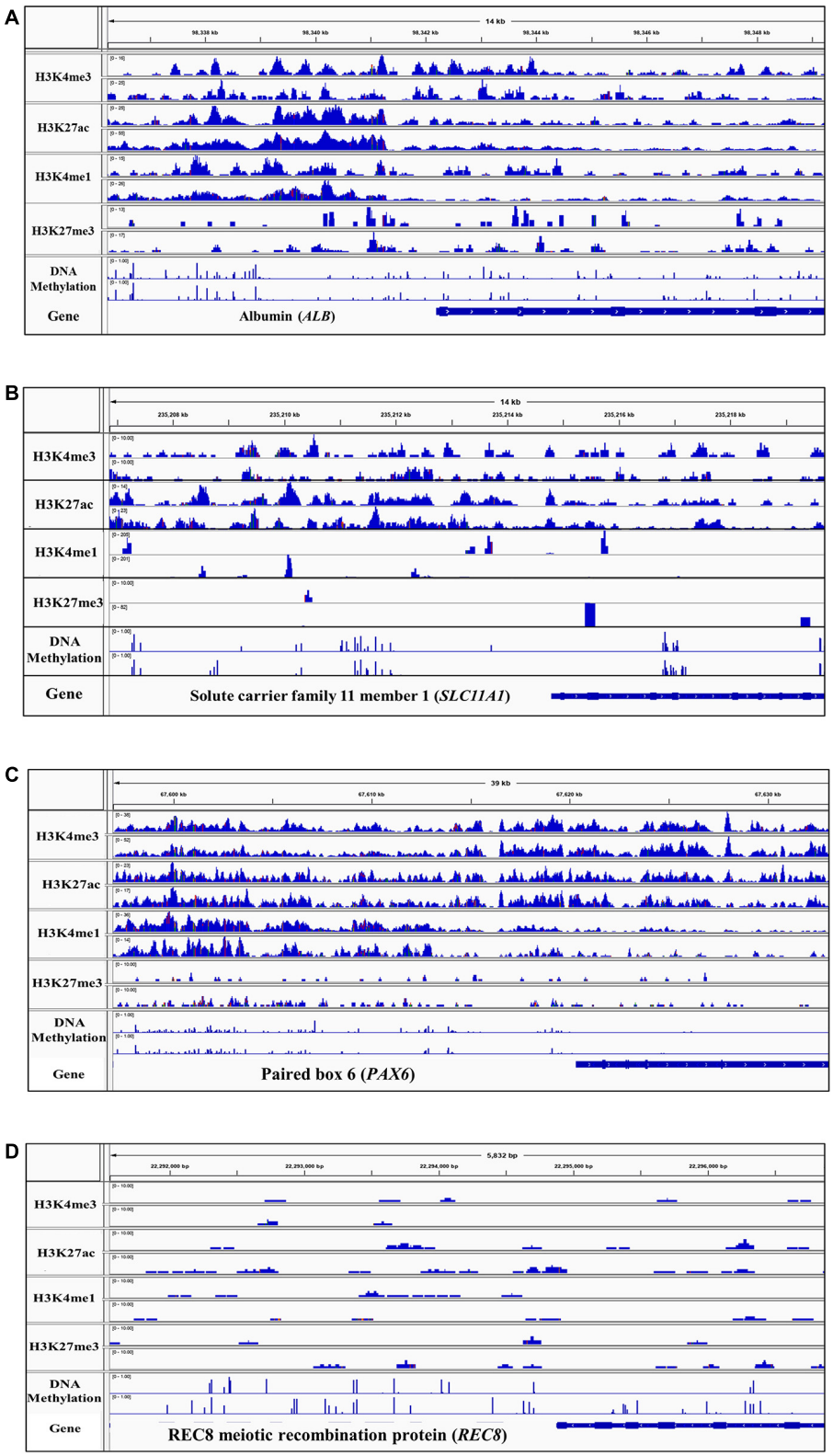


FIGURE 3 | Integrative genomics viewer (IGV) screenshot of sequence pileup normalized with the input control for active and repressive histone marks and DNA methylation in two representative samples (M1 and M2) for **(A)** positive control Albumin (*ALB*) gene in the liver, **(B)** positive control Solute carrier family 11 member 1 (*SLC11A1*) in the spleen, **(C)** positive control Paired box 6 (*PAX6*) in the cerebellum, and **(D)** negative control REC8 gene (*REC8*) in all three tissues.

and multiple enhancers both upstream and downstream of the gene (Ha et al., 2015; Divya et al., 2016). Furthermore, *PAX6* has greater expression in the cerebellum than other tissues in sheep which is supported by the sequence pileup of active histone marks H3K4me3 and H3K27ac, with some activity of H3K4me1 and little DNA methylation (**Figure 3C**) (Jiang et al., 2014). In contrast, the REC8 meiotic recombination protein (*REC8*) is a gene that encodes a meiosis-specific protein involved in the synapsis of sister chromatids that is not expected to be active in the liver, spleen, or cerebellum (Xu et al., 2005). This gene location shows no sequence pileup in all four sheep in the liver, spleen, or cerebellum and several methylated regions (**Figure 3D**).

Variability in Histone Marks Between Animals

Correlations were calculated for histone marks and for DNA methylation between samples to evaluate interanimal variation in sequence pileup signal for the liver, spleen, and cerebellum (Friedman and Alm, 2012; Siska and Kechris, 2017). Correlations of ChIP-seq data (Spearman) and DNA methylation data (Pearson) averages for all four animals and males only (in parentheses) are provided in **Table 1**. The narrow mark H3K4me3 was moderately correlated between all four animals in the liver (0.66) and spleen (0.54) and highly correlated in the cerebellum (0.85). In males, H3K4me3 was highly correlated in the liver (0.86), spleen (0.71), and cerebellum (0.88). The narrow mark H3K27ac was highly correlated between samples across all three tissues in the liver (0.89 overall and 0.95 in males), spleen (0.78 overall and 0.84 in males), and cerebellum (0.70 overall and 0.91 in males).

The broad mark H3K4me1 also showed high correlation in two tissues, namely the liver (0.71 overall and 0.93 in males) and cerebellum (0.82 overall and 0.91 in males), but the correlation in the spleen was markedly lower (0.47 overall and 0.56 in males), and overall, the correlations between the spleen samples were lower than the liver and cerebellum for all four histone marks. This is evident in H3K27me3 in the spleen (0.37 overall and 0.44 in males) than in the liver (0.58 overall and 0.74 in males) and cerebellum (0.72 overall and 0.83 in males). The correlations of DNA methylation signal between samples ranged from 0.70 to 0.76, with the liver and cerebellum displaying the greatest correlation between the two males (0.76). However, sex differences in correlations were not observed, as each female has a moderate to high correlation with both the other female (0.54–0.84) and both males (0.44–0.92) for each mark within all three tissues.

Principal Component Analysis of DNA Methylation

A principal component analysis was performed with the DNA methylation data to investigate similarity and differences between samples and tissues. Eigenvalues were calculated based on the position of CG methylation signal in all animals for all three tissues, and the first two eigenvalues (PC1 and PC2) were plotted (**Figure 4**). Samples cluster distinctly by tissue type rather than by sex or individual animal. The greatest spread of points within a tissue was observed in the liver. The first eigenvalue (PC1, 27.56%) shows separation of the liver, spleen, and cerebellum. The second eigenvalue (PC2, 12.16%) shows another dimension of separation of the cerebellum and liver from the spleen.

Methylation Level at CG and Non-CG Sites

Average methylation levels were calculated and compared in each of the three tissues in both the CG and non-CG sites (**Figure 5A**). Non-CG sites are defined as CHG and CHH where H is either A/T/C. CG sites have an average methylation level ranging between 70 and 81% across different tissues. Specifically, cerebellum samples have an average methylation level of 81.4%, whereas liver and spleen samples have an average methylation level of 70.3 and 76.9%, respectively. The average methylation level of cytosines at non-CG contexts (CHG and CHH) is nine-fold higher in the cerebellum (1.7–2.1%) than in spleen and liver samples (0.2%) (**Figure 5B**).

Chromatin State Assignment and Correlation With Methylation Status

The relative positions of the combination of specific histone marks provide a more complete definition of the overall regulatory chromatin state than individual peak calling. Regulatory elements were defined for two animals (M1 and M2) using a hidden Markov model employed by ChromHMM which assigns 200 bp bins across the genome to a given number of chromatin states based on the combined histone modification signal profiles (Ernst and Kellis, 2010, 2017). The genome was categorized into two through 20 chromatin states using ChromHMM. The optimal number of states was determined to be nine, as it was the lowest number of states that had greater than 0.95 correlation of all samples to 20 states, which captures the complexity of the data with fewer states (see **Supplementary Figure 2**) (Gorkin et al., 2017, 2020). These nine chromatin states are categorized as follows: promoter, active enhancer, poised enhancer, repressed enhancer, CTCF, and three or four states of

TABLE 1 | Average correlations of sequencing signal between all four animals.

Tissue	H3K4me3	H3K27ac	H3K4me1	H3K27me3	DNA methylation
Liver	0.66 (0.86)	0.89 (0.95)	0.71 (0.93)	0.58 (0.74)	0.72 (0.76)
Spleen	0.54 (0.71)	0.78 (0.84)	0.47 (0.56)	0.37 (0.44)	0.70 (0.74)
Cerebellum	0.85 (0.88)	0.70 (0.91)	0.82 (0.91)	0.72 (0.83)	0.73 (0.76)

Spearman correlations were used for ChIP-seq data and Pearson correlations were used for DNA methylation data. Parentheses indicate correlations between the replicates used in the ChromHMM chromatin state analysis.

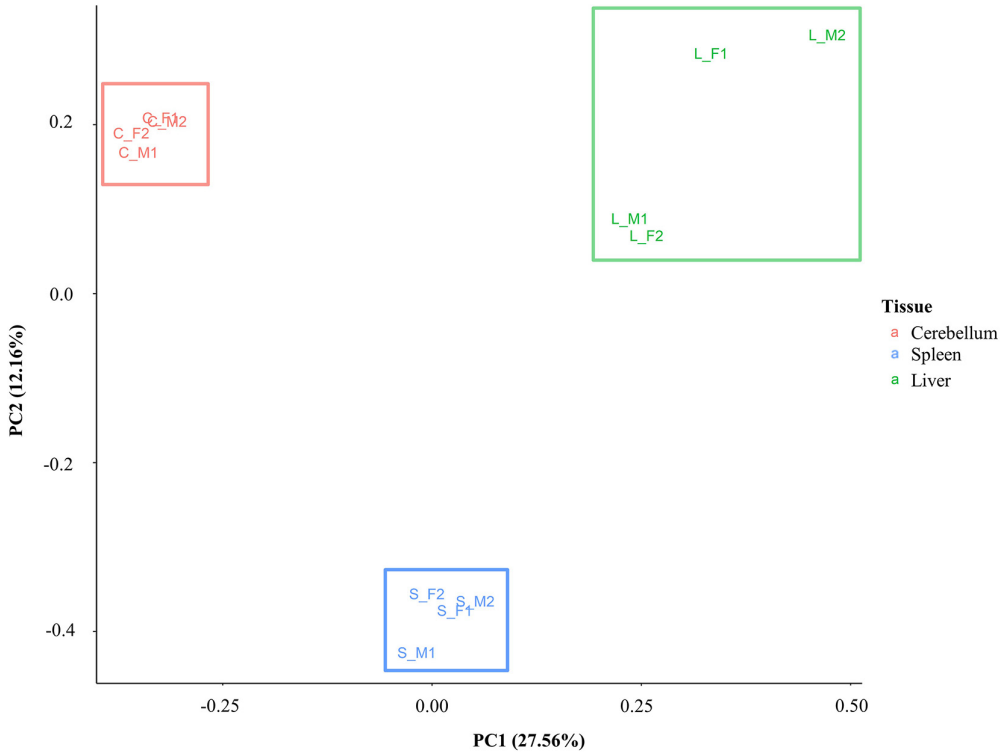


FIGURE 4 | Principal component analysis plot based on CG methylation. Four animals are labeled as F1, F2, M1, and M2. The cerebellum, liver, and spleen samples are labeled as C, L, and S, respectively.

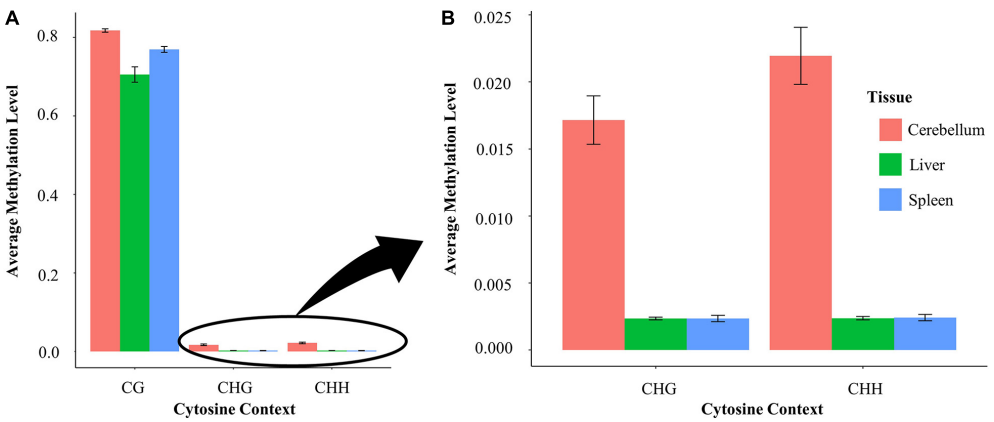


FIGURE 5 | (A) Methylation level at CG compared with non-CG sites in the liver, spleen, and cerebellum and **(B)** methylation level at non-CG (CHG and CHH) sites in each tissue enlarged.

quiescent/low signal. The consensus of chromatin states assigned to both M1 and M2 was used for further analyses.

The signal of all the histone marks and the nine chromatin states for each tissue is displayed as heatmaps in **Figure 6**. Regions with primarily H3K4me3 signal often overlapping with H3K27ac are considered promoters, regions with strong H3K27ac signal are considered active enhancers, regions with H3K4me1 often paired with weak H3K27me3 signal are considered poised enhancers, and regions with strong H3K27me3 signal are considered repressed enhancers (Wang et al., 2008; Creyghton et al., 2010; Core et al., 2014; Carelli et al., 2018). All four of these categories of regulatory elements were observed and displayed in the heatmaps, with the addition of a weak poised enhancer state in the spleen and weak repressed enhancer state in the cerebellum which both displayed lower but still distinguishable signal. In addition, regions with CTCF signal which overlap with other marks including H3K4me1 and H3K27me3 were observed in the liver and

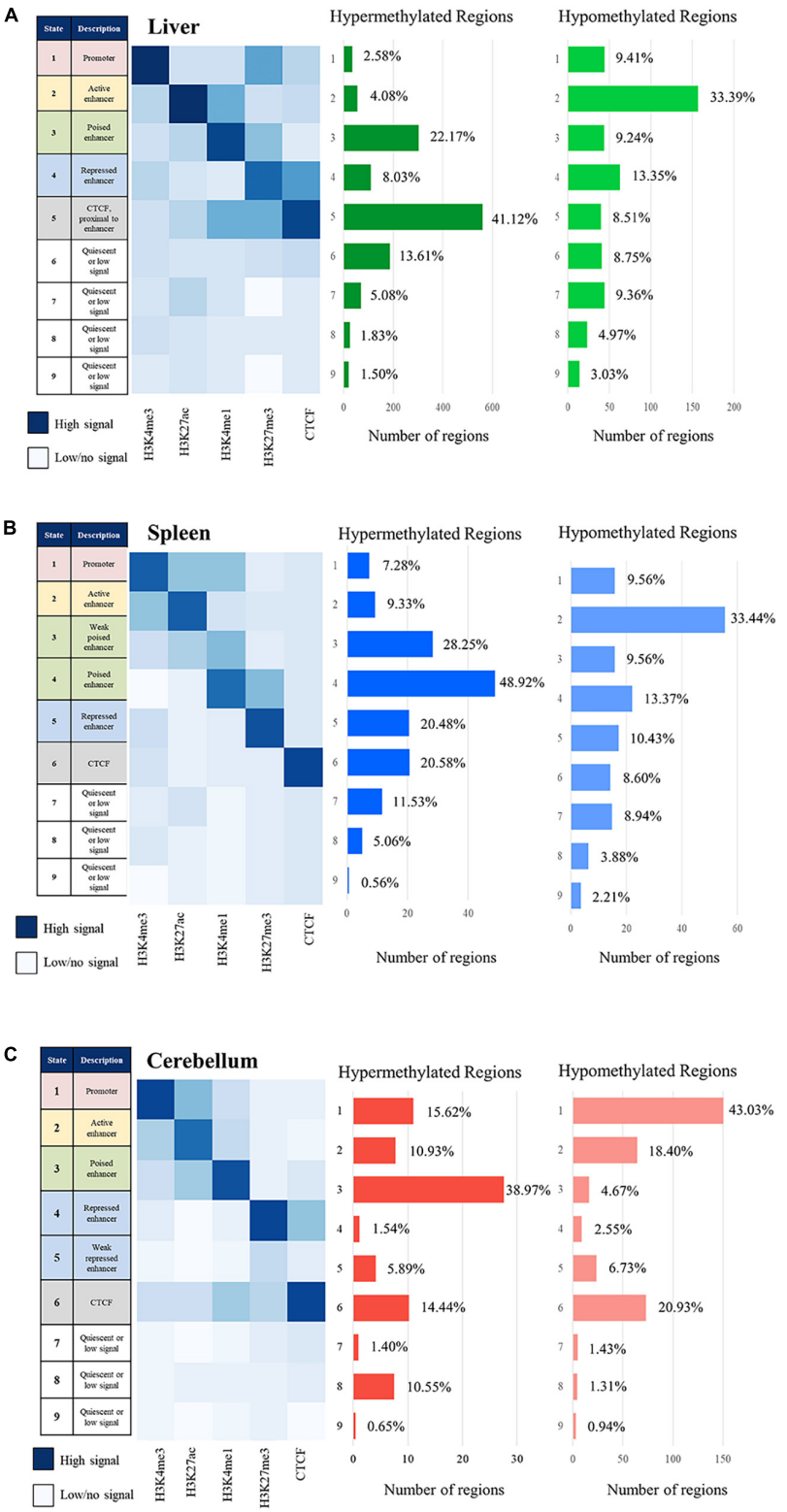


FIGURE 6 | Chromatin state description and ChromHMM heatmap with histone mark signal overlap consensus from M1 and M2 compared with the number of hypermethylated regions and hypomethylated region consensus per Mb for M1 and M2 for the **(A)** liver, **(B)** spleen, and **(C)** cerebellum.

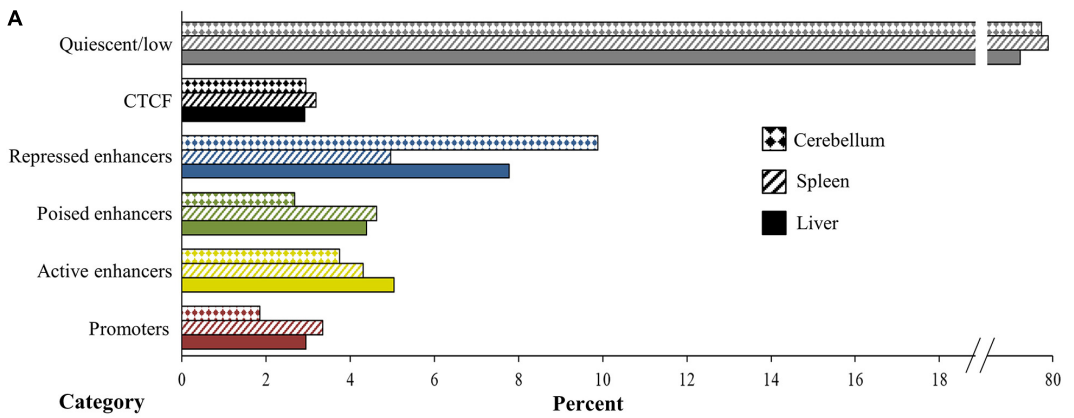
cerebellum. Lastly, quiescent/low states had very little signal in any of the five marks.

The correlation of DNA methylation status with predicted chromatin state was examined by estimating the number of hyper- and hypomethylated regions per Mb within the boundaries of the regulatory elements in the nine defined chromatin states. The greatest number of hypomethylated regions was observed in active enhancer regions in the liver and spleen and in active promoter regions in the cerebellum, as expected if our process was correctly identifying regulatory elements and classifying them as actively transcribed genes. The greatest number of hypermethylated regions was observed in poised enhancers and CTCF in the liver, weak poised and poised enhancer regions in the spleen, and poised enhancer regions in the cerebellum, also consistent with the process correctly classifying regulatory elements.

Distribution of Chromatin States in the Genome and Proximity to TSS

The chromosomal segments spanned by regulatory elements, as defined by the histone mark peaks, were combined and summarized to estimate the overall extent and percent of the genome representing regulatory elements and their chromatin state among the three tissues examined. Chromatin states from the ChromHMM analyses were categorized and combined into promoter, active enhancer, poised enhancer including weak poised enhancers, repressed enhancer including weak repressed enhancers, and quiescent or low signal categories and averaged for each tissue (**Figure 7**). Promoters comprise 2.95% of the genome in the liver, 3.35% in the spleen, and 1.85% in the cerebellum, and active enhancers occupy 5.04% of the genome in the liver, 4.30% in the spleen, and 3.74% in the cerebellum. In addition, 4.38% of the genome in the liver, 4.63% in the spleen, and 2.68% in the cerebellum are categorized as poised enhancers, while 7.78% of the genome in the liver, 4.96% in the spleen, and 9.89% in the cerebellum are considered repressed enhancers. The percent of the genome that had primarily CTCF signal was 2.92% in the liver, 3.19% in the spleen, and 2.94% in the cerebellum. Cumulatively, states considered as enriched with histone mark and CTCF signal intensity by ChromHMM, which includes the promoter, enhancer, and CTCF functional elements, comprise approximately 23.08% of the genome in the liver, 20.44% in the spleen, and 21.10% in the cerebellum. Not surprisingly, the largest percent of the genome, 76.91% in the liver, 79.56% in the spleen, and 78.90% in the cerebellum, was categorized as quiescent or low signal.

The locations of assigned promoter chromatin states were compared with TSS generated from CAGE data for the liver, spleen, and cerebellum. Both the signal distribution and heatmap plots display a strong signal before and after the TSS in all three tissues (**Supplementary Figure 7**). This signal is similar to the distribution of the H3K4me3 peak signal before and after



B

Category	Liver	Spleen	Cerebellum
Quiescent/low	76.91%	79.56%	78.90%
CTCF	2.92%	3.19%	2.94%
Repressed enhancers	7.78%	4.96%	9.89%
Poised enhancers	4.38%	4.63%	2.68%
Active enhancers	5.04%	4.30%	3.74%
Promoters	2.95%	3.35%	1.85%

FIGURE 7 | Percent of the genome in the liver, spleen, and cerebellum (from M1 and M2) assigned to each category of quiescent/low (gray), CTCF (black), repressed enhancer (blue), poised enhancer (green), active enhancer (gold), and promoter (red) depicted visually in panel **(A)** the bar graph and numerically in panel **(B)** the table.

TSS, which is not surprising as the ChromHMM model assigns promoter states based on the presence of H3K4me3 signal. It is worth noting that the CAGE data used in this study were generated from the reference genome animal, a Rambouillet, which is different from the crossbred animals used in this study and may explain some of the signal noise.

Similarities and Differences of Chromatin States Between Tissues

Similarities and differences of promoters, enhancers, and methylated regions within and between tissues were examined and percentages of overlap are displayed in **Figure 8**. Active promoters were 64.76% similar between the liver and spleen, 25.39% between the liver and cerebellum, and 35.69% between the spleen and cerebellum. The liver had 81.09 and 51.10% of active enhancers in common with the spleen and cerebellum, respectively. The spleen and cerebellum had 53.85% similarity of active enhancers. Poised enhancers were shared 51.90% between the liver and spleen, 52.72% between the liver and cerebellum, and 38.27% between the spleen and cerebellum. The percent of repressed enhancers that overlapped between the liver and spleen was 56.05%. The liver and cerebellum repressed enhancers overlapped 67.90%, and the spleen and cerebellum repressed enhancers overlapped 41.66%. Hypermethylated genomic locations overlapped 4.42% and hypomethylated regions overlapped 56.05% between the liver and spleen. The liver and cerebellum displayed more similar hypermethylated and hypomethylated regions, 75.42 and 72.89%, respectively, than the spleen and cerebellum, 19.44 and 32.51%, respectively.

CTCF-Binding Motifs

The insulator CTCF is often present at the boundaries of topologically associated domains (TADs), compartments of chromatin interactions, across the genome (Beagan and Phillips-Cremins, 2020). The location of significant ($P < 0.00001$) CTCF-binding motifs both known from previous research and *de novo* was identified across the genome in the liver, spleen, and cerebellum (Heinz et al., 2010). Of these, 13 were present in at least three animals (**Table 2**). Three motifs, MYB3R4, MYB3R1, and Pdx1, were significantly enriched in the liver, spleen, and cerebellum tissues. The liver and spleen exhibited the most significantly enriched CTCF motifs in common (TAGL, Six2, RRTF1, Sox6, SVP, and TGA2). One motif, ZBTB19, was enriched in the spleen and cerebellum. The cerebellum had three enriched motifs (Elk4, Pho2, and BZR1) not present in the liver or spleen. In addition, *de novo* motifs were identified in all three tissues. The top three most significant *de novo* motifs per sample in the liver, spleen, and cerebellum are reported in **Tables 3–5**, respectively. Of the total number of *de novo* motifs, 16, 13, and 21 were identified as unique to the liver, spleen, and cerebellum, respectively. Sixteen *de novo* motifs were identified in both the liver and spleen, while the cerebellum had only three *de novo* motifs in common with the other tissues.

DISCUSSION

The goal of this study was to characterize regulatory elements in ovine liver, spleen, and cerebellum using ChIP-seq and WGBS. The three selected tissues, the liver, spleen, and

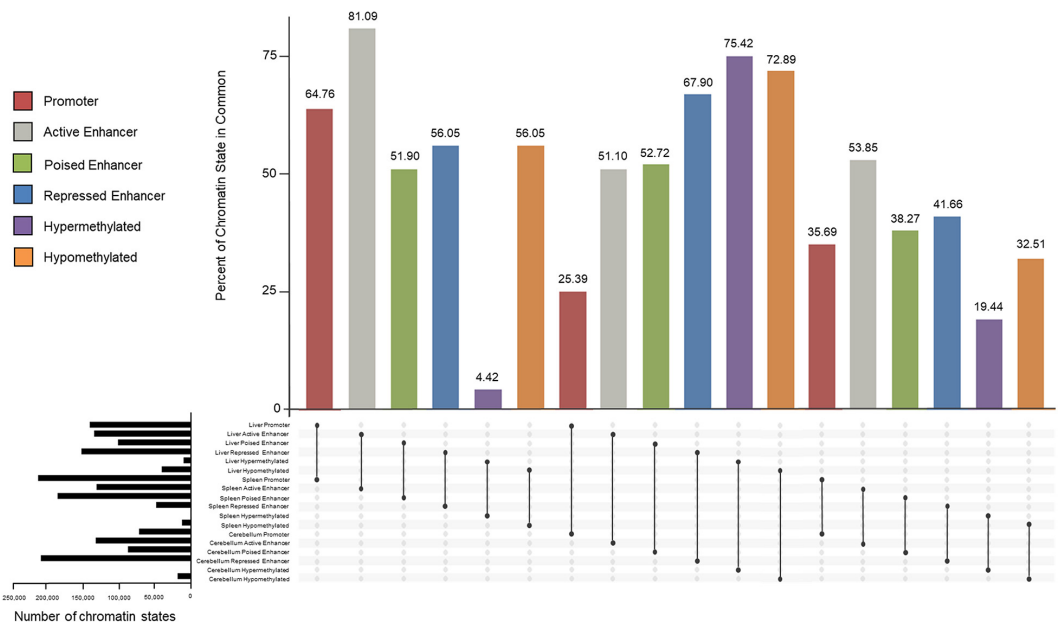


FIGURE 8 | Percent of overlapping promoter (red), active enhancer (gray), poised enhancer (green), and repressed enhancer (blue) chromatin state categories and hypermethylated (purple) and hypomethylated (orange) regions between the liver, spleen, and cerebellum tissues of the consensus categories from M1 and M2. The total number of chromatin states for each tissue is displayed in black horizontal bars.

TABLE 2 | Known CTCF motifs present in the top 10 most significant motifs across multiple samples.

Known motif name	Known motif	Tissue	Number of samples	Mean *P* value	Mean percent of target sequences with motif	Mean percent of background sequences with motif
MYB3R4 (MYB)		Liver, spleen, and cerebellum	7	1E-2612	13.54%	1.27%
TAGL1 (MADS)		Liver and spleen	6	1E-2167	44.33%	18.98%
MYB3R1 (MYB)		Liver, spleen, and cerebellum	6	1E-1632	12.85%	2.42%
Pdx1 (homeobox)		Liver, spleen, and cerebellum	6	1E-1475	37.21%	17.82%
Six2 (homeobox)		Liver and spleen	5	1E-1486	30.93%	13.64%
RRTF1 (APTEREBP)		Liver and spleen	4	1E-1655	7.16%	0.55%
Sox6 (HMG)		Liver and spleen	4	1E-931	40.23%	23.32%
ZBTB19 (Zf)		Spleen and cerebellum	4	1E-418	8.27%	3.07%
SVP (MADS)		Liver and spleen	3	1E-1897	28.39%	9.82%
TGA2 (bZIP)		Liver and spleen	3	1E-1792	16.14%	3.27%
Elk4 (ETS)		Cerebellum	3	1E-61	3.69%	2.07%
Pho2 (bHLH)		Cerebellum	3	1E-32	1.72%	0.92%
BZR1 (BZR)		Cerebellum	3	1E-29	0.68%	0.25%

TABLE 3 | Top three *de novo* CTCF motifs present in each sample in the liver.

Animal	*De novo* motif	*P* value	Percent of target sequences with motif	Percent of background sequences with motif
F1		1E-3278	31.57%	3.37%
		1E-3012	23.91%	1.74%
		1E-2873	23.31%	1.76%
F2		1E-1388	7.03%	0.62%
		1E-1349	6.55%	0.54%
		1E-1345	6.11%	0.45%
M1		1E-8604	21.40%	0.41%
		1E-7739	26.14%	1.17%
		1E-7299	21.19%	0.63%
M2		1E-10234	44.68%	4.07%
		1E-8614	34.87%	2.59%
		1E-8422	42.53%	4.89%

TABLE 4 | Top three *de novo* CTCF motifs present in each sample in the spleen.

Animal	*De novo* motif	*P* value	Percent of target sequences with motif	Percent of background sequences with motif
F2		1E-12441	29.41%	0.72%
		1E-12221	38.23%	2.03%
		1E-12174	38.03%	2.01%
M1		1E-7022	24.88%	1.16%
		1E-6916	30.15%	2.24%
		1E-6704	25.65%	1.42%
M2		1E-5921	24.42%	0.88%
		1E-5710	24.12%	0.93%
		1E-5440	20.87%	0.61%

cerebellum, each represent a different developmental origin and are important to metabolism, immune response, and motor control, respectively. We have demonstrated the successful application of the micrococcal nuclease ChIP protocol across these tissues and the bioinformatic pipeline for the analysis of ChIP-seq in sheep. Furthermore, this study has incorporated the value of coupled histone modification and DNA methylation data toward a better understanding of regulatory regions in the sheep genome.

Micrococcal nuclease was used to shear the chromatin because it provided a consistent and reproducible shearing across samples and tissue types. A limitation of the micrococcal nuclease may be increased likelihood of the appearance of duplicated reads due to similarity of cut sites in the chromatin; however, several studies have not found substantial bias when duplicates were removed (Allan et al., 2012; David et al., 2017; Gutiérrez et al., 2017; Chereji et al., 2019). Furthermore, shearing with micrococcal nuclease to approximately 1–2 nucleosome lengths may contribute to slightly different characteristics, including width, of peaks called from these experiments.

Sequence read pileups were examined in IGV near genes known to be active and inactive in humans and expected

TABLE 5 | Top three *de novo* CTCF motifs present in each sample in the cerebellum.

Animal	*De novo* motif	*P* value	Percent of target sequences with motif	Percent of background sequences with motif
F1		1E-910	2.92%	0.01%
		1E-756	2.50%	0.01%
		1E-735	2.44%	0.01%
F2		1E-1078	3.49%	0.04%
		1E-875	1.85%	0.00%
		1E-842	2.01%	0.01%
M1		1E-946	1.42%	0.00%
		1E-800	1.24%	0.01%
		1E-697	1.25%	0.01%
M2		1E-677	1.02%	0.00%
		1E-565	0.88%	0.01%
		1E-509	0.80%	0.01%

to be conserved across species. This provided a means of examining genes with known promoters and expression patterns as positive and negative controls for both ChIP-seq experiments and WGBS and provided insight into the potential similarity of regulatory elements across species. Several genes known to be active across different mammalian species in the liver, spleen, and cerebellum showed a sequence pileup of active histone marks which likely indicated the presence of active regulatory elements. Inversely, genes known to be active during meiotic processes and quiescent during adult stages in several mammalian species showed no sequence pileup of histone marks and presence of DNA methylation, which likely indicates inactivity of regulatory elements.

Consistency of regulatory element identification by ChIP-seq and DNA methylation for each tissue between the four individual animals was evaluated by calculating Spearman and Pearson correlations, respectively. Correlations between samples for both ChIP-seq and DNA methylation were within the ranges previously reported with sequence data (Peng et al., 2010; Siska and Kechris, 2017). Furthermore, correlations between ChIP-seq biological replicates have been reported as low as 0.3–0.4, with technical replicates reported as high as 0.9 (Friedman and Alm, 2012; Siska and Kechris, 2017). The results for these sheep tissues therefore achieve equivalent or improved results compared with previously reported pipelines for regulatory element identification and characterization and demonstrate a tissue-specific moderate variation across biological replicates. The spleen displayed the highest variation between biological replicates, with correlations between 0.44 and 0.84 among histone marks, although DNA methylation was consistent across replicates including the spleen. Given that splenic tissue is an acutely responsive immunological tissue, perhaps it is not surprising that we observed greater variation in the biological replicates.

The CG methylation signal for all four samples clustered distinctly by tissue in a principal component analysis, indicating clear differences in DNA methylation between tissues (**Figure 3**). This finding is supported by others that have shown that the greatest differences in methylation occur between tissue types rather than between individuals (Pai et al., 2011; Zhang et al., 2013) and consistent with the requirement for a particular set of genes to be active and therefore demethylated depending on tissue function. Cerebellum samples demonstrated a higher level of both CG and non-CG methylation compared to the liver and spleen. Brain tissues are known to differ from other tissues in methylation patterns in other species, and furthermore, the cerebellum has been shown to be different than other brain tissues (Gibbs et al., 2010; Cantrell et al., 2019).

The enrichment of individual histone marks was examined by identifying peaks in each sample. The number of peaks identified in these sheep liver, spleen, and cerebellum samples was consistent with other studies in sheep adipose, cattle liver, cattle muscle, cattle rumen epithelium, human liver, and mouse liver (**Supplementary Table 3**) (Villar et al., 2015; Zhao et al., 2015; Naval-Sanchez et al., 2018; Fang et al., 2019). Many chromosomes had differences in peak numbers normalized by chromosome length between tissues, indicating potential tissue specificity of some peaks. Narrow marks H3K3me3, H3K27ac, and CTCF had a shorter average width than broad marks H3K4me1 and H3K27me3, which may be influenced by the program and statistical model used to call peaks as well as by the shearing method (Zhang et al., 2008, 2009). Because micrococcal nuclease was used for shearing, the length of the narrow peaks more closely resembles the size of a single nucleosome.

Trimethylation of histone 3 lysine 4 peaks were enriched annotated TSS in all three tissues. The peaks and heatmap signature signals are similar to several other ChIP-seq experiments in human PBMCs and CD14+ cells, as well as mouse liver (Schones et al., 2008; Quinodoz et al., 2014; Uchiyama et al., 2018). Peaks from all histone modifications and CTCF were also annotated with regions defined in the *Oar_rambouillet_v1.0* genome. In the liver, spleen, and cerebellum, the most TSS were identified near (within 2 kb of distance on either side) to H3K4me3 peaks, which is not surprising. Many H3K27ac and H3K4me1 peaks, which indicate the presence of active or poised enhancers, were located in intronic regions. Repressed enhancers marked by H3K27me3 were located mostly in intergenic regions, along with CTCF, which may be indicative of insulated TAD boundaries not in close proximity of genes. Further work with additional animals in combination with RNA expression and TSS analyses is needed to examine regulatory element activity outside of previously annotated regions of the sheep genome.

The genomic segments identified by histone mark peaks were evaluated for overlap between marks and CTCF binding. This broader view of the regulatory landscape lends a better understanding of gene regulation at each location than individual

marks (Park, 2018). Active promoters have been shown to exhibit greater enrichment of H3K4me3 than other histone marks in addition to the often present H3K27ac (Wang et al., 2008; Creyghton et al., 2010; Carelli et al., 2018). However, if lysine 4 is monomethylated (H3K4me1), this indicates the presence of a poised enhancer, in which enrichment of lysine 27 can be acetylated or trimethylated depending on the state and activity of the enhancer (Heintzman et al., 2007; Wang et al., 2008; Creyghton et al., 2010; Carelli et al., 2018). Low H3K4me3 coincident with high H3K27ac signal has been reported to be common at enhancers near genes undergoing highly active transcription (Core et al., 2014; Carelli et al., 2018). Repressed enhancers are generally characterized by H3K27me3 signal (Carelli et al., 2018). However, H3K27me3 has also been shown to be enriched near the promoter or gene body in genes being expressed at a relatively low rate (Young et al., 2011; Flensburg et al., 2014). The chromatin states characterized in this study are similar to what others have previously described in cattle (Fang et al., 2019). Furthermore, the weak poised enhancer category detected in the spleen and the weak repressed enhancer category detected in the cerebellum demonstrate that different tissues may have varying chromatin states, which supports the importance of characterizing chromatin states across tissues within a species.

Hypermethylated and hypomethylated regions of the sheep genome were defined across liver, spleen, and cerebellum tissues. The number of hypermethylated and hypomethylated regions per Mb in each of the nine chromatin states was quantified. The data presented in this study demonstrate an enrichment of hypermethylated regions in chromatin states with prominent H3K4me1 (primarily poised enhancers) and hypomethylated regions in active enhancers and promoters enriched with H3K27ac and H3K4me3. These results agree with previous research in humans and mice which indicate that active enhancer activity is inversely correlated with DNA methylation (Aran and Hellman, 2013; Barwick et al., 2016; Bell and Vertino, 2017). Interestingly, the presence of H3K4me1 was found to be positively correlated with DNA methylation, specifically intermediate methylation (25–75%), in mice (Zhang et al., 2009; Teng and Tan, 2012; Sharifi-Zarchi et al., 2017). Furthermore, enhancers enriched with H3K27ac and promoters enriched with H3K4me3 had less DNA methylation than other regions (Sharifi-Zarchi et al., 2017).

Approximately 20% of the sheep genome was assigned to a chromatin state category including promoters; active, poised, and repressed enhancers; and CTCF in the liver, spleen, and cerebellum. In cattle, a previous study similarly assigned approximately 30% of the genome to either a chromatin state or areas with open chromatin in rumen epithelium (Fang et al., 2019). The locations of many regulatory elements were similar between the liver and spleen in this study; however, a greater difference was observed in active enhancers and promoters between the cerebellum compared with the liver and spleen. Since distinct differences in gene expression and regulation have been observed between the cerebellum and other tissues in sheep, this difference is not surprising (Jiang et al., 2014).

The CCCTC-binding factor (CTCF) along with cohesins was shown to be present at the boundaries of TADs in humans and mice (Dixon et al., 2012; Phillips-Cremins et al., 2013; Rao et al., 2014; Vietri Rudan et al., 2015; Szabo et al., 2019). Depending on the cell type, 75–95% of TAD boundaries defined by Hi-C chromatin capture have shown CTCF signal in mice (Bonev et al., 2017; Szabo et al., 2019). The chromatin states in this study that display primarily CTCF could be representative of these domain boundaries; however, Hi-C data are required to confirm which will be possible for the data produced in the FAANG study of the reference ewe, where Hi-C data are also available. In addition to helping define TAD boundaries, CTCF has also been identified near enhancers and promoters within TADs in humans and mice, which then form smaller loop domains with cohesins and the protein YY1 (Phillips-Cremins et al., 2013; Weintraub et al., 2017; Szabo et al., 2019). The chromatin state analysis may be detecting some of these within-TAD loop interactions, with overlap between CTCF and H3K27me3 as well as H3K4me1 signal shown in the chromatin state heatmaps in the liver and cerebellum. Signal from CTCF, H3K27me3, and H3K4me1 marks within one chromatin state was also observed in another study in cattle rumen epithelial tissue and Madin–Darby bovine kidney epithelial cells (Fang et al., 2019).

Motif analysis of CTCF resulted in both known and *de novo* motifs identified in more than one tissue. A large number of CTCF-binding motifs are similar in sequence across mammalian species including cattle (Filippova et al., 1996; Schmidt et al., 2012; Wang et al., 2018). Wang and associates identified putative CTCF-binding motifs in the bovine genome with 82 CTCF motif profiles with similar sequence in human, mouse, dog, and macaque (Schmidt et al., 2012; Wang et al., 2018). In this study, significant motifs identified in ovine liver, spleen, and cerebellum were also identified in human, mouse, fly (*Drosophila melanogaster*), and yeast (*Saccharomyces cerevisiae*) within the HOMER motif database (Heinz et al., 2010).

This experiment examines regulatory elements in multiple sheep tissues and individuals with ChIP-seq and WGBS methylation assays. These data provide putative categories of biological functions for regulatory DNA and will facilitate the identification of epigenetic variation that controls phenotypic traits in sheep. Epigenetic annotation is especially important for revealing the biology behind interesting complex traits since genetic variation does not always reveal the entire story. Epigenetic variation may play a larger role in traits uniquely expressed in a specific tissue or recently evolved rare traits. Identification of causal regulatory variants will allow more rapid genetic improvement for health and production traits in the meat, milk, and wool industries across sheep populations. Causal variants have the highest utility across breeds and allow more efficient assimilation of genetic markers into marker-assisted selection and genomic selection algorithms. The protocols and analysis pipeline optimized here for validation and the eventual annotation of DNA regulatory elements are valuable resources for the Ovine FAANG Project Consortium and the International Sheep Genomics Consortium.

MEMBERS OF THE OVINE FAANG PROJECT CONSORTIUM (LISTED BY INSTITUTION)

Brenda Murdoch (University of Idaho)
Kimberly Davenport (University of Idaho)
Stephen White (USDA, ARS, ADRU, Washington State University)
Michelle Mousel (USDA, ARS, ADRU, Washington State University)
Alisha Massa (Washington State University)
Kim Worley (Baylor College of Medicine)
Alan Archibald (The Roslin Institute, University of Edinburgh)
Emily Clark (The Roslin Institute, University of Edinburgh)
Brian Dalrymple (University of Western Australia)
James Kijas (CSIRO)
Shannon Clarke (AgResearch)
Rudiger Brauning (AgResearch)
Timothy Smith (USDA, ARS, MARC)
Tracey Hadfield (Utah State University)
Noelle Cockett (Utah State University)

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 BioProject, accession no: PRJEB41457.

ETHICS STATEMENT

Sheep were housed and cared for under Animal Subject Approval Form (ASAF) 04618 titled "Maintenance of MCF and OPPV free sheep flocks" approved by the Washington State University Institutional Animal Care and Use Committee (IACUC) and euthanized for tissue collection under ASAF 6003.

AUTHOR CONTRIBUTIONS

MM, MH, SW, SM, TS, and BM designed the study. KD, AM, MM, MH, and BM collected the samples. BM, KD, MM, SW, AM, SM, and SB facilitated the ChIP-Seq and methylation experiments, data analyses, and drafted the manuscript. TS facilitated ChIP-seq library preparation and sequencing. KD, AM, SB, SM, MM, MH, SW, TS, and BM discussed and interpreted results. SM, MM, SW, NC, TS, and BM acquired funding. All authors contributed to the article and approved the final version.

FUNDING

This project was supported by Agriculture and Food Research Initiative grant no. 2017-67016-26301, Hatch grant no. IDA01566 from the USDA National Institute of Food and Agriculture, USDA-NIFA-2016-67016-24766 for DNA methylation, USDA-ARS Project 3040-31000-100-00-D for sequencing, and USDA-ARS Project 2090-32000-36-00-D for the care and management of sheep.

ACKNOWLEDGMENTS

The authors would like to thank Codie Durfee and Kaneesha Hemmerling for assistance with tissue collection; Jacky Carnahan, Kristen Kuhn, and Kelsey McClure for assistance with sequencing; and Mingrui Duan for assistance with the bioinformatics analysis pipeline.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.628849/full#supplementary-material

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, 537–544. doi: 10.1093/nar/gky379

Albert, F. W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16, 197–212. doi: 10.1038/nrg3891

Al-Mamun, H. A., Clark, S. A., Kwan, P., and Gondro, C. (2015). Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep. *Genet. Sel. Evol.* 47:90. doi: 10.1186/s12711-015-0169-6

Allan, J., Fraser, R. M., Owens-Hughes, T., and Keszenman-Pereyra, D. (2012). Micrococcal nuclease does not substantially bias nucleosome mapping. *J. Mol. Biol.* 417, 152–164. doi: 10.1016/j.jmb.2012.01.043

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project. *Gen. Biol.* 16:57. doi: 10.1186/s13059-015-0622-4

Aran, D., and Hellman, A. (2013). DNA methylation of transcriptional enhancers and cancer predisposition. *Cell* 154, 11–13. doi: 10.1016/j.cell.2013.06.018

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837. doi: 10.1016/j.cell.2007.05.009

Barwick, B.G., Scharer, C.D., Bally, A.P., and Boss, J.M. (2016). Plasma cell differentiation is coupled to division-dependent DNA hypomethylation and gene regulation. *Nat. Immunol.* 17, 1216–1225. doi: 10.1038/ni.3519

Beagan, J. A., and Phillips-Cremins, J. E. (2020). On the existence and functionality of topologically associating domains. *Nat. Genet.* 52, 8–16. doi: 10.1038/s41588-019-0561-1

Bell, J. S. K., and Vertino, P. M. (2017). Orphan CpG islands define a novel class of highly active enhancers. *Epigenetics* 12, 449–464. doi: 10.1080/15592294.2017.1297910

Bernardi, M., Maggioli, C., and Zaccherini, G. (2012). Human albumin in the management of complications of liver cirrhosis. *Crit. Care* 16:211. doi: 10.1186/cc11218

Bonev, B., Cohen, N. M., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., et al. (2017). Multiscale 3D genome rewiring during mouse neural development. *Cell* 171, 557–572. doi: 10.1016/j.cell.2017.09.043

Broad Institute (2019). *Picard Toolkit*. Cambridge, MA: Broad Institute.

Cantrell, B., Lachance, H., Murdoch, B., Sjoquist, J., Funston, R., Weaber, R., et al. (2019). Global DNA methylation in the limbic system of cattle. *Epigenomes* 3:8. doi: 10.3390/epigenomes3020008

Carelli, F. N., Liechti, A., Halbert, J., Warnefors, M., and Kaessmann, H. (2018). Repurposing of promoters and enhancers during mammalian evolution. *Nat. Commun.* 9:4066. doi: 10.1038/s41467-018-06544-z

Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* 10, 295–304. doi: 10.1038/nrg2540

Chereji, R. V., Bryson, T. D., and Henikoff, S. (2019). Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biol.* 20:198. doi: 10.1186/s13059-019-1815-z

Clark, E. L., Bush, S. J., McCulloch, M., Farquhar, I. L., Young, R., Lefevre, L., et al. (2017). A high resolution atlas of gene expression in the domestic sheep (Ovis aries). *PLoS Genet.* 13:e1006997. doi: 10.1371/journal.pgen.1006997

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., and Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* 46, 1311–1320. doi: 10.1038/ng.3142

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* 107, 21931–21936. doi: 10.1073/pnas.1016071107

David, S. A., Piégu, B., Hennequet-Antier, C., Pannetier, M., Aguirre-Lavin, T., Crochet, S., et al. (2017). An assessment of fixed and native chromatin preparation methods to study histone post-translational modifications at a whole genome scale in skeletal muscle tissue. *Biol. Proc. Online* 19:10. doi: 10.1186/s12575-017-0059-0

Deaton, A. M., and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev.* 25, 1010–1022. doi: 10.1101/gad.2037511

Divya, T., Lalitha, S., Parvathy, S., Subashini, C., Sanalkumar, R., Bindu Dhanesh, S., et al. (2016). Regulation of Tlx3 by Pax6 is required for the restricted expression of Chrnα3 in cerebellar granule neuron progenitors during development. *Sci. Rep.* 6:30337. doi: 10.1038/srep30337

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

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* 28, 817–825. doi: 10.1038/nbt.1662

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–216. doi: 10.1038/nmeth.1906

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* 12, 2478–2492. doi: 10.1038/nprot.2017.124

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., et al. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteom.* 13, 397–406. doi: 10.1074/mcp.M113.035600

Fang, L., Liu, S., Liu, M., Kang, X., Lin, S., Li, B., et al. (2019). Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biol.* 17:68. doi: 10.1186/s12915-019-0687-8

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* 7, 1728–1740. doi: 10.1038/nprot.2012.101

Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., et al. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol. Cell. Biol.* 16, 2802–2813. doi: 10.1128/MCB.16.6.2802

Flensburg, C., Kinkel, S. A., Keniry, A., Blewitt, M. E., and Oshlack, A. (2014). A comparison of control samples for ChIP-seq of histone modifications. *Front. Genet.* 5:329. doi: 10.3389/fgene.2014.00329

Frain, M., Hardon, E., Ciliberto, G., and Sala-Trepat, J. M. (1990). Binding of a liver-specific factor to the human albumin gene promoter and enhancer. *Mol. Cell. Biol.* 10, 991–999. doi: 10.1128/mcb.10.3.991

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687

Ghirlando, R., and Felsenfeld, G. (2016). CTCF: making the right connections. *Genes Dev.* 30, 881–891. doi: 10.1101/gad.277863.116

Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S. L., et al. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in the human brain. *PLoS Genet.* 6:e1000952. doi: 10.1371/journal.pgen.1000952

Giuffra, E., Tuggle, C. K., and Faang Consortium (2019). Functional Annotation of Animal Genomes (FAANG): current achievements and roadmap. *Ann. Rev. Anim. Biosci.* 7, 65–88. doi: 10.1146/annurev-animal-020518-114913

Gorkin, D. U., Barozzi, I., Zhang, Y., Lee, A. Y., Zhao, Y., Wildberg, A., et al. (2017). Systematic mapping of chromatin state landscapes during mouse development. *bioRxiv* [Preprint] doi: 10.1101/166652

Gorkin, D. U., Barozzi, I., Zhao, Y., Zhang, Y., Huang, H., Lee, A. Y., et al. (2020). An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* 583, 744–751. doi: 10.1038/s41586-020-2093-3

Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M. Q., et al. (2013). BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* 14:774. doi: 10.1186/1471-2164-14-774

Guo, W., Zhu, P., Pellegrini, M., Zhang, M. Q., Wang, X., and Ni, Z. (2018). CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics* 34, 381–387. doi: 10.1093/bioinformatics/btx595

Gutiérrez, G., Millán-Zambrano, G., Medina, D. A., Jordán-Pla, A., Pérez-Ortín, J. E., Peñate, X., et al. (2017). Subtracting the sequence bias from partially digested MNase-seq data reveals a general contribution of TFIIS to nucleosome positioning. *Epigenetics Chromatin* 10:58. doi: 10.1186/s13072-017-0165-x

Ha, T., Swanson, D., Larouche, M., Glenn, R., Weeden, D., Zhang, P., et al. (2015). CbGRiTs: cerebellar gene regulation in time and space. *Dev. Biol.* 397, 18–30. doi: 10.1016/j.ydbio.2014.09.032

Hayashi, Y., Chan, J., Nakabayashi, H., Hashimoto, T., and Tamaoki, T. (1992). Identification and characterization of two enhancers of the human albumin gene. *J. Biol. Chem.* 267, 14580–14585.

Hedges, J. F., Kimmel, E., Snyder, D. T., Jerome, M., and Jutila, M. A. (2013). Solute carrier 11A1 is expressed by innate lymphocytes and augments their activation. *J. Immunol.* 190, 4263–4273. doi: 10.4049/jimmunol.1200732

Hegde, N. G. (2019). Livestock development for sustainable livelihood of small farmers. *Asian J. Res. Anim. Vet. Sci.* 3, 1–17.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318. doi: 10.1038/ng1966

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Jenuwein, T., and Allis, C. D. (2001). Translating the histone code. *Science* 293, 1074–1080. doi: 10.1126/science.1063127

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J. F., Faraut, T., et al. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* 344, 1168–1173. doi: 10.1126/science.1252806

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316, 1497–1502. doi: 10.1126/science.1141319

Kingsley, N. B., Kern, C., Creppe, C., Hales, E. N., Zhou, H., Kalbfleisch, T. S., et al. (2019). Functionally annotating regulatory elements in the equine genome using histone mark ChIP-Seq. *Genes* 11:3. doi: 10.3390/genes11010003

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22, 1813–1831. doi: 10.1101/gr.136184.111

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.

Lee, B. K., and Iyer, V. R. (2012). Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. *J. Biol. Chem.* 287, 30906–30913. doi: 10.1074/jbc.R111. 324962

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph*. 20, 1983–1992. doi: 10.1109/TVCG.2014.2346248

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Meadows, J. R., Chan, E. K., and Kijas, J. W. (2008). Linkage disequilibrium compared between five populations of domestic sheep. *BMC Genet.* 9:61. doi: 10.1186/1471-2156-9-61

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454, 766–770. doi: 10.1038/nature07107

Micsinai, M., Parisi, F., Strino, F., Asp, P., Dynlacht, B. D., and Kluger, Y. (2012). Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res.* 40:e70. doi: 10.1093/nar/gks048

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560. doi: 10.1038/nature06008

Naval-Sanchez, M., Nguyen, Q., McWilliam, S., Porto-Neto, L. R., Tellam, R., Vuocolo, T., et al. (2018). Sheep genome functional annotation reveals proximal regulatory elements contributed to the evolution of modern breeds. *Nat. Commun.* 9:859. doi: 10.1038/s41467-017-02809-1

Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K., and Gilad, Y. (2011). A genome-wide study of DNA Methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.* 7:e1001316. doi: 10.1371/journal.pgen.1001316

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet*. 10, 669–680. doi: 10.1038/nrg2641

Park, H. S. (2018). A short report on the markov property of DNA sequences on 200-bp genomic units of roadmap genomics ChromHMM annotations: a computational perspective. *Genomics Inform.* 16:e27. doi: 10.5808/GI.2018.16.4.e27

Pauler, F. M., Sloane, M. A., Huang, R., Regha, K., Koerner, M. V., Tamir, I., et al. (2009). H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res.* 19, 221–233. doi: 10.1101/gr.080861.108

Peng, S., Kuroda, M. I., and Park, P. J. (2010). Quantized correlation coefficient for measuring reproducibility of ChIP-chip data. *BMC Bioinformatics* 11:399. doi: 10.1186/1471-2105-11-399

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* 153, 1281–1295. doi: 10.1016/j.cell.2013.04.053

Quinlan, A. R. (2014). BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* 47, 11.12.1–34. doi: 10.1002/0471250953.bi1112s47

Quinodoz, M., Naef, F., Gobet, C., and Gustafson, K. (2014). Characteristic bimodal profiles of RNA polymerase II at thousands of active mammalian promoters. *Genome Biol.* 15:R85. doi: 10.1186/gb-2014-15-6-r85

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* 42, 187–191. doi: 10.1093/nar/gku365

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

Rexroad, C., Vallet, J., Matukumalli, L. K., Reecy, J., Bickhart, D., Blackburn, H., et al. (2019). Genome to phenome: improving animal health, production, and well-being - a new USDA blueprint for animal genome research 2018–2027. *Front. Genet.* 10:327. doi: 10.3389/fgene.2019.00327

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4, 651–657. doi: 10.1038/nmeth1068

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Salavati, M., Caulton, A., Clark, R., Gazova, I., Smith, T. P. L., Worley, K. C., et al. (2020). Global analysis of transcription start sites in the new ovine reference genome (*Oar rambouillet v1.0*). *Front. Genet.* 11:580580. doi: 10.3389/fgene.2020.580580

Schmidt, D., Schwalie Petra, C., Wilson Michael, D., Ballester, B., Gonçalves, Â, Kutter, C., et al. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 148, 335–348. doi: 10.1016/j.cell.2011.11.058

Schones, D. E., Cui, K., Cuddapah, S., Roh, T. Y., Barski, A., Wang, Z., et al. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132, 887–898. doi: 10.1016/j.cell.2008.02.022

Sharifi-Zarchi, A., Gerovska, D., Adachi, K., Totonchi, M., Pezeshk, H., Taft, R.J., et al. (2017). DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. *BMC Genomics.* 18, 964. doi: 10.1186/s12864-017-4353-7

Siska, C., and Kechris, K. (2017). Differential correlation for sequencing data. *BMC Res. Notes* 10:54. doi: 10.1186/s13104-016-2331-9

Song, Q., Decato, B., Hong, E. E., Zhou, M., Fang, F., Qu, J., et al. (2013). A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* 8:e81148. doi: 10.1371/journal.pone.0081148

Szabo, Q., Bantignies, F., and Cavalli, G. (2019). Principles of genome folding into topologically associating domains. *Sci. Adv.* 5:1668. doi: 10.1126/sciadv.aaw1668

Teng, L., and Tan, K. (2012). Finding combinatorial histone code by semi-supervised biclustering. *BMC Genomics* 13:301. doi: 10.1186/1471-2164-13-301

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

Thomas, R., Thomas, S., Holloway, A. K., and Pollard, K. S. (2017). Features that define the best ChIP-seq peak calling algorithms. *Brief. Bioinform.* 18, 441–450. doi: 10.1093/bib/bbw035

Tuggle, C. K., Giuffra, E., White, S. N., Clarke, L., Zhou, H., Ross, P. J., et al. (2016). GO-FAANG meeting: a gathering on functional annotation of animal genomes. *Anim. Genet.* 47, 528–533. doi: 10.1111/age.12466

Uchiyama, R., Kupkova, K., Shetty, S. J., Linford, A. S., Pray-Grant, M. G., Wagar, L. E., et al. (2018). Histone H3 lysine 4 methylation signature associated with human undernutrition. *Proc. Natl. Acad. Sci. U.S.A.* 115, E11264–E11273. doi: 10.1073/pnas.1722125115

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., et al. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* 10, 1297–1309. doi: 10.1016/j.celrep.2015.02.004

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., et al. (2015). Enhancer evolution across 20 mammalian species. *Cell* 160, 554–566. doi: 10.1016/j.cell.2015.01.006

Wang, M., Hancock, T. P., Chamberlain, A. J., Vander Jagt, C. J., Pryce, J. E., Cocks, B. G., et al. (2018). Putative bovine topological association domains and CTCF binding motifs can reduce the search space for causative regulatory variants of complex traits. *BMC Genomics* 19:395. doi: 10.1186/s12864-018-4800-0

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* 40, 897–903. doi: 10.1038/ng.154

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

Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hannett, N. M., Day, D. S., et al. (2017). YY1 is a structural regulator of enhancer-promoter loops. *Cell* 171, 1573–1588.e28. doi: 10.1016/j.cell.2017.11.008

Xiang, R., Berg, I., MacLeod, I. M., Hayes, B. J., Prowse-Wilkins, C. P., Wang, M., et al. (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl. Acad. Sci. U.S.A.* 116, 19398–19408. doi: 10.1073/pnas.1904159116

Xu, H., Beasley, M. D., Warren, W. D., van der Horst, G. T., and McKay, M. J. (2005). Absence of mouse REC8 cohesin promots synapsis of sister chromatids in meiosis. *Dev. Cell* 8, 949–961.

Young, E. T., and Sinsheimer, R. L. (1965). A comparison of the initial actions of spleen deoxyribonuclease and pancreas deoxyribonuclease. *J. Biol. Chem.* 240, 1274–1280.

Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., et al. (2011). ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.* 39, 7415–7427. doi: 10.1093/nar/gkr416

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 25, 1952–1958. doi: 10.1093/bioinformatics/btp340

Zhang, B., Zhou, Y., Lin, N., Lowdon, R. F., Hong, C., Nagarajan, R. P., et al. (2013). Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Res.* 23, 1522–1540. doi: 10.1101/gr.156539.113

Zhang, X., Bernatavichute, Y. V., Cokus, S., Pellegrini, M., and Jacobsen, S. E. (2009). Genomewide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis Thaliana*. *Genome Biol.* 10:R62. doi: 10.1186/gb-2009-10-6-r62

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

Zhao, C., Carrillo, J. A., Tian, F., Zan, L., Updike, S. M., Zhao, K., et al. (2015). Genome-wide H3K4me3 analysis in angus cattle with divergent tenderness. *PLoS One* 10:e0115358. doi: 10.1371/journal.pone.0115358

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.

Copyright © 2021 Davenport, Massa, Bhattarai, McKay, Mousel, Herndon, White, Cockett, Smith and Murdoch. 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: 16 June 2021
doi: 10.3389/fgene.2021.641788

Successful ATAC-Seq From Snap-Frozen Equine Tissues

*Sichong Peng[1], Rebecca Bellone[1,2], Jessica L. Petersen[3], Theodore S. Kalbfleisch[4] and Carrie J. Finno[1]**

[1] Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, Davis, CA, United States, [2] Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California, Davis, Davis, CA, United States, [3] Department of Animal Science, University of Nebraska-Lincoln, Lincoln, NE, United States, [4] Department of Veterinary Science, Gluck Equine Research Center, University of Kentucky, Lexington, KY, United States

An assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) has become an increasingly popular method to assess genome-wide chromatin accessibility in isolated nuclei from fresh tissues. However, many biobanks contain only snap-frozen tissue samples. While ATAC-seq has been applied to frozen brain tissues in human, its applicability in a wide variety of tissues in horse remains unclear. The Functional Annotation of Animal Genome (FAANG) project is an international collaboration aimed to provide high quality functional annotation of animal genomes. The equine FAANG initiative has generated a biobank of over 80 tissues from two reference female animals and experiments to begin to characterize tissue specificity of genome function for prioritized tissues have been performed. Due to the logistics of tissue collection and storage, extracting nuclei from a large number of tissues for ATAC-seq at the time of collection is not always practical. To assess the feasibility of using stored frozen tissues for ATAC-seq and to provide a guideline for the equine FAANG project, we compared ATAC-seq results from nuclei isolated from frozen tissue to cryopreserved nuclei (CN) isolated at the time of tissue harvest in liver, a highly cellular homogenous tissue, and lamina, a relatively acellular tissue unique to the horse. We identified 20,000–33,000 accessible chromatin regions in lamina and 22–61,000 in liver, with consistently more peaks identified using CN isolated at time of tissue collection. Our results suggest that frozen tissues are an acceptable substitute when CN are not available. For more challenging tissues such as lamina, nuclei extraction at the time of tissue collection is still preferred for optimal results. Therefore, tissue type and accessibility to intact nuclei should be considered when designing ATAC-seq experiments.

Keywords: FAANG, horse, cryopreserved, chromatin, epigenetics

OPEN ACCESS

Edited by:
Christopher K. Tuggle, Iowa State University, United States

Reviewed by:
Michelle Halstead, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), France
Eveline M. Ibeagha-Awemu, Agriculture and Agri-Food Canada (AAFC), Canada

***Correspondence:**
Carrie J. Finno
cjfinno@gmail.com; cjfinno@ucdavis.edu

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

Received: *14 December 2020*
Accepted: *21 May 2021*
Published: *16 June 2021*

Citation:
Peng S, Bellone R, Petersen JL, Kalbfleisch TS and Finno CJ (2021) Successful ATAC-Seq From Snap-Frozen Equine Tissues. Front. Genet. 12:641788. doi: 10.3389/fgene.2021.641788

INTRODUCTION

The completion of the equine genome assembly (Wade et al., 2009; Kalbfleisch et al., 2018) has enabled research leading to novel discoveries concerning the health and reproduction of horses (Finno and Bannasch, 2014; Ghosh et al., 2018; Raudsepp et al., 2019). However, despite having the same genomic sequence, differential regulation of gene expression leads to tissue-specific profiles. A lack of understanding of gene regulation has largely stalled research of complex traits in horses.

In humans and mice, the Encyclopedia of DNA Elements (ENCODE) project has provided an abundance of data for understanding gene regulation and its role in complex diseases and traits (Qu and Fang, 2013). Unfortunately, limited resources are currently available in the horse. The Functional Annotation of Animal Genome (FAANG) initiative (The FAANG Consortium et al., 2015) is an international collaboration aimed to bridge this gap between genotype and phenotype. The equine FAANG project has successfully generated a biobank of over 80 tissues and bodily fluids of two reference animals (Burns et al., 2018). RNA-seq of 32 tissues (unpublished, data access: PRJEB26787), as well as the identification of tissue specific histone marks for eight prioritized tissues (Kingsley et al., 2019), from this biobank has been performed. Additional projects are underway to identify tissue specific chromatin states to integrate all of these datasets and build a robust tissue specific functional annotation atlas in the horse (Giuffra et al., 2019).

An important component of gene expression and regulation is chromatin accessibility. Active genes and regulatory elements are typically found within or near regions of the DNA accessible to transcription factors. Therefore, identifying open chromatin regions is a crucial step to identify and categorize tissue specific regulatory elements in order to advance our understanding of complex traits in the horse. An assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) (Buenrostro et al., 2015) is commonly used to identify regions of open chromatin. A typical ATAC-seq protocol requires nuclei extracted from fresh tissues. Halstead et al. (2020b) proposed a modified ATAC-seq protocol to allow long-term storage of cryopreserved nuclei (CN) extracted from fresh tissues. Still, the intensive efforts needed to prepare and cryopreserve nuclei during a large-scale tissue collection prove to be difficult. Alternatively, Corces et al. (2017) successfully applied a modified ATAC-seq (Omni-ATAC) protocol on frozen human brain tissues. However, the applicability of Omni-ATAC has not been tested in a wide variety of tissues in horse where nuclei extraction may prove challenging. Additionally, it has been shown that in cultured cells cryopreservation is preferable to flash-freezing process in order to preserve native chromatin structures (Milani et al., 2016). To our knowledge, no studies have investigated the effect of snap freezing on tissues for ATAC-seq library generation in comparison to CN preps. Additionally, the library preparation step is a major source of variation in RNA-seq studies (McIntyre et al., 2011), particularly at low read depth. As a result, RNA-seq data generated from different laboratories or at different times cannot often be directly compared. For a collaborative project, it is important to assess the effect of technical variations to better inform project planning and analytical decisions for data integration.

To address these gaps of knowledge in the applicability of ATAC-seq in snap-frozen horse tissues, and to provide a guide for future ATAC-seq studies to assess chromatin accessibility, we compared data from CN prepared from fresh tissue to that of nuclei extracted from snap-frozen tissues collected from the two mares from the initial equine FAANG biobank study (Burns et al., 2018). In order for this comparison to be informative and applicable to a wide range of tissues, we utilized both liver, a highly cellular and homogenous tissue type, and lamina, a relatively acellular tissue unique to the horse. Equine laminae are highly vascularized interdigitated dermal and epidermal tissues in the equine foot that form the attachment between the hoof wall and the third phalanx. Inflammation of laminae in horses (i.e., laminitis) is a devastating disease that impacts many breeds of horses and often leads to euthanasia. Therefore, gene regulation in laminae is of particular interest to equine geneticists and veterinary practitioners as this debilitating and life-threatening disease estimated to impact up to 34% of the horse population (Wylie et al., 2011). Laminitis is also the primary clinical consequence of equine metabolic syndrome (EMS) (Durham et al., 2019). EMS is a complex syndrome that requires constant veterinarian care and diet control, impacting an estimated 18 to 27 percent of horse population (Durham et al., 2019). Liver is the primary metabolic organ with a homogenously cellular structure. Detailed knowledge of gene expressions and regulations in healthy liver provides a baseline for studying impaired metabolism in horses with EMS. Additionally, to assess the effect of library preparation techniques, snap-frozen tissues and CN from this pilot study were sent to two different core laboratories for library generation and subsequent sequencing. We hypothesized that (1) ATAC-seq using frozen tissues would identify comparable peaks to those using CN from fresh tissues, (2) libraries generated from liver will have better quality than those from laminae, and (3) similar to what was found in RNA-seq studies there will be a significant amount of variation between the libraries generated by two laboratories.

MATERIALS AND METHODS

Tissue Collection and Nuclei Isolation

Liver and lamina tissues from two mares (AH2 and AH1) were collected as described in Burns et al. (2018). Briefly, two healthy adult Thoroughbred mares (AH1: 5 years old; AH2: 4 years old) were closely examined by veterinarians prior to tissue collection. Nuclei were isolated from liver and lamina tissues immediately following tissue collection and cryopreserved following protocols published in Halstead et al. (2020a) with some modifications for lamina. Briefly, additional incubation periods with collagenase were added to assist in homogenization (see **Supplementary Material**). These are referred to as CN. Additionally, at time of collection, approximately 1 g aliquots of tissue were snap frozen in liquid nitrogen for nuclei extraction at a later time. These are referred to as frozen tissue-derived nuclei (FTDN).

ATAC-Seq Library Preparation and Sequencing

Both snap frozen tissues and CN were stored at −80°C for 3 years until shipped on dry ice overnight to two commercial laboratories (L1 and L2) for library preparation. Nuclei were extracted from frozen tissues using each laboratory's internally optimized protocol (see **Supplementary Material**). Extracted Nuclei (FTDN) and CN were used to prepare ATAC libraries (**Supplementary Methods** and **Supplementary Table 1**). Libraries were sequenced on an Illumina HiSeq 4000, paired-end

2 × 75 bp (L1) or NextSeq 500, paired-end 2 × 42 bp (L2) with a targeted depth of 30 million read pairs.

ATAC-Seq Data Analysis

Read QC was carried out using FastQC (Andrews, 2010). Adapters and low-quality ends were trimmed using TrimGalore (Krueger, 2019) and Cutadapt (Martin, 2011). Reads were then aligned to reference genome EquCab3 using BWA-MEM algorithm from BWA (Li and Durbin, 2009) using default parameters. Post-alignment filtering was employed to remove low mapping quality reads, mitochondrial reads, and PCR duplicates using Samtools (Li et al., 2009) and Sambamba (Tarasov et al., 2015). Genome coverage was analyzed using deepTools (Ramírez et al., 2016). Specifically, bamCoverage was used to convert bam files to bigwig files, using RPKM to normalize coverage with exact scaling (–normalizeUsing RPKM –exactScaling). Then multibigwigSummary was used to calculate average coverage across 1,000 bp windows (-bs 1,000). plotPCA was used to calculate eigen values based on all genomic windows (–ntop 0) and top 2 principle components were plotted using matplotlib (Caswell et al., 2020). Custom scripts were used to analyze sample correlation, clustering, and correlation with ChIP-seq data and annotated genes using Python packages numpy (Harris et al., 2020), scipy (SciPy 1.0 Contributors et al., 2020), pandas (Reback et al., 2020), and matplotlib (Caswell et al., 2020). Open regions were identified using HMMRATAC (–threshold 2 –score fc -u 20 -l 10) (Tarbell and Liu, 2019) and MACS2 (-q 0.05 -B –broad -f BAMPE) (Zhang et al., 2008). Jaccard indices were calculated using pybedtools (Quinlan and Hall, 2010; Dale et al., 2011)Quinlan and Hall, 2010) for each pair of biologic replicates with default parameters. More detailed pipeline is available at https://github.com/SichongP/FAANG_ATACseq.

Histone ChIP-Seq Data Processing

Histone ChIP-seq data were downloaded from FAANG data repository[1] under accession PRJEB35307. Histone marks were determined according to Kingsley et al. (2019) and compared with open chromatin regions analyzed in this study for both liver and lamina.

ATAC-Seq Peak Validation With Histone Marks

ATAC-seq peaks called by HMMRATAC and MACS2 were validated using histone ChIP-seq data following (Tarbell and Liu, 2019) with modifications to utilize available data in the horse. First, the following sets of peaks were generated from Kingsley et al. (2019) data:

Real positive set (RP): peaks from either H3K4me1 or H3K4me3 that overlap H3K27ac peaks
Real negative set (RN): peaks from H3K27me3 data

Then, following metrics were calculated for each dataset:

$$TP = number\ of\ bases\ in\ called\ ATAC-seq\ peaks\ overlapping\ RP$$

$$FP = number\ of\ bases\ in\ called\ ATAC-seq\ peaks\ overlapping\ RN$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{RP}$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{RN}$$

Increasing quality scores as produced by MACS2 or HMMRATAC were used as the cutoff score to filter peaks before the remaining peaks were used to calculate above metrics. Changes in the metrics as the cutoff score increased were used to identify the thresholds at which to filter final sets of open chromatin peaks.

RNA-Seq Data Processing

RNA-seq reads from liver and lamina of the same two animals were available from a separate project under European Nucleotide Archive accession PRJEB26787. Briefly, RNA was isolated from liver or lamina tissues using Trizol chloroform phase separation followed by a column cleanup using Zymo Research Direct-Zol Mini columns. TruSeq mRNA libraries were prepared at Minnesota Genomics Center (Minneapolis, MN, United States) and sequenced at 125 bp paired-end. These reads were quantified against Equcab3 Ensembl annotated genes (Kalbfleisch et al., 2018; Cunningham et al., 2019) using Salmon (Patro et al., 2017) mapping-based mode. Transcript level counts were aggregated into gene level using the R package tximport (Soneson et al., 2015) and final counts were normalized using the variance-stabilizing transformation method from DESeq2 vst function (Love et al., 2014).

ATAC-Seq Peak Validation With RNA-Seq Data

Ensembl annotated genes were classified as open or closed depending on whether their presumed promoter regions (1 kb upstream of annotated gene start) overlapped with identified ATAC-seq peaks. These genes were then compared to their RNA abundance estimated using FAANG data.

RESULTS

Libraries prepared by two laboratories (L1 and L2) using nuclei isolated from snap-frozen tissues (FTDN) or cryopreserved from tissues at time of collection (CN) from liver and lamina of two animals (AH1 and AH2, Thoroughbred adult mares) were sequenced at PE75 on an Illumina HiSeq 4000 (L1) or PE42 on an Illumina NextSeq 500 (L2). **Figure 1** shows a schematic of the experimental design.

[1] https://data.faang.org/home

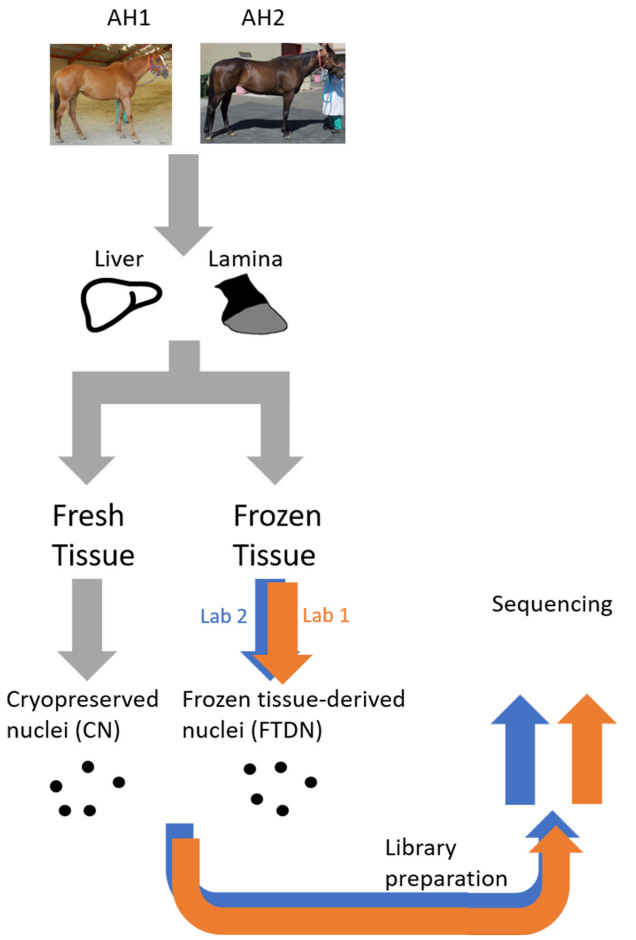


FIGURE 1 | A schematic of the experimental design. All samples were prepared at UC Davis prior to shipment to the core laboratories. Samples used were obtained from an equine biobank of two horses (AH1 and AH2), as previously described (Burns et al., 2018).

Library Fragmentation

ATAC-seq libraries are expected to present a laddering pattern that corresponds to different nucleosome-bound fragments. **Supplementary Figures 1**, **2** show fragment size distributions of ATAC libraries as determined by sequencing and Agilent Fragment Analyzer (L1) or TapeStation (L2) from L1 and L2, respectively. In general, liver libraries showed distinguishable laddering pattern while in lamina libraries, only the fragment size corresponding to nucleosome-free fragments was observed.

Sequencing Read Lengths

Since libraries from L1 and L2 were sequenced at different lengths (75 and 42 bp, respectively), we trimmed longer reads from L1 from 3′ down to 42 bp and compared read alignment statistics to those obtained using full length reads (75 bp), after appropriate quality trimming. There were no significant changes in read alignment statistics, with less than 0.02% fewer reads aligned and less than 0.3% fewer reads identified as duplicates for each library after length trimming. Therefore, we proceeded with data analysis using original full length reads from both laboratories.

Duplication Rate and Mitochondrial Contamination

Overall, liver libraries have higher mitochondrial contamination than lamina libraries, likely due to higher metabolic activities in liver (**Supplementary Figure 3A**). Among liver samples, CN libraries prepared by L1 contained 56 and 81% duplicates, with 37 and 23% mitochondrial reads in AH1 and AH2, respectively. In comparison, the CN libraries from L2 contained 31 and 24% duplicates, with 23 and 10% mitochondrial reads from AH1 and AH2, respectively, (**Supplementary Figure 3A**). It was suspected that the higher amount of mitochondrial contamination contributed to the higher duplication rate and led to lower library complexity. To test this hypothesis, resequencing was performed for the liver CN libraries from L1. The number of unique nuclear reads from AH2 largely remained unchanged despite increasing read depth three-fold. For AH1, however, twice the number of unique nuclear reads was obtained after the total read depth was increased (**Supplementary Figure 3B**). Both the fingerprint plot and fraction of reads in peaks (FRiP) identified a decrease in enrichment for AH1 with increased sequencing depth but little change for AH2 (**Supplementary Figure 3C** and **Supplementary Table 2**). This suggests that, in the AH1 library, while further sequencing increased the number of unique reads, it did not substantially improve peak detection. Lowered enrichment in the resequenced AH1 library suggests that a majority of additional unique reads are less enriched background reads. In the AH2 library, however, resequencing did not significantly improve library complexity, due to more cycles of amplification during library preparation and therefore, higher PCR duplication rate in the library.

Genome Coverage and Enrichment

To assess which part of the ATAC-seq protocol contributed more to library variations and complexities, we compared genome coverage and enrichment (**Figure 2**). Principle component analysis (PCA) revealed that liver libraries generally clustered closely together, while more variation was observed for the lamina libraries (**Figure 2A**). Within the lamina libraries, there is a clear clustering based on which laboratory prepared the libraries. The lamina libraries from L2 clustered closely with each other and with liver libraries while the lamina libraries from L1 clustered further away from liver libraries (**Figure 2A**). Heatmaps of the genome coverage Pearson correlation showed that liver CN libraries yielded well-correlated results, with the exception of that from AH2 by L1 (**Figure 2B**). This is consistent with low complexity of that library shown in **Supplementary Figure 3**. On the other hand, little correlation is observed among lamina library preparations (**Figure 2B**). Since no input libraries were used for ATAC-seq experiments (Buenrostro et al., 2015), synthetic Jensen-Shannon distance (SJSD) was used, together with Area Under Curve (AUC) from fingerprint plots, to assess the enrichment of each library (**Figure 2C** and **Supplementary Table 3**). In general, liver libraries showed higher enrichment

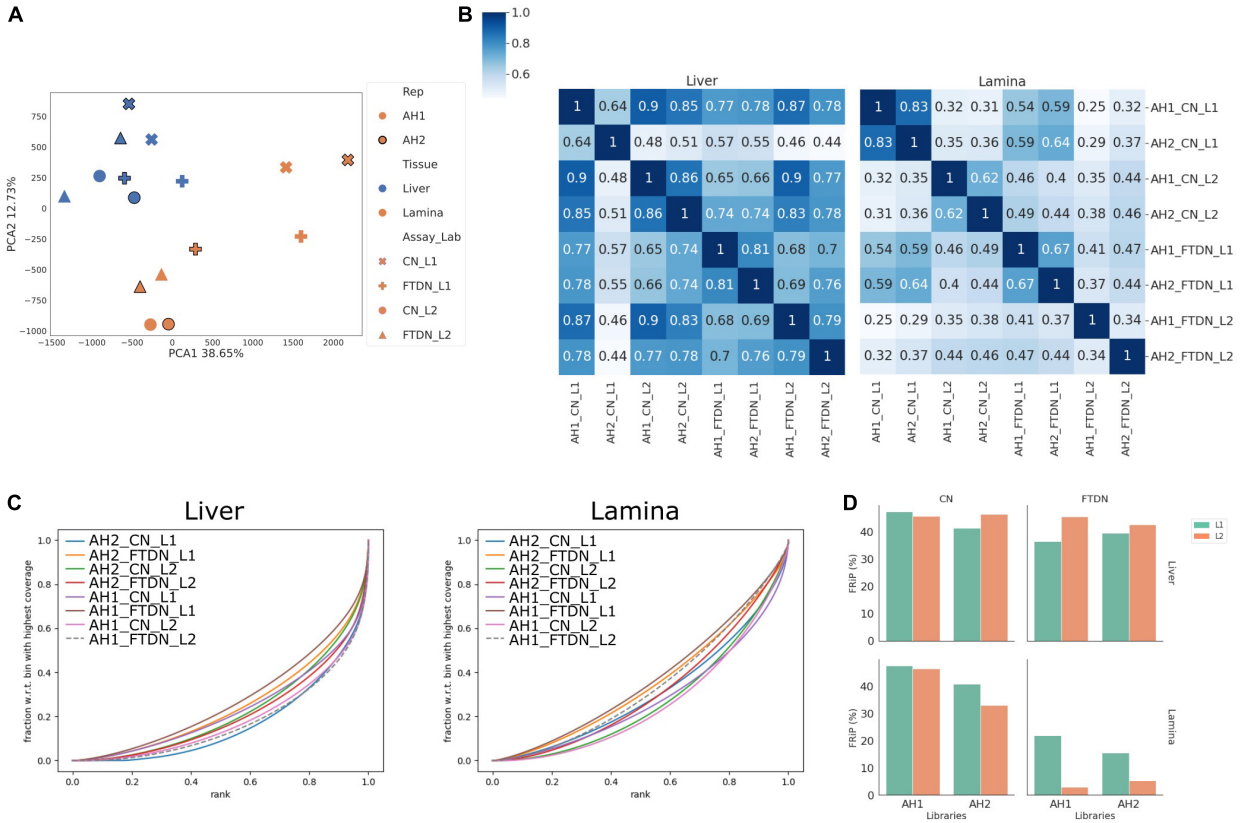


FIGURE 2 | Read coverage correlation between libraries. Read depth was normalized across all libraries. **(A)** Principal component analysis of genome coverage, showing the first two principal components. **(B)** Pearson correlation of genome coverage in liver (left) and lamina (right) libraries. Linkage was calculated using Farthest Point Algorithm. **(C)** Fingerprint plot of genome coverage in liver (left) and lamina (right) libraries. **(D)** Enrichment as measured by FRiP in each library.

than lamina libraries. Within liver libraries, CN libraries were more enriched than FTDN libraries from L1, while both libraries from L2 showed similar enrichment. Within lamina libraries, both laboratories generated more enriched libraries from CN than from FTDN. This is further exemplified in **Figure 2D**, showing the FRiP in each library.

Peak Calling

To identify accessible chromatin regions, MACS2 (Zhang et al., 2008) and HMMRATAC (Tarbell and Liu, 2019) were used to call peaks and results from both programs were compared. To control for sequencing depth, all libraries were down-sampled to 60 million unique reads that are suitable for peak calling using sambamba view function. Using MACS2 (-q 0.05 -B–broad -f BAMPE), 31,000–721,000 peaks were identified. While using HMMRATAC (–threshold 2 –score fc -u 20 -l 10), 14,000–514,000 peaks were identified. Overall, using HMMRATAC, peaks identified from lamina libraries had lower quality [fewer (**Figure 3A**) and shorter peaks (**Figure 3B**) with lower scores (**Figure 3C**)] than those from liver libraries. For liver libraries, CN generated comparable results to FTDN while, in lamina libraries, CN outperformed FTDN (**Figure 3D**). Similar results were obtained when peaks were called using MACS2 (**Supplementary Figures 4A,D**).

To better assess the quality of peaks, we used histone mark ChIP-seq data generated from the same samples as described in Kingsley et al. (2019). A set of metrics, precision, recall, and false positive rate (FPR), were generated for different cutoff scores as described in Methods. These metrics were then plotted against cutoff scores. Consistent with the observation of peak lengths and scores, peaks called using HMMRATAC from liver libraries had higher precision and recall rates and lower false positive rates (**Figure 3E**) than lamina (**Figure 3F**). Consistent with observations of library quality, CN liver libraries of AH2 from L1 have lower recall and precision rates than that from L2 or that of AH1, despite having same unique read depth (**Figure 3E**). Comparing peaks identified by two programs, HMMRATAC identified peaks with higher recall and precision rates than MACS2 (**Supplementary Figures 4E,F**).

ATAC-Seq Peak Validation

Despite higher quality from L2 in liver AH2 CN library, L1 produced the only libraries from laminae with high quality peaks (**Figure 3F**). Therefore, to maximize usable data, libraries from L1 were chosen for all further analyses. HMMRATAC was used as it produced generally better metrics and because it allowed interrogation of nucleosome-bound regions vs. nucleosome-free regions for future studies.

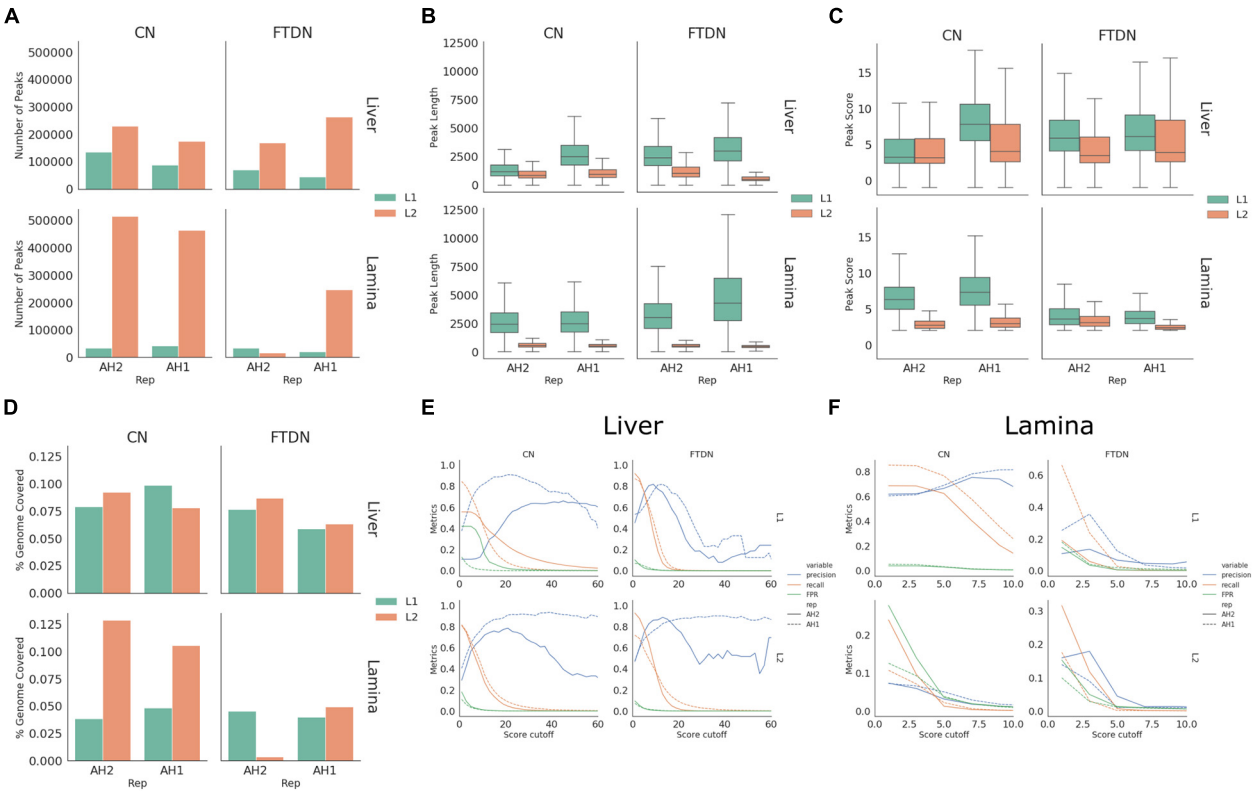


FIGURE 3 | HMMRATAC peak calling statistics. **(A)** Number of peaks, **(B)** peak length distribution, **(C)** peak score distribution, and **(D)** percent of genome covered by peaks for each library. **(E,F)** Peak metrics assessed using ChIP-seq dataset in liver **(E)** and lamina **(F)** libraries.

A cutoff score, where the precision and recall lines intercept, was used for each sample set to filter peaks identified by HMMRATAC. Final peak counts are shown in **Table 1**. Consistent with previous observations, liver samples generated the most high-quality peaks, while CN libraries outperformed FTDN libraries. Using UpSetPlot (Nothman, 2020) based on (Lex et al., 2014), we identified overlapping peaks in each dataset (**Figure 4A**). AH1 liver CN library generated the most unique peaks, consistent with the previous observation that this library has highest library complexity. Since 17,347 unique peaks were identified from this library only, a precision score of these unique peaks was calculated using histone ChIP-seq data mentioned above. A precision score of 18.4% was observed in these peaks, suggesting a high rate of false positive peaks. This further highlights the importance of replicates in an ATAC-seq experiment. FTDN libraries did not yield significant number of unique peaks that were not detected in CN libraries. Despite a relatively low quality of the lamina libraries, 12,256 unique peaks were detected from the lamina libraries.

TABLE 1 | Cutoff used to filter peaks and metrics of filtered peaks.

Tissue	Rep	Nuclei prep	Cutoff score	Count	AvePeakLen	MedianPeakLen	Bases covered	Confirmed count	Jaccard index
Liver	AH1	CN	6	61,473	2,937.0	2,600	180,547,240	3,646	0.05
Liver	AH2	CN	16	3,810	2,428.8	2,090	9,253,670		
Liver	AH1	FTDN	6	22,588	3,701.6	3,300	83,611,751	18,596	0.35
Liver	AH2	FTDN	6	33,782	3,059.1	2,650	103,343,612		
Lamina	AH1	CN	6	28,418	3,106.6	2,650	88,284,203	23,439	0.51
Lamina	AH2	CN	4	30,906	2,883.5	2,480	89,117,724		
Lamina	AH1	FTDN	2	19,886	5,061.4	4,300	100,651,092	17,619	0.35
Lamina	AH2	FTDN	2	33,762	3,361.9	3,010	113,504,835		

Filtered peaks and their corresponding cutoff scores in each library. AvePeakLen, Average peak length; MedianPeakLen, Median peak length; Bases covered, number of bases covered by all peaks in a library; Confirmed count, overlapping peaks in both biological replicates; Jaccard index, jaccard index of two biological replicates.

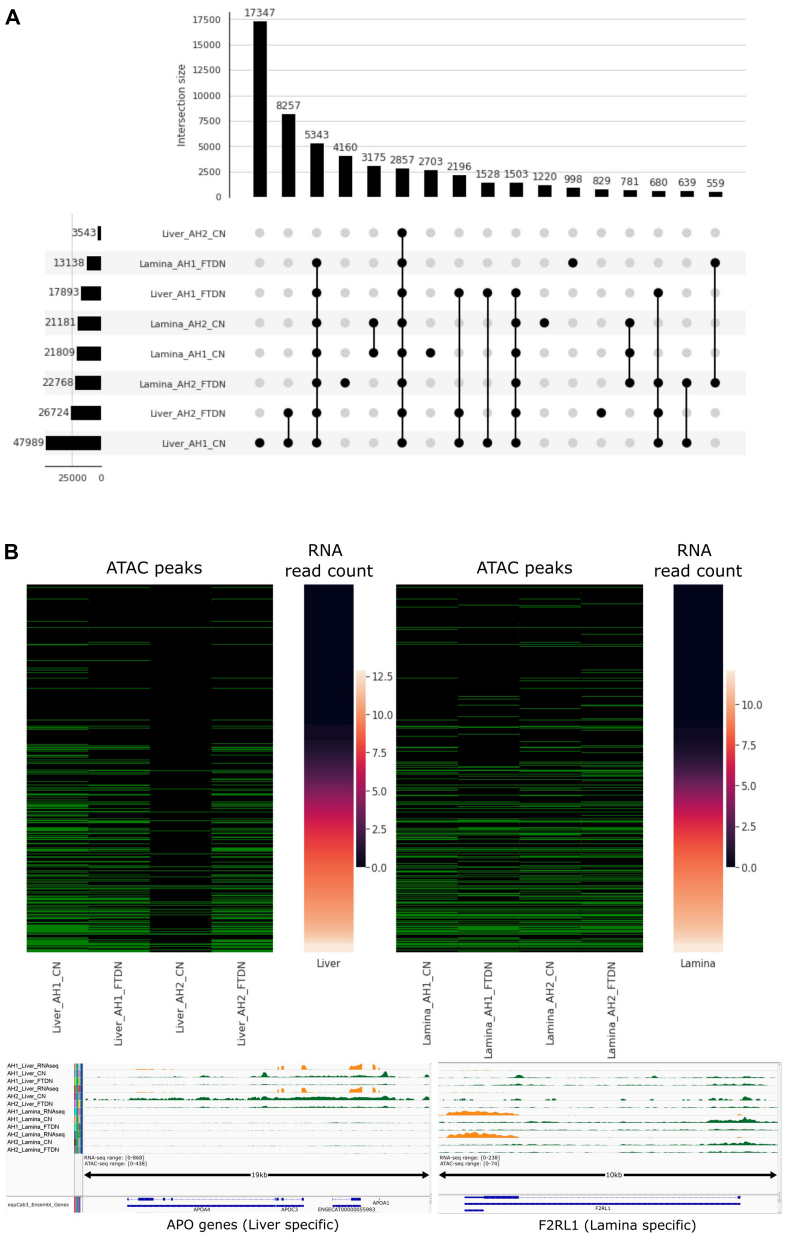


FIGURE 4 | Filtered ATAC-seq peaks. **(A)** Intersection plot of quality filtered peaks from each library. Bottom left panel shows filtered peak count in each library; bottom right panel shows different intersections (BedTools, 1 bp minimum) of peaks where filled dots indicate presence of peaks in corresponding library; Top panel shows peak count in each intersection. **(B)** Relationship between promoter accessibility and gene expression (mean vst transformed count) in liver (top left) and lamina (top right). Green cell in ATAC peaks indicate presence of ATAC peaks and black cells indicate absence. Bottom panel shows bigwig tracks of RNA-seq and ATAC-seq read abundance (normalized using RPKM)near *APO genes* (left, liver specific) and *F2RL1* (right, lamina specific) transcription start sites.

As an *in silico* validation of the results, peaks were overlapped with Ensembl gene annotation for EquCab3 (Kalbfleisch et al., 2018) at promoter regions (1 kb upstream of annotated gene start) to classify each promoter as open or closed. These classified promoter regions were then compared to RNA abundance at the corresponding gene level (**Figure 4B**). In liver, AH1 CN identified more open promoters where RNA expression levels are high but the results from the two assays (CN and FTDN) were highly comparable for this sample in liver. Fewer peaks were identified from AH2 CN, due to low library quality and issues in repeat freeze thaw cycles as outlined in the discussion. In lamina, CN assays identified more open promoters than FTDN. Manual inspection of some highly abundant genes in liver and laminae validate accurate identification of open chromatin in each tissue (**Figure 4B**).

Overall, our results confirm that extracting nuclei from snap-frozen tissues for ATAC-seq library preparations negatively

affects the library quality, resulting in fewer peaks detected. However, when CN from freshly collected tissue are not available, these data show that snap-frozen tissues can be used to prepare ATAC-seq libraries to give reliable peak calls, with the caveat that some regions of open chromatin will be missed. However, results from laminae suggest that for more challenging tissue types, fresh tissue extraction is a requirement.

DISCUSSION

In this pilot study, we compared two tissues (liver and laminae, representing homogenous cellular and relatively acellular nuclei extraction, respectively) from the equine FAANG project for ATAC-seq library generation, using two nuclei extraction methods. Nuclei extracted and cryopreserved immediately after tissue collection and nuclei isolated from snap-frozen tissues were used to determine suitable methods for performing ATAC-seq to identify accessible chromatin regions in a wide variety of equine tissues for functional annotation. Similar to what was identified by Halstead et al. (2020a), we determined that ATAC-seq can be used to characterize open chromatin in animal tissue but optimization is necessary to have a robust data set across tissues. Further, we found that while CN generally yield more peaks, frozen tissues can still be used to isolate nuclei and identify accessible regions. However, the quality of libraries generated by the frozen tissue protocol suffered when nuclei were extracted from a more challenging, relatively acellular tissue, such as laminae. Therefore, for challenging tissues, care should be taken at time of collection to prioritize those tissues for nuclei extraction and cryopreservation when possible.

We also showed that the frozen tissue protocol is more prone to variations introduced at the library preparation step. Specifically, FTDN liver libraries generated at two different laboratories only have a moderate correlation (0.68 for AH1 and 0.76 for AH2). Our analysis suggests that, similar to RNA-seq experiments, library preparation can introduce large variation that will impact subsequent data quality, specifically peak detection for ATAC-seq studies. However, since the two commercial laboratories used different internally optimized protocols, it is impossible to determine whether the variation was protocol-specific or lab-specific. Nonetheless, it is advisable for all ATAC-seq library preparations to be performed at a single site using the same protocols to minimize variability in datasets when trying to integrate information.

During library preparation, the CN aliquot from AH2 was partially thawed twice by L1 (first for an optimization experiment (data not shown) and then a second time to perform the data collection). The nuclei obtained during the second partial thawing were used in this study. Due to the precipitation of nuclei and contaminating mitochondria, this was likely the cause of low quality observed in that library preparation. The effect of different read lengths used by two laboratories was investigated and deemed to have no significant impact on read alignment. Our analysis suggested a detrimental impact on data quality by this practice and resequencing of this particular library also did not improve data quality nor was this resequencing effort able to identify more peaks. Therefore, it is advisable to avoid repeated partial thawing of CN aliquots.

Library fragment size screening using gel electrophoresis proved to be predictive of final fragment size distribution in sequencing results and data quality. As indicated in **Supplementary Figures 1**, **2**, a strong signature corresponding to nucleosome-free fragments without accompanying signatures for nucleosome-bound regions does not necessarily mean a high enrichment of nucleosome-free fragments. It could also indicate high levels of mitochondria contamination or fragmentation of chromatins before tagmentation, which are likely the cases in lamina libraries from L2.

We identified 20–33,000 accessible chromatin regions in lamina and 22–61,000 in liver, largely in line with observations of liver ATAC-seq from studies in other species (Ackermann et al., 2016; Foissac et al., 2019; Liu et al., 2019; Halstead et al., 2020b). As a preliminary study, we opted to include laboratory replicates in lieu of technical replicates in order to assess the effect of technical variations introduced during the library preparation step. Technical replicates would allow further validation of tissue specific open-chromatin. Following ENCODE standard (Landt et al., 2012) for ChIP-seq experiments, two biological replicates were collected for the FAANG project. However, more replicates would have allowed a more robust comparison between different protocols.

In this study, we demonstrated the feasibility of using snap-frozen tissues for ATAC-seq experiments for the equine FAANG project. For acellular tissues, more optimization is required for ATAC-seq experiments. We also showed that significant variation can be introduced during library preparation. This study provides important guidelines for planning future ATAC-seq experiments using equine FAANG tissues. We will use the guidelines established here to conduct ATAC-seq experiments on six other prioritized tissues in the mares. Furthermore, following these guidelines should enable the most meaningful integration of datasets across studies thus building a reliable functional tissue specific atlas of the equine genome which would advance our understanding of complex traits in the horse.

DATA AVAILABILITY STATEMENT

ATAC-seq data used in this study are available from the European Nucleotide Archive under the accession PRJEB41317. RNA-seq data from the liver and lamina tissues of the same two animals used in this study can be found from the European Nucleotide Archive under the accession PRJEB26787.

AUTHOR CONTRIBUTIONS

CF and SP were responsible for the conceptualization, methodology, investigation, formal analysis, resources, and writing of the manuscript. CF, RB, JP, and TK were responsible for the funding acquisition. All authors have reviewed the final manuscript.

FUNDING

Funding was provided by the Grayson Jockey Club Foundation, USDA NRSP-8, and the UC Davis Center for Equine Health. Funding Support for CF was provided by the National Institutes of Health (NIH) (L40 TR001136). None of the funding agencies had any role in the design of the study, analysis, interpretation of the data, or writing of the manuscript. Salary support for SP was provided by the Ann T. Bowling Fellowship at the UC Davis Veterinary Genetics Laboratory.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Diagenode Epigenomic service team and Active Motif service team for providing partial financial support and completing the assays for this study. The preliminary results of this study were presented at Plant and Animal Genome conference in San Diego in 2019.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.641788/full#supplementary-material

Supplementary Figure 1 | Fragment size distributions of libraries from L1 as determined by sequencing and Fragment Analyzer.

Supplementary Figure 2 | Fragment size distributions of libraries from L2 as determined by sequencing and tapestation.

Supplementary Figure 3 | Duplication and mitochondrial contamination rates. **(A)** Total, mitochondrial, and unique nuclear read counts of all libraries; **(B)** Comparison between first sequencing run (left) and combined reads (right) from L1 liver CN libraries; **(C)** Fingerprint plot of L1 CN liver libraries.

Supplementary Figure 4 | MACS2 peak calling statistics. **(A)** Number of peaks, **(B)** peak length distribution, **(C)** peak score distribution, and **(D)** percent of genome covered by peaks for each library. **(E,F)** Peak metrics assessed using ChIP-seq dataset in liver **(E)** and lamina **(F)** libraries.

REFERENCES

Ackermann, A. M., Wang, Z., Schug, J., Naji, A., and Kaestner, K. H. (2016). Integration of ATAC-Seq and RNA-Seq identifies human alpha cell and beta cell signature genes. *Mol. Metab.* 5, 233–244. doi: 10.1016/j.molmet.2016.01.002

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed September 2016).

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

Burns, E. N., Bordbari, M. H., Mienaltowski, M. J., Affolter, V. K., Barro, M. V., Gianino, F., et al. (2018). Generation of an equine biobank to be used for functional annotation of animal genomes project. *Anim. Genet.* 49, 564–570. doi: 10.1111/age.12717

Caswell, T. A., Droettboom, M., Lee, A., Hunter, J., Firing, E., Stansby, D., et al. (2020). *Matplotlib/Matplotlib v3.1.3* (version v3.1.3). Zenodo. doi: 10.5281/ZENODO.3633844

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., et al. (2017). An improved ATAC-Seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* 14, 959–962. doi: 10.1038/nmeth.4396

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., et al. (2019). Ensembl 2019. *Nucleic Acids Res.* 47, D745–D751. doi: 10.1093/nar/gky1113

Dale, R. K., Pedersen, B. S., and Quinlan, A. R. (2011). Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics* 27, 3423–3424. doi: 10.1093/bioinformatics/btr539

Durham, A. E., Frank, N., McGowan, C. M., Menzies-Gow, N. J., Roelfsema, E., Vervuert, I., et al. (2019). ECEIM consensus statement on equine metabolic syndrome. *J. Vet. Int. Med.* 33, 335–349. doi: 10.1111/jvim.15423

Finno, C. J., and Bannasch, D. L. (2014). Applied equine genetics. *Equine Vet. J.* 46, 538–544. doi: 10.1111/evj.12294

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., et al. (2019). Multi-Species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol.* 17:108. doi: 10.1186/s12915-019-0726-5

Ghosh, M., Sharma, N., Singh, A. K., Gera, M., Pulicherla, K. K., and Jeong, D. K. (2018). Transformation of animal genomics by next-generation sequencing technologies: a decade of challenges and their impact on genetic architecture. *Crit. Rev. Biotechnol.* 38, 1157–1175. doi: 10.1080/07388551.2018.1451819

Giuffra, E., Tuggle, C. K., and Faang Consortium (2019). Functional Annotation of Animal Genomes (FAANG): current achievements and roadmap. *Ann. Rev. Anim. Biosci.* 7, 65–88. doi: 10.1146/annurev-animal-020518-114913

Halstead, M. M., Kern, C., Saelao, P., Chanthavixay, G., Wang, Y., Delany, M. E., et al. (2020a). Systematic alteration of ATAC-Seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues. *Sci. Rep.* 10:5230. doi: 10.1038/s41598-020-61678-9

Halstead, M. M., Kern, C., Saelao, P., Wang, Y., Chanthavixay, G., Medrano, J. F., et al. (2020b). A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. *BMC Genomics* 21:698. doi: 10.1186/s12864-020-07078-9

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

Kalbfleisch, T. S., Rice, E. S., DePriest, M. S., Walenz, B. P., Hestand, M. S., Vermeesch, J. R., et al. (2018). Improved reference genome for the domestic horse increases assembly contiguity and composition. *Commun. Biol.* 1:197. doi: 10.1038/s42003-018-0199-z

Kingsley, N. B., Kern, C., Creppe, C., Hales, E. N., Zhou, H., Kalbfleisch, T. S., et al. (2019). Functionally annotating regulatory elements in the equine genome using histone mark ChIP-Seq. *Genes* 11:3. doi: 10.3390/genes11010003

Krueger, F. (2019). *Trim Galore!* (version 0.6.5). Available online at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed February 2020).

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., et al. (2012). ChIP-Seq guidelines and practices of the ENCODE and ModENCODE Consortia. *Genome Res.* 22, 1813–1831. doi: 10.1101/gr.136184.111

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans. Visual. Comp. Graph.* 20, 1983–1992. doi: 10.1109/TVCG.2014.2346248

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liu, C., Wang, M., Wei, X., Wu, L., Xu, J., Dai, X., et al. (2019). An ATAC-Seq atlas of chromatin accessibility in mouse tissues. *Sci. Data* 6:65. doi: 10.1038/s41597-019-0071-0

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:10. doi: 10.14806/ej.17.1.200

McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J., et al. (2011). RNA-Seq: Technical Variability and Sampling. *BMC Genomics* 12:293. doi: 10.1186/1471-2164-12-293

Milani, P., Escalante-Chong, R., Shelley, B. C., Patel-Murray, N. L., Xin, X., Adam, M., et al. (2016). Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. *Sci. Rep.* 6:25474. doi: 10.1038/srep25474

Nothman, J. (2020). *UpSetPlot* (version 0.4.1). Available online at: https://github.com/jnothman/UpSetPlot (accessed December 2020).

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi: 10.1038/nmeth.4197

Qu, H., and Fang, X. (2013). A brief review on the human encyclopedia of DNA elements (ENCODE) project. *Genomics Proteomics Bioinformatics* 11, 135–141. doi: 10.1016/j.gpb.2013.05.001

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., et al. (2016). DeepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44, W160–W165. doi: 10.1093/nar/gkw257

Raudsepp, T., Finno, C. J., Bellone, R. R., and Petersen, J. L. (2019). Ten years of the horse reference genome: insights into equine biology, domestication and population dynamics in the Post-genome Era. *Anim. Genet.* 50, 569–597. doi: 10.1111/age.12857

Reback, J., McKinney, W., jbrockmendel, Van Den Bossche, J., Augspurger, T., Cloud, P., et al. (2020). *Pandas-Dev/Pandas: Pandas 1.1.3* (version v1.1.3). Zenodo, doi: 10.5281/ZENODO.3509134

SciPy 1.0 Contributors, Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* 17, 261–272. doi: 10.1038/s41592-019-0686-2

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for RNA-Seq: transcript-level estimates improve gene-level inferences. *F1000Research* 4:1521. doi: 10.12688/f1000research.7563.1

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31, 2032–2034. doi: 10.1093/bioinformatics/btv098

Tarbell, E. D., and Liu, T. (2019). HMMRATAC: A hidden Markov ModeleR for ATAC-Seq. *Nucleic Acids Res.* 47:e91. doi: 10.1093/nar/gkz533

The FAANG Consortium, Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project. *Genome Biol.* 16:57. doi: 10.1186/s13059-015-0622-4

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science (New York, N.Y.)* 326, 865–867. doi: 10.1126/science.1178158

Wylie, C. E., Collins, S. N., Verheyen, K. L., and Richard Newton, J. (2011). Frequency of equine laminitis: a systematic review with quality appraisal of published evidence. *Vet. J.* 189, 248–256. doi: 10.1016/j.tvjl.2011.04.014

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

Conflict of Interest: The cost of library preparation and sequencing was partially covered by two core laboratories as part of collaboration to optimize ATAC-seq protocol using horse tissues.

Copyright © 2021 Peng, Bellone, Petersen, Kalbfleisch and Finno. 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

TECHNOLOGY AND CODE
published: 17 June 2021
doi: 10.3389/fgene.2021.639238

The FAANG Data Portal: Global, Open-Access, "FAIR", and Richly Validated Genotype to Phenotype Data for High-Quality Functional Annotation of Animal Genomes

Peter W. Harrison, Alexey Sokolov, Akshatha Nayak, Jun Fan, Daniel Zerbino, Guy Cochrane and Paul Flicek*

European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom

The Functional Annotation of ANimal Genomes (FAANG) project is a worldwide coordinated action creating high-quality functional annotation of farmed and companion animal genomes. The generation of a rich genome-to-phenome resource and supporting informatic infrastructure advances the scope of comparative genomics and furthers the understanding of functional elements. The project also provides terrestrial and aquatic animal agriculture community powerful resources for supporting improvements to farmed animal production, disease resistance, and genetic diversity. The FAANG Data Portal (https://data.faang.org) ensures Findable, Accessible, Interoperable and Reusable (FAIR) open access to the wealth of sample, sequencing, and analysis data produced by an ever-growing number of FAANG consortia. It is developed and maintained by the FAANG Data Coordination Centre (DCC) at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI). FAANG projects produce a standardised set of multi-omic assays with resulting data placed into a range of specialised open data archives. To ensure this data is easily findable and accessible by the community, the portal automatically identifies and collates all submitted FAANG data into a single easily searchable resource. The Data Portal supports direct download from the multiple underlying archives to enable seamless access to all FAANG data from within the portal itself. The portal provides a range of predefined filters, powerful predictive search, and a catalogue of sampling and analysis protocols and automatically identifies publications associated with any dataset. To ensure all FAANG data submissions are high-quality, the portal includes powerful contextual metadata validation and data submissions brokering to the underlying EMBL-EBI archives. The portal will incorporate extensive new technical infrastructure to effectively deliver and standardise FAANG's shift to single-cellomics, cell atlases, pangenomes, and novel phenotypic prediction models. The Data Portal plays a key role for FAANG by supporting high-quality functional annotation of animal genomes, through open FAIR sharing of data, complete with standardised rich metadata. Future Data Portal features developed by the DCC will support new technological developments for continued improvement for FAANG projects.

Keywords: FAANG, functional annotation, phenotype to genotype, FAIR data, agricultural genomics, Data Portal, open access, metadata validation

OPEN ACCESS

Edited by:
Christopher K. Tuggle,
Iowa State University, United States

Reviewed by:
Zhi-Liang Hu,
Iowa State University, United States
Ted Kalbfleisch,
University of Kentucky, United States

****Correspondence:***
Peter W. Harrison
peter@ebi.ac.uk

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

Received: *08 December 2020*
Accepted: *04 May 2021*
Published: *17 June 2021*

Citation:
Harrison PW, Sokolov A, Nayak A, Fan J, Zerbino D, Cochrane G and Flicek P (2021) The FAANG Data Portal: Global, Open-Access, "FAIR", and Richly Validated Genotype to Phenotype Data for High-Quality Functional Annotation of Animal Genomes. Front. Genet. 12:639238. doi: 10.3389/fgene.2021.639238

INTRODUCTION

The Functional Annotation of Animal Genomes Project (FAANG) is a coordinated action to improve availability of high-quality functional annotation of farmed and companion animal genomes (Andersson et al., 2015; Tuggle et al., 2016; Giuffra et al., 2019; Clark et al., 2020). Rich genome-to-phenome resources are of particular importance in domesticated animals of commercial importance for efforts to increase agricultural production, but the available resources also impact upon the fundamental understanding of functional elements, biomedical science, evolution, and the environment. The FAANG project comprises multiple globally distributed consortia working across a growing range of species and committed to high-quality data production and interpretation. The FAANG Data Coordination Centre (DCC) at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) ensures that all data generated by the project is richly described, consistently reported, openly available, reusable, and clearly presented (Harrison et al., 2018). The FAANG Data Portal[1] plays a pivotal role by coordinating and presenting the wealth of data generated by the project to the scientific community. Its primary purpose is to provide a searchable, unified view of the multi-omic FAANG data held across specialised EMBL-EBI archives. Its web interface and Application Programming Interface (API) supports the identification and download of FAANG generated and associated community datasets, as well as exploring the associated rich validated metadata, protocols, and publications. Here we describe the key features of this rich genome to phenome resource and look to future developments that will expand it as research advances.

METHODS

The FAANG Data Portal[1] comprises a modern technology stack with a microservice architecture design. Establishing each component as a separate microservice enables more flexible, scalable, and faster maintenance and development of new features. The front-end Data Portal is written in Angular, an open-source web application framework. Continuous integrated deployment is managed by Kubernetes, an open-source system to manage the Docker containers which contain all of the code and software required to launch and run the Data Portal components (**Figure 1**). It is deployed on the EMBL-EBI Embassy Cloud infrastructure[2], enabling a flexible and efficient use of computational resources. The selection of open-source frameworks matches the FAANG projects and EMBL-EBI's ethos of open development. The code has permissive Apache 2.0 licencing to allow the community to reuse and benefit from any of the codebase. The use of Kubernetes and Docker containers would permit the Data Portal to be easily deployed onto other cloud infrastructures, such as Google Cloud or Amazon Web Services, if required, aiding long-term sustainability. Another advantage of deployment on the EMBL-EBI Embassy Cloud is that the underlying datasets are held in the same data centres for rapid access. Metadata search, from both the Data Portal and programmatic API, is supported by a local backend Elasticsearch metadata database. The Elasticsearch database is essentially a customised full-text search engine built explicitly for indexing the FAANG metadata and supports partial word search (through custom tokenisation) with high specificity. Rather than hosting duplicate copies of the data, the underlying data files remain located in the specialised data archives such as the European Nucleotide Archive (Amid et al., 2020) and European Variation Archive. The Data Portal includes direct links and bulk download support to directly obtain the relevant data files. It also supports its own FTP site for hosting protocol and presentation files not suitable for archive submission including track hub files. Track data hubs provide a mechanism to store third-party genome annotations within specifically formatted files for distribution and display. Track hub file formats are optimised for display through web-based genome browsers. FAANG utilises the Track Hub Registry[3] to seamlessly enable FAANG community generated trackhub annotations to be made discoverable for use with the UCSC (Lee et al., 2020) and Ensembl Genome Browsers (Yates et al., 2020). The Data Portal infrastructure is already in place to support track hub submissions, utilising the FAANG Data Portal FTP site and EMBL-EBI Track Hub Registry (https://trackhubregistry.org/). Presentation and linking to FAANG track hubs will be improved in future releases of the Data Portal, so that track hubs are clearly associated with FAANG datasets and users can more quickly view them in the UCSC (Lee et al., 2020) and Ensembl Genome Browsers (Yates et al., 2020).

Metadata standards are held in GitHub in JSON schema for ease of rendering on the Data Portal and use in the validation system. This also allows anyone in the community to propose a metadata change through a pull request. The validation and submission brokering data processing is performed in python with asynchronous data processing also hosted in an Embassy Cloud instance with user interaction through the FAANG Data Portal interface. Documentation is managed through a readthedocs GitHub instance[4] that allows updating site text without the need for redeployment of the full Data Portal.

RESULTS

The FAANG Data Portal

The FAANG Data Portal's primary function is to collate and clearly present the wealth of FAANG data to the community. FAANG data is divided into clear record sections within the Data Portal comprising organism metadata BioSample records, specimen metadata BioSample records, full datasets, individual raw data files, and individual processed analysis files. The portal enables browsing of any of these interconnected sections making it possible to navigate to all of the data records that stem from a given organism. The record table views (**Figure 2**) provide preconfigured filtering to narrow down the search, for example,

[1]https://data.faang.org
[2]https://www.embassycloud.org/
[3]https://trackhubregistry.org/
[4]https://dcc-documentation.readthedocs.io/en/latest/faq/

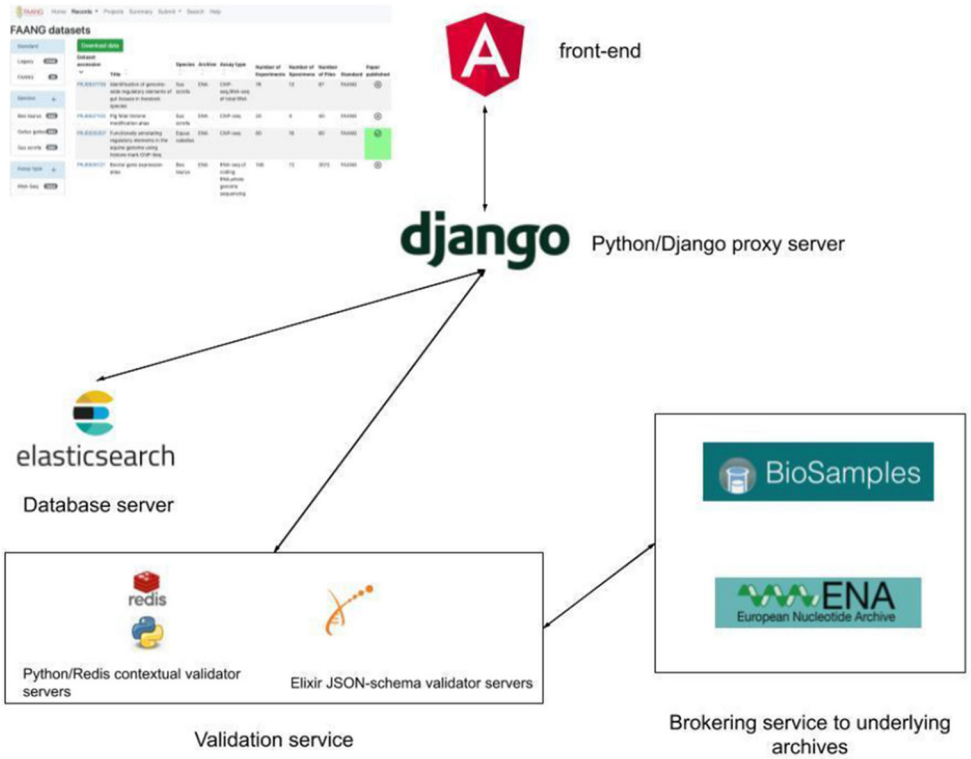


FIGURE 1 | FAANG Data Portal architecture with local Elasticsearch metadata database, python, and JSON-schema contextual validation and brokering of validated data to underlying public archives.

by species or assay type. The filters show the number of records in each field.

Tables are sortable by columns, and the filtered table can be downloaded into CSV or tabular format. As well as the data records, it is also possible to browse and search the extensive collection of protocols. The FAANG data portal automatically scans for publications associated with any of the record identifiers contained in the FAANG Data Portal and automatically downloads and links these publications to their associated records. Other key Data Portal features include global summary statistics, predictive text search, and detailed documentation. The site is supported by an active helpdesk backed by the EMBL-EBI FAANG DCC that supports users with data validation, submission, and retrieval of data. Additionally, the Data Portal collates relevant publicly available functional data generated outside of the FAANG consortium into comparable tables. When these data sets do not meet current stringent FAANG standards, they are clearly labelled with a Legacy tag.

The identification and acquisition of data relevant to a user's scientific interests is the main use case for the portal. The site provides extensive preconfigured filters for exploring data tables, search functionality, and an API. For example, a user interested in obtaining specimens from *Equus caballus* females from the liver left lateral lobe can utilise the preconfigured filters on the specimen data exploration table to select species, sex, and organism part and exclude legacy data (**Figure 3**). The same search is possible using the highly customisable API (https://dcc-documentation.readthedocs.io/en/latest/api/) to return the results programmatically. The current Data Portal search function ranks data based on keyword prioritisation, so for the above example the search would return a range of other datasets that share one or more of the keywords. In a future Data Portal release, the search page (https://data.faang.org/search) will be updated to use an advanced query language that would allow for more advanced text searches. Once records of interest have been identified, data files can be downloaded from within the data portal without having to navigate away to the underlying archives. Continuing with the above example, the 18 files that include Chip-Seq, methylation profiling, and RNA-Seq can be initiated for download from within the Data Portal from the specific specimen pages (https://data.faang.org/specimen/SAMEA104728881).

A recent new feature is project-specific subportal views, initially developed to support the EU-funded Horizon 2020 FAANG projects GENE-SWiTCH, BovReg, and AQUA-FAANG. This feature is now available for any current or future FAANG consortia. These project-specific portals distribute the full functionality of the Data Portal with datasets pre-filtered to show only those from the specific project (**Figure 4**). A customised project page includes relevant information such as a social media stream and can be further tailored with features to support the project's data presentation requirements. Project specific

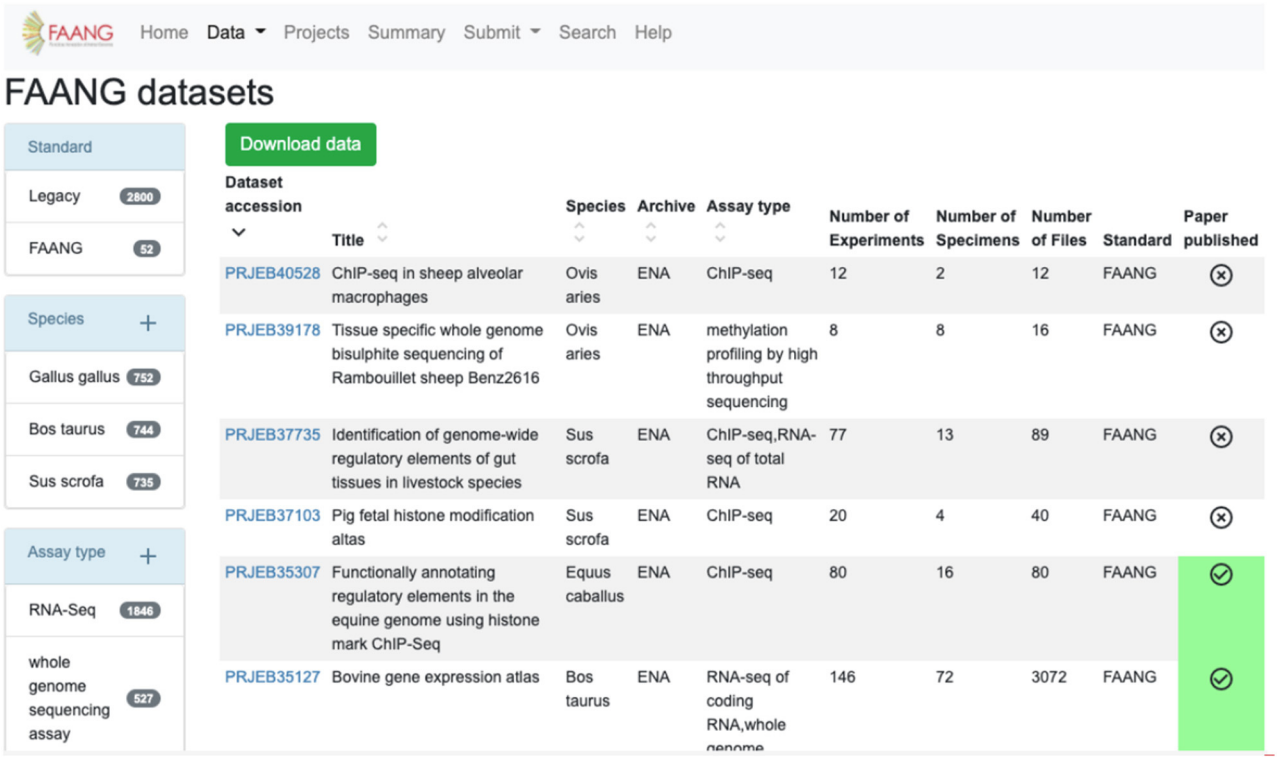


FIGURE 2 | FAANG Data Portal presenting rich 'omic datasets to the community complete with preconfigured data filters, automated literature scraping, and direct links to data files in underlying archives (https://data.faang.org/dataset).

portal pages are constructed using standardised files on GitHub[5], allowing for new project pages to be quickly established.

Metadata Validation and Data Submission Brokering

Metadata validation is fully integrated into the FAANG Data Portal with an improved web interface based on user experience testing (**Figure 5**). The improvements further standardise FAANG submissions, streamline the submission process, and lower the barrier to open data sharing. At the same time, we added fully brokered submissions into the underlying data archives, whereby users who supply their EMBL-EBI credentials can have submissions made on their behalf. This simplifies and accelerates the submission of FAANG data. The Data Portal includes clear guidance for using FAANG's rich metadata standards and provides an intuitive validation interface that ensures every FAANG submission meets this high standard by annotating and requiring improvements prior to submission. A metadata spreadsheet accompanies the data for each submission. This spreadsheet is validated against the FAANG standards and then used to construct the file format required for a brokered archive submission. The Data Portal will also host the corresponding protocols in its FTP directory, thus ensuring long-term availability and cross project standardisation.

The submission brokering service handles the full submission process of sample metadata to the EMBL-EBI BioSamples archive. For sequencing and analysis submissions, the submission brokering system handles the metadata registration and study creation and just requires additionally that the user uploads the sequencing or analysis files direct to the European Nucleotide Archive (ENA) FTP submission server. This ensures that the files are available when the brokering system makes the submission and can correctly associate the files with the created study. FAANG supports all of the data types and file formats currently accepted by the BioSamples and ENA archives and periodically checks for the requirement to support new technologies and file formats. The Data Portal handles any immediate submission errors and presents them back to the submitter. Errors subsequently discovered during post-submission processing by the underlying archives go straight to the submitters' registered email address. Upon a successful submission, the submitter is provided with a receipt that contains all of the assigned identifiers for their submission, which can then be referenced in their publications. Data files typically appear on the FAANG data portal within 48 h, once they have been made publicly available by the underlying archives.

[5]https://github.com/FAANG/comm-data-portal-projects/tree/master/projects

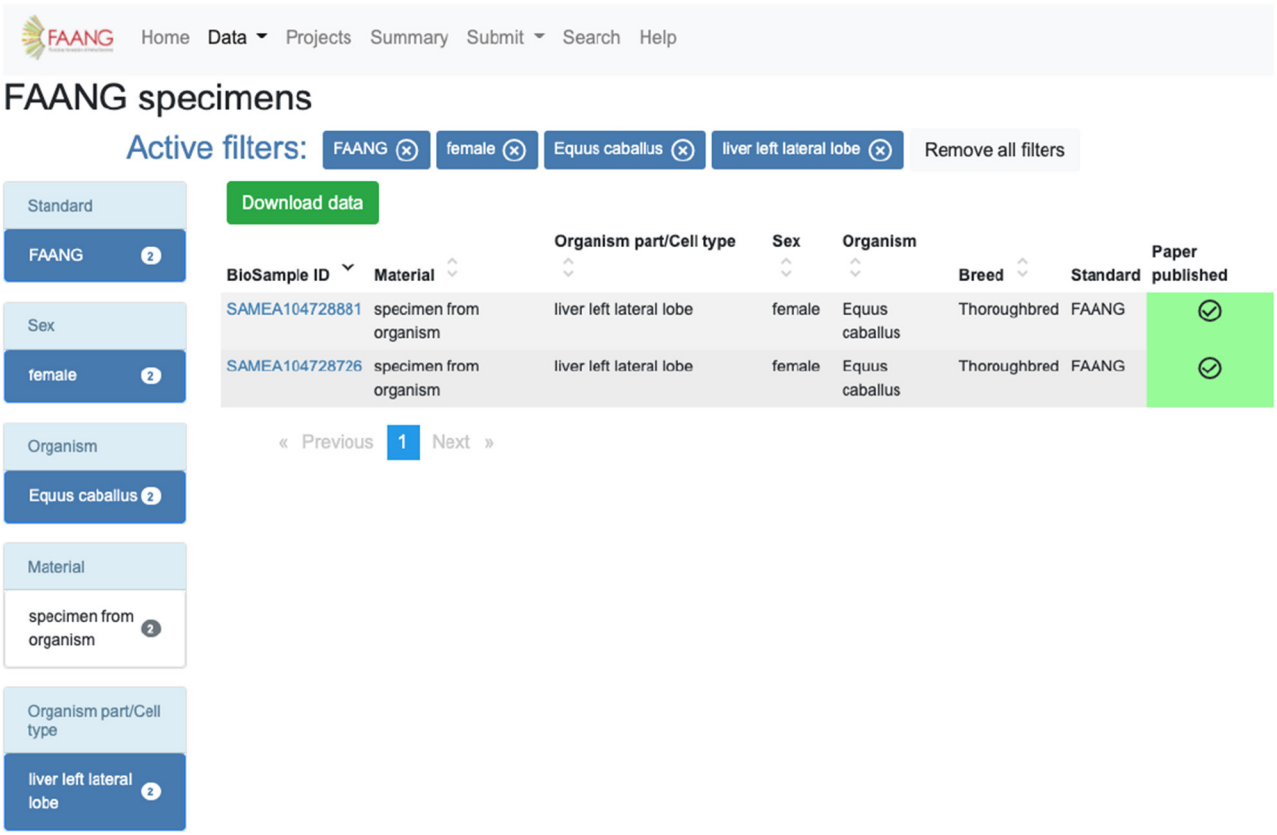


FIGURE 3 | FAANG Data Portal specimen table utilising filters to obtain specimens from *Equus caballus* females from the liver left lateral lobe (https://data.faang.org/specimen?standard=FAANG&sex=female&organism=Equus%20caballus&organismpart_celltype=liver%20left%20lateral%20lobe).

Open-Access and FAIR Data

The DCC strives to meet the highest standards of open and FAIR data recording (Wilkinson et al., 2016). All FAANG data is easily findable as it is assigned a persistent globally recognised identifier by the EMBL-EBI archives, to which the Data Portal brokers submitted datasets. The FAANG metadata standards and associated validation tooling ensures that all data has associated rich metadata, as it holds submission until all standards are appropriately met. The data records are easily found on the Data Portal using any of the interconnected identifiers or a range of preconfigured filters or using the powerful keyword search. The FAANG data is accessible by humans and machines with the persistent identifiers linking directly to the underlying data archives. The data is interoperable through use of open and widely accepted data formats, and all records are ensured to have met the high metadata standards that make extensive use of ontologies for standardising data descriptions. The Data Portal ensures data is reusable by associating mandatory detailed user submitted sampling, sequencing, and analysis protocols to each submitted data record. This provides highly detailed information on each study's methodology, and in even greater detail than the already rich metadata records. A specific protocol browsing page[6] also assists the community in designing future experiments and further standardising how FAANG data is generated for future compatibility of comparative studies.

The FAANG data reuse policy is clearly associated with all FAANG records, and the prepublication data policy is supported through the clear labelling of data within the portal that has been published and thus is free of constraint for further research use. The automated association of publications with data records is particularly important for users to know what records are free from publication restriction in accordance with FAANG's data use policy[7]. This prohibits publication with obtained datasets until the data owners themselves have first published. To aid this, the Data Portal clearly displays with a green tick all datasets that have an associated publication (**Figure 2**). All of the codes for the FAANG Data Portal, data processing, and brokering are freely available under an Apache 2.0 licence[8].

Future Developments

The FAANG project is now moving into its next major phase, with a greater focus on harnessing functional 'omic data from larger populations and leveraging recent technological

[6]https://data.faang.org/protocol/samples

[7]https://www.faang.org/data-share-principle

[8]https://github.com/FAANG/dcc-portal-frontend/blob/master/LICENSE

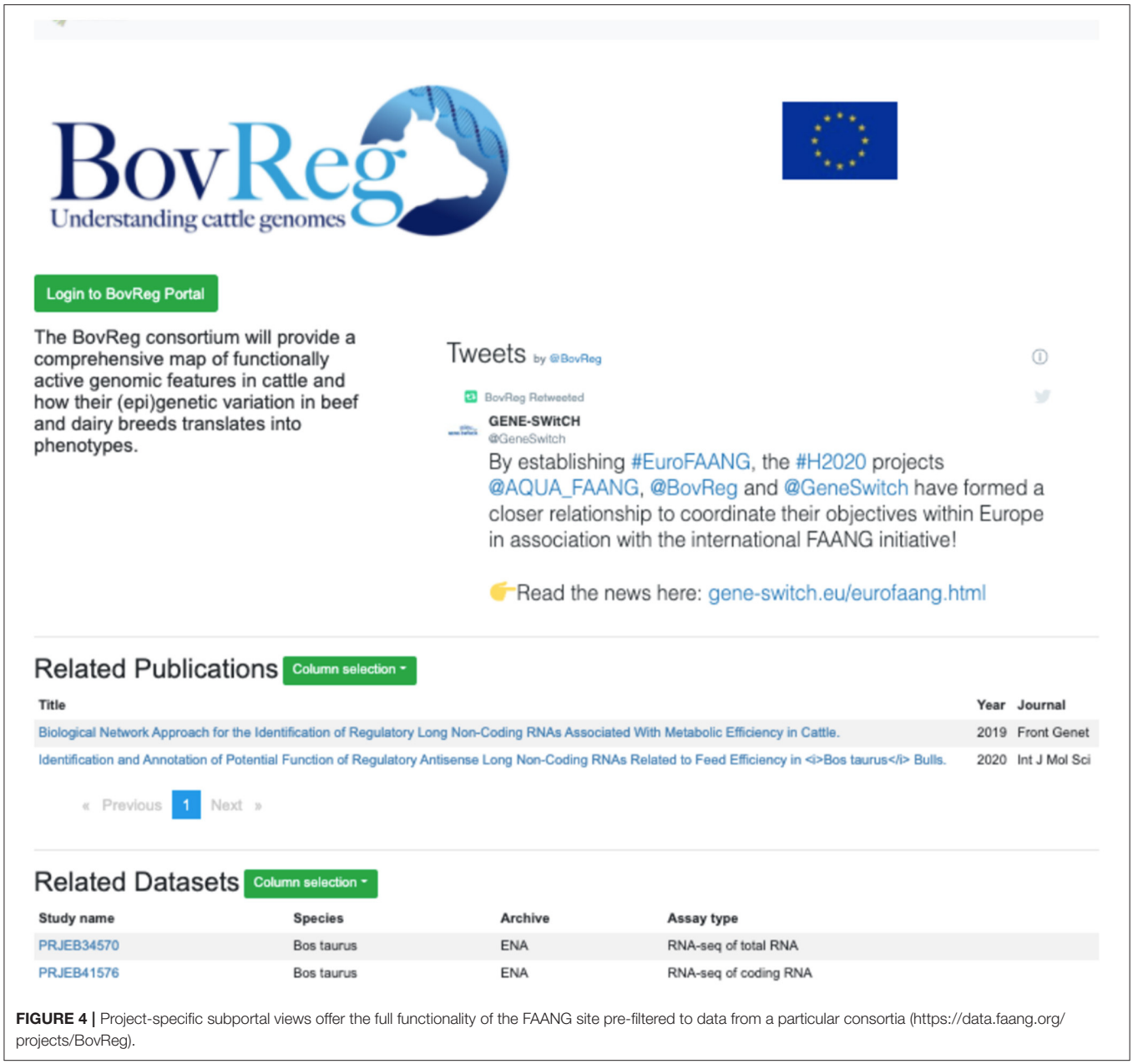


FIGURE 4 | Project-specific subportal views offer the full functionality of the FAANG site pre-filtered to data from a particular consortia (https://data.faang.org/projects/BovReg).

advances such as single-cell 'omics, species and tissue-cell atlases, pangenomes, and novel phenotypic prediction models. Collectively, these will further improve animal genome genotype-to-phenotype annotation and its translation to industrial applications to improve animal production. The Data Portal will continue to evolve alongside the communities' research priorities, and the DCC will develop new infrastructure and site features to effectively deliver and standardise these new data types. For agricultural single-cell 'omics and cell atlases, the DCC will take advantage of the significant prior developments of the ENCODE consortium (Davis et al., 2018) and Human Cell Atlas projects (Regev et al., 2017).

The DCC and the wider FAANG bioinformatics community are focussed on ensuring open reproducible science. There is significant ongoing community effort to create reproducible analysis pipelines. To support this, the Data Portal is already preparing a functionality to link each analysis file to the reusable pipeline that produced it. A new browser page would also create a catalogue of standardised containerised FAANG pipelines to users for further downstream analysis. The Data Portal will look to develop links, wrappers, and infrastructure to enable rapid launching of cloud-based analysis through a range of providers. Discussions are ongoing to support mirroring of FAANG datasets and host the FAANG community's curated bioinformatics pipelines. Alongside the technical infrastructure and standardised pipelines, there is a need to train the current and next generation of scientists to effectively implement them. The FAANG Data Portal could

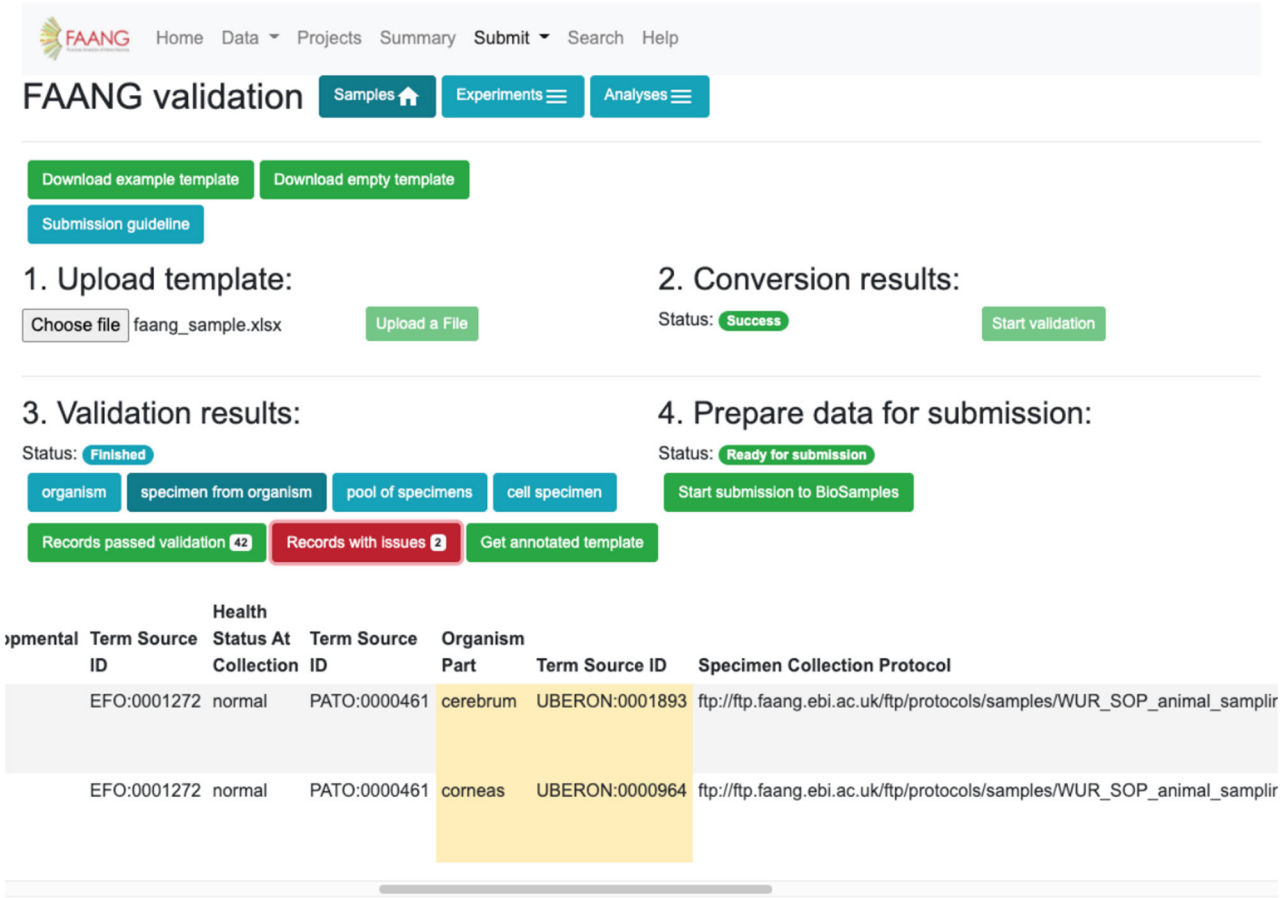


FIGURE 5 | Data validation and submission brokering service flag metadata errors and improvements for correction before submission can be made (https://data.faang.org/validation/samples).

release an online training resource that collates documentation, videos, and webinars on FAANG analysis methods and protocols. This will host resources produced by FAANG projects and links to training upcoming courses. This training, and the distributed data and analysis infrastructure, will be crucial for the successful application of functional data to farmed animal breeding programmes.

The current implementation of search within the data portal is based on inclusive keyword ranking. It is powerful for simple searches to get a view across the different FAANG data types but lacks the desired specificity when multiple search terms are provided. Currently specific multi-term searches need to be performed using the preconfigured filters on the data exploration tables (**Figure 3**) or using the API. To address this, an advanced query language search will be developed that provides a similar search customisation power to that available to programmatic users in the API. The new query language search will allow the user to search by multiple terms linked to specific fields to accurately narrow the returned results to the desired data records. This will mean that a search for "sex = female" and "species = sus scrofa" will only returning female pig records, rather than the current "female" and/or "sus scrofa" search that would return any female or any pig records. This will complement the existing API and data portal table filter searches that already make multiple-field/value search possible.

Automated and standardised visualisation of data across species, systems, tissues, and cell types will also be a key focal area. This includes both automated DCC-generated and user-provided visualisations. FAANG comparative and phenotypic datasets continue to increase in complexity, driving a need to build open-source systems to interrogate and visualise them for research and industrial applications. The Data Portal will also support improved cross referencing and linking to established biorepositories, breeding resources, and key data resources of phenotypic, climate, and functional data. Through its continued key role for FAANG, the Data Portal will continue to support high-quality functional annotation of animal genomes through open sharing of data complete with FAIR standardised rich metadata, and new portal features to support new technological developments for continued improvement in functional annotation of farmed and companion animal genomes.

PROJECT LINKS

The FAANG Data Portal—https://data.faang.org/

The FAANG Data Portal frontend code—https://github.com/FAANG/dcc-portal-frontend

The FAANG Data Portal validation and file conversion code—https://github.com/FAANG/dcc-validate-metadata

The FAANG metadata raw files—https://github.com/FAANG/dcc-metadata

The FAANG Data Portal documentation—https://dcc-documentation.readthedocs.io/en/latest/faq/.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analysed in this study. This data can be found at: https://data.faang.org/home.

AUTHOR CONTRIBUTIONS

PH authored the paper. PH, AN, AS, and JF developed the Data Portal. PH, DZ, GC, and PF managed the project. All authors read and approved the final manuscript.

FUNDING

The FAANG Animal Epigenomes DCC was supported by the Biotechnology and Biological Sciences Research Council (BBSRC; grant nos. BB/N019563/1 and BB/N019202/1), the Horizon 2020 Programme of the European Union under the following awards Aqua FAANG (grant no. 817923), GENE-SWitCH (grant no. 817998), BovReg (grant no. 815668), and the European Molecular Biology Laboratory.

REFERENCES

Amid, C., Alako, B. T. F., Balavenkataraman Kadhirvelu, V., Burdett, T., Burgin, J., Fan, J., et al. (2020). The European nucleotide archive in 2019. *Nucleic Acids. Res.* 48, D70–D76. doi: 10.1093/nar/gkz1063

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* 16:57. doi: 10.1186/s13059-015-0622-4

Clark, E. L., Archibald, A. L., Daetwyler, H. D., Groenen, M. A. M., Harrison, P. W., Houston, R. D., et al. (2020). From FAANG to fork: application of highly annotated genomes to improve farmed animal production. *Genome Biol.* 21:285. doi: 10.1186/s13059-020-02197-8

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46, D794–D801. doi: 10.1093/nar/gkx1081

Giuffra, E., Tuggle, C. K., and Consortium, F. (2019). Functional Annotation of Animal Genomes (FAANG): current achievements and roadmap. *Annu. Rev. Anim. Biosci.* 7, 65–88. doi: 10.1146/annurev-animal-020518-114913

Harrison, P. W., Fan, J., Richardson, D., Clarke, L., Zerbino, D., Cochrane, G., et al. (2018). FAANG, establishing metadata standards, validation and best practices for the farmed and companion animal community. *Anim. Genet.* 49, 520–526. doi: 10.1111/age.12736

Lee, C. M., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J. N., et al. (2020). UCSC genome browser enters 20th year. *Nucleic Acids Res.* 48, D756–D761. doi: 10.1093/nar/gkz1012

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., et al. (2017). The human cell atlas. *Elife* 6:e27041. doi: 10.7554/eLife.27041

Tuggle, C. K., Giuffra, E., White, S. N., Clarke, L., Zhou, H., Ross, P. J., et al. (2016). GO-FAANG meeting: a Gathering On Functional Annotation of Animal Genomes. *Anim. Genet.* 47, 528–533. doi: 10.1111/age.12466

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., et al. (2020). Ensembl 2020. *Nucleic Acids Res.* 48, D682–D688. doi: 10.1093/nar/gkz966

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.

Copyright © 2021 Harrison, Sokolov, Nayak, Fan, Zerbino, Cochrane and Flicek. 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

METHODS
published: 21 June 2021
doi: 10.3389/fgene.2021.666265

Tissue Resources for the Functional Annotation of Animal Genomes

Michèle Tixier-Boichard[1†], Stéphane Fabre[2], Sophie Dhorne-Pollet[1], Adeline Goubil[1], Hervé Acloque[1], Silvia Vincent-Naulleau[3], Pablo Ross[4], Ying Wang[4], Ganrea Chanthavixay[4], Hans Cheng[5], Catherine Ernst[6], Vicki Leesburg[7], Elisabetta Giuffra[1†] and Huaijun Zhou[4†] and the Collaborative Working Group*

[1] University Paris-Saclay, INRAE, AgroParisTech, GABI, Jouy-en-Josas, France, [2] GenPhySe, INRAE, ENVT, ENSAT, Castanet-Tolosan, France, [3] University Paris-Saclay, CEA, IRCM, Fontenay-aux-Roses, France, [4] UC Davis Department of Animal Science, University of California, Davis, Davis, CA, United States, [5] USDA-ARS Avian Disease and Oncology Laboratory, East Lansing, MI, United States, [6] Department of Animal Science, Michigan State University, East Lansing, MI, United States, [7] USDA-ARS, Fort Keogh Livestock and Range Research Laboratory, Miles City, MT, United States

OPEN ACCESS

Edited by:
Sonia Andrade,
University of São Paulo, Brazil

Reviewed by:
Stephen J. Bush,
University of Oxford, United Kingdom
Alex Clop,
Centre for Research in Agricultural Genomics (CRAG), Spain

***Correspondence:**
Michèle Tixier-Boichard
michele.tixier-boichard@inrae.fr

[†] These authors share senior authorship

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

Received: 09 February 2021
Accepted: 29 April 2021
Published: 21 June 2021

Citation:
Tixier-Boichard M, Fabre S, Dhorne-Pollet S, Goubil A, Acloque H, Vincent-Naulleau S, Ross P, Wang Y, Chanthavixay G, Cheng H, Ernst C, Leesburg V, Giuffra E, Zhou H and the Collaborative Working Group (2021) Tissue Resources for the Functional Annotation of Animal Genomes. Front. Genet. 12:666265. doi: 10.3389/fgene.2021.666265

In order to generate an atlas of the functional elements driving genome expression in domestic animals, the Functional Annotation of Animal Genome (FAANG) strategy was to sample many tissues from a few animals of different species, sexes, ages, and production stages. This article presents the collection of tissue samples for four species produced by two pilot projects, at INRAE (National Research Institute for Agriculture, Food and Environment) and the University of California, Davis. There were three mammals (cattle, goat, and pig) and one bird (chicken). It describes the metadata characterizing these reference sets (1) for animals with origin and selection history, physiological status, and environmental conditions; (2) for samples with collection site and tissue/cell processing; (3) for quality control; and (4) for storage and further distribution. Three sets are identified: set 1 comprises tissues for which collection can be standardized and for which representative aliquots can be easily distributed (liver, spleen, lung, heart, fat depot, skin, muscle, and peripheral blood mononuclear cells); set 2 comprises tissues requiring special protocols because of their cellular heterogeneity (brain, digestive tract, secretory organs, gonads and gametes, reproductive tract, immune tissues, cartilage); set 3 comprises specific cell preparations (immune cells, tracheal epithelial cells). Dedicated sampling protocols were established and uploaded in https://data.faang.org/protocol/samples. Specificities between mammals and chicken are described when relevant. A total of 73 different tissues or tissue sections were collected, and 21 are common to the four species. Having a common set of tissues will facilitate the transfer of knowledge within and between species and will contribute to decrease animal experimentation. Combining data on the same samples will facilitate data integration. Quality control was performed on some tissues with RNA extraction and RNA quality control. More than 5,000 samples have been stored with unique identifiers, and more than 4,000 were uploaded onto the Biosamples database, provided that standard ontologies were available to describe the sample. Many tissues have already been used to implement FAANG assays, with published results. All

samples are available without restriction for further assays. The requesting procedure is described. Members of FAANG are encouraged to apply a range of molecular assays to characterize the functional status of collected samples and share their results, in line with the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles.

Keywords: tissue sampling, repository, mammals, bird, cryopreservation, genome

INTRODUCTION

A coordinated genome-wide identification of functional elements in multiple species represents an invaluable resource for the dissection of genotype-to-phenotype relationships. The Functional Annotation of Animal Genome (FAANG) initiative (Andersson et al., 2015; Giuffra et al., 2019) supports the international community in the production of comprehensive maps of functional elements in the genomes of domesticated animal species. An early aspiration of the FAANG Consortium was to create a framework for organizing data standardization, collection, and sharing from many groups (Tuggle et al., 2016). The FAANG data portal[1] has been established to ensure high-quality and rich supporting metadata to describe its farmed and companion animals, samples, and related data sets (Harrison et al., 2018).

In order to generate an atlas of the functional elements driving genome expression in different biological conditions, the FAANG strategy has been to sample many tissues from different species, sexes, ages, and production stages. A consensus was reached at a workshop convened at Plant and Animal Genome (2014) as reported by Andersson et al. (2015). Two FAANG pilot projects [FRAGENCODE for INRAE (Institut national de la recherche agronomique), France, and FarmENCODE for University of California, Davis (UCD), United States] were initially funded to support this effort. This article details the sampling and storage procedures and describes the metadata collection of the reference samples collections for four livestock species (cattle, pig, chicken, goat) realized by these two pilot projects, as well as the guidelines for their possible future use.

ANIMALS

Species and Population of Origin

A prerequisite was to sample species with high-quality genome assemblies. Then, taxonomic diversity was considered: mammals and birds have been sampled, and among mammals, ruminants and non-ruminants have been selected. This article describes the sampling done for *Bos taurus*, *Capra hircus*, *Sus scrofa*, and *Gallus gallus* by INRAE and UCD.

A large choice of breeds is available within each species worldwide. Well-characterized breeds were prioritized and selected for sampling. Regarding cattle, Holstein breed is the most widely used dairy cattle as is Hereford for beef cattle. Regarding goat, one of the two mostly used dairy breeds (the Alpine) was sampled, to allow for comparisons between two species of ruminants for milking traits. Regarding pigs, the sampling included Large White as a dam line and Yorkshire as a sire line. Regarding chickens, the White Leghorn breed was chosen as it provides the genetic basis for numerous experimental lines and is widely used for white egg production. A control line from a selection experiment was sampled, as well as an F1 crossbred obtained from two highly inbred White Leghorn lines differing in disease resistance.

Selection History of the Animal

Animals were chosen so as to be representative of their breed in order to be used as a reference for future studies. They all had a known pedigree, and some of them could also have produced progeny. If possible, frozen semen was collected from males to be able to produce progeny in the future.

Both sexes were sampled, two males and two females for each genetic type (**Table 1**). Adult animals were sampled for all species, considering they were in a stable period for gene expression. They already had performance records, obtained in known environmental conditions. A limited number of physiological states were recommended in order to have more assays from the same tissue in the same individual and have replicates across laboratories and to maximize comparisons across species. For Alpine goats and Large White pigs, blood cells were also sampled at different ages in the young males, in order to allow for a longitudinal analysis of immune traits in the same individual. These young males were the progeny of the adult females that were slaughtered or were closely related. For Large White pigs, blood cells were sampled monthly from weaning at 1 month of age until slaughter at 8 months of age. For Alpine goats, only two blood samples could be collected in the young males.

Flock/Herd/Owner

Animals sampled by INRAE came from its experimental facilities, except for the two Holstein bulls with registration numbers FR2832014033 and FR4934530986 that were purchased from a breeding center in France (*Origen plus*). There is no legal uncertainty regarding the ownership of the biological material sampled from experimental animals, neither for the bulls and their semen, sold by *Origen plus* for research use, without any further conditions.

For the UCD project, two bulls with registration number of 43497294 and 43496857 and two heifers with registration numbers of 43497060 and 43496864 were raised for 12 months in animal facility at Fort Keogh Livestock and Range Research Laboratory in the US Department of Agriculture–Agricultural Research Service (USDA-ARS) and were then transferred to Animal Facility at UCD for another 2 months before the samples collection. Two male and two female chickens were raised

[1] https://data.faang.org/home

TABLE 1 | Number of animals sampled according to species and sex, showing age at sampling according to sex, and reproductive stage for females.

Species	Breed	Female	Age (days)	Stage	Male	Age (days)	Castrated male	Age
Bos taurus	Holstein	2	1250, 1532	Lactating	2	539, 602	–	–
Bos taurus	Hereford (Line 1)	2	420	Post-ovulatory	2	420	–	
Capra hircus	Alpine	2	1697, 2072	Pregnant and lactating	2	Days 49 to 246	–	
Sus scrofa	Large White	2	592, 595	lactating	2	Days 25 to 255		
Sus scrofa	Yorkshire						2	170
Gallus gallus	White Leghorn	2	387	Laying	2	387		
Gallus gallus	White Leghorn F1 crossbred (Lines 6x7)	2	140	Sexually mature but not yet in lay	2	140		

at the USDA, ARS, Avian Disease and Oncology Laboratory (ADOL). Two littermate male pigs were provided by the Michigan State University Swine Teaching and Research Center in East Lansing, MI.

Environmental Conditions

Production System

Cattle: Holstein cows were provided by the INRAE facility Le Pin (latitude 48°44′6.6′ North; longitude 0°9′58.8′′ East). They were raised with a mixed system: in closed barns with freedom of movement from November to April, on grassland from May to October. Inside the barn, they were fed *ad libitum* with a "winter" diet composed of 48% maize silage, 24% green silage with dehydrated pulp, 21% concentrate, 7% rapeseed meal, and 150 g minerals. On grassland, they received daily 2 kg of concentrate with additional complementation with maize silage if necessary. Hereford bulls and heifers were raised at the Fort Keogh Livestock and Range Research Laboratory of USDA-ARS (latitude 45°47′15.4896′′ North, longitude 108°29′21.4944′′ West) with the same mixed system. Inside the barn, they were fed *ad libitum* with a bull's ration of 20% corn, 10% hay, 5% supplement, and 65% silage and a Heifer ration of 39.5% hay, 3.5% supplement, and 57% silage.

Goats: Alpine goats and bucks were provided by the INRAE facility located in Bourges (latitude 47°1′59.98′′ North; longitude 2°39′0′′ East). They were raised in closed barns. Females were fed *ad libitum* with dry hay composed of *Dactylis* and alfalfa. Lactating females received in addition 1.2 kg of concentrate (19% total proteins, 5.3% lipids, 26% starch, 9% raw cellulose, 1% calcium) per day. Males were fed *ad libitum* with dry hay from grass and received 0.6 kg of the same concentrate.

Pigs: Large White pigs were provided by the INRAE facility located in Rouillé (latitude 46°25′0.02′′ North; longitude 0°3′0′′ East). They were housed in groups on straw and fed twice a day with a complete diet (13.5% total proteins, 3% lipids, 6.8% raw cellulose, and 6% ashes, supplemented with minerals, vitamins, and amino acids lysine and methionine) with a total amount of 2.7 to 3.2 kg/day for females, according to pregnancy stage, and of 2.7 to 3.5 kg/day for males according to body weight. Water was provided *ad libitum*. Boars were isolated at the time of semen collection. Two littermate castrated male Yorkshire pigs were provided by the Michigan State University Swine Teaching and Research Center in East Lansing, MI (latitude 42°44′15.5472′′ North; longitude 84°29′′1.6368′′ West). Following weaning at 21 days of age, pigs were housed in groups of 10 with other castrated male pigs, on rounded metal slat flooring with fiberglass-gated sides. Pigs were moved to a grow-finish pen at 65 days of age in groups of 14 pigs, with metal-gated sides and a fully slatted concrete floor. Pigs were fed *ad libitum* with a commercial diet meeting or exceeding the National Research Council (2012) nutritional requirements for each stage of development. Feed was delivered using one self-feeder per pen with 0.61 m of linear feeder space in the nursery and 0.77 m of linear feeder space in the grow-finish pen. Water was provided *ad libitum* with a single nipple drinker in each pen.

Chickens: White Leghorn chickens were provided by the INRAE experimental unit facility located in Nouzilly (latitude 47°32′38′′ North; longitude 0°44′41′′ East). Adults were kept in individual cages for pedigree control and egg recording. Females received 16 h of light per day in a single cycle, and males received 10 h of light per day. Ambient temperature was set at 20°C for females and 19°C for males. They were fed *ad libitum* with a complete diet containing either 17.5% total proteins (supplemented with methionine, lysine, cysteine), 3.3% lipids, 2.6% cellulose, 40% starch, 13% ashes, and 4% calcium for females, or 12.5% total proteins (supplemented with lysine and methionine), 2.8% lipids, 4.2% cellulose, 4.4% ashes, and 0.75% calcium for males (detailed list of compounds can be provided upon request). The Line 6 × 7 F1 chickens were provided by the USDA, ARS, ADOL (latitude 42°44′15.5472′′ North; longitude 84°29′1.6368′′ West) located in East Lansing, MI. Male adults were housed in Horsfall-Bauer isolation units that received 8 h of light per day and kept at 21–27°C. They were fed *ad libitum* "starter" feed crumbles.

Vaccination Program

Cattle: cows were vaccinated against pulmonary infections and enteric diseases during the rearing phase, and each year thereafter for enteric diseases, at the start of the winter period (Rispoval®

RS-BVD) and at the end of it (Coglavax®). In addition, they were vaccinated against neonatal diseases (Trivacton® 6) in the last month of pregnancy.

Female goats were vaccinated against blue tongue virus, and a serological test was performed to check for the absence of brucellosis. As kids, they received only a treatment against coccidiosis (Vecoxan® or equivalent).

Pigs were not vaccinated, but serological tests were performed to check absence of brucellosis, parvovirus, Aujeszky virus, and porcine reproductive and respiratory syndrome.

Chickens received a complete vaccination program from hatch to the adult stage, with vaccinations against Marek disease, Newcastle disease, Gumboro disease, infectious bronchitis, rhinotracheitis, infectious anemia, encephalomyelitis, coccidiosis, and egg drop syndrome (detailed list of vaccines with the calendar can be provided upon request).

Physiological Status at Sampling

Date of birth was recorded for each animal so that age at sampling was precisely known (**Table 1**). Reproductive stage was determined for females; early pregnancy stage was detected in goats upon sampling the uterus (**Table 1**). Animals were fasted for at least 12 h before slaughter.

TISSUE COLLECTION

For INRAE, collection took place after slaughter that was realized according to the authorized practices, without chemical anesthesia. As tissue sampling after death is not submitted to an official permit, the ethical approval was needed in case of blood sampling on live animals. For mammals, blood samples were collected in the context of the approvals (APAFIS/project#): 334-2015031615255004_v4 and 333-2015031613482601_v4 (pigs), 3066- 201511301610897_v2 (cattle), 03936.02, and 8613-2017012013585646_v4 (goats). Chicken immune cells were obtained from spleen sampled after slaughter (no need for animal experiment authorization). For UCD, tissues were collected following Protocol for Animal Care and Use #18464, approved by the Institutional Animal Care and Use Committee (IACUC), UCD. Collection protocols are available from https://data.faang.org/protocol/samples, and a link to each standard protocol is included along with collection description.

For male Large White pigs, scalding was not performed as slaughter took place in an experimental facility at INRAE. Scalding was performed for Yorkshire pigs that were slaughtered in a commercial slaughterhouse. Scalding may expose testis to high temperature stress, but these pigs were castrated, so that no effect of scalding was encountered.

Small cubes of $0.5 \times 0.5 \times 0.5$ cm^3 were sampled from all solid tissues at INRAE (INRA_SOP_tissue_sampling_1a_20160721.pdf) with a total of 2 to 20 replicates from each tissue. At UCD, cross sections of tissue were minced/homogenized using scalpel/scissors to collect subsamples (UCD_SOP_50_TissueCollection_20160520.pdf). In addition, for some complex tissues or special cell preparations, specific protocols were developed: they are mentioned in *Tissue collection* and listed in **Supplementary Table 1**.

The aim of tissue collection was to cover a wide range of tissues for a comprehensive approach of genome annotation. Because of anatomical differences between species, particularly between mammals and birds, some tissues were not collected in all species. Parallel sampling by experts was performed to minimize time to sample preservation; the order of sampling was recorded because it provided an estimate of time elapsed since death. It may be estimated that sampling time varied from 30 min for one chicken to 2 h for one cow. Tissues known to be more susceptible to degradation, such as pancreas or brain, were sampled first. Thus, the majority of tissues were sampled within 30 min postmortem; it was shorter than 30 min for pancreas (<10 min) and brain in mammals, as well as for all tissues in chickens. It was within an hour for digestive and reproductive tracts for cattle, sheep, and pig, except for Holstein where it was up to 2 h.

To classify tissues into different sets, the following parameters were considered including functional importance, standardization of sampling, realized assays, and specific cell preparations. As a consequence, several sets of tissues have been identified (**Table 2**).

Set 1: The Standard Set

This set corresponds to the tissues for which collection is easy to standardize and which will be studied with several assays. It included liver, spleen, lung, heart, skin, fat depot, muscle, and peripheral blood mononuclear cells (PBMCs, i.e., lymphocytes).

Liver samples were taken from the edge of the organ, avoiding proximity with gallbladder and avoiding blood vessels. Gallbladder was collected in the Hereford cow.

The entire spleen was extracted from the abdominal cavity for mammals. Capsule part was removed, and cubes of 0.5-cm-long edges were isolated from the bands. For birds, spleen tissue was either processed the same (UCD project), or a specific procedure was implemented to separate spleen cells from red blood cells, in order to avoid contamination of immune cells by platelet cells (INRA_SOP_chicken_splenocytes_sampling_20160721.pdf).

Lung samples were taken from the edge of the organ, avoiding large bronchioles. Left and right lobes were separately collected in the Hereford cow.

Heart muscle was collected for all animals, with separate collection of left and right ventricles and atria at UCD.

In mammals, a large piece of skin (15×10 cm) from the groin of the right leg was extracted from the carcass. This location was chosen to limit the presence of hairs or bristles. Nevertheless, the entire piece of skin was shaved with a scalpel and finally rinsed with phosphate-buffered saline (PBS) $1 \times$ solution to remove hairs. First, using a circular skin biopsy punch tool (8 mm in diameter) and thereafter a scissor to separate epidermis and dermis from subcutaneous fat tissue, individual skin biopsies were isolated and frozen. When the coat color exhibited different types of pigmentation, biopsies were sampled from contrasted areas, either white or black areas in the Holstein and pale or

TABLE 2 | List of tissues or tissue sections of a given organ, with number of aliquots in the collection, according to species and functional tract.

Tissue	Tissue set	*Bos taurus*	*Capra hircus*	*Sus scrofa*	*Gallus gallus*	Total	Functional Tract
Abdominal fat	1	59	44	48	44	195	Adipose tissue
Subcutaneous fat	1	44	44	44		132	Adipose tissue
Heart	1	64	40	44	40	188	Cardio-respiratory tract
Lung	1	56	41	44	20	161	Cardio-respiratory tract
Trachea	2	8		4	8	20	Cardio-respiratory tract
Abomasum	2	24				24	Digestive system
Cecum	2	24			24	48	Digestive system
Colon	2	68	44	46		158	Digestive system
Duodenum	2	64	40	44	20	168	Digestive system
Esophagus	2	8		4	8	20	Digestive system
Gall bladder	2	8				8	Digestive system
Gizzard	2				39	39	Digestive system
Ileum	2	64	40	44	20	168	Digestive system
Jejunum	2	64	40	44	20	168	Digestive system
Liver	1	88	79	76	60	303	Digestive system
Omasum	2	8				8	Digestive system
Reticulum	2	8				8	Digestive system
Rumen	2	8				8	Digestive system
Salivary gland	2	8				8	Digestive system
Stomach	2			4		4	Digestive system
Pigment epithelium ey	2	8				8	Epithelium
Skin	1	96	48	44	40	228	Epithelium
Tongue superficial	2	8				8	Epithelium
Trachea epithelium	3	44	20	40		104	Epithelium
Bone marrow	2	52	43	47	16	158	Immune tissue
Lymph nodes	2	32	14	16		62	Immune tissue
Peyer's patches	2	30	44	44		118	Immune tissue
Spleen	1	56	56	52	24	188	Immune tissue
Thymus	2	30	21	34		85	Immune tissue
Cerebellum	2	49	41	44	12	146	Nervous system
Frontal lobe (cortex)	2	68	44	48	16	176	Nervous system
Hypothalamus	2	16	8	12	12	48	Nervous system
Medulla	2	8				8	Nervous system
Olfactory lobe	2	32	20	44		96	Nervous system
Pons	2	8				8	Nervous system
Spinal cord	2	8				8	Nervous system
Thalamus	2	8				8	Nervous system
Cervical lining	2	4				4	Female reproductive system
Fornix vagina	2	4				4	Female reproductive system
Infundibulum oviduct	2	8			10	18	Female reproductive system
Isthmus fallopian tube/hen isthmus	2	13			10	23	Female reproductive system
Corpus luteum	2	22	19	2		43	Female reproductive system
Ovarian cortex	2	26	20	19	4	69	Female reproductive system
Ovarian follicle	2	4			10	14	Female reproductive system
Oviduct	2	16	22	22		60	Female reproductive system
Oviductal ampulla/hen magnum	2	18			10	28	Female reproductive system
Theca ovarian follicles	2				10	10	Female reproductive system
Uterus/shell gland	2	32	22	22	10	86	Female reproductive system
Uterovaginal gland	2				10		Female reproductive system
Prostate	2	4				4	Male reproductive system
Bulbourethral	2	4				4	Male reproductive system
Epididymis	2	34	10	10		54	Male reproductive system
Seminal vesicle	2	26	22	22		70	Male reproductive system
Spermatozoon	2	20	10	10	72	112	Male reproductive system

(Continued)

TABLE 2 | Continued

Tissue	Tissue set	*Bos taurus*	*Capra hircus*	*Sus scrofa*	*Gallus gallus*	Total	Functional tract
Testis	2	26	40	40	22	128	Male reproductive system
Vas deferens	2	4			7	11	Male reproductive system
Adrenal cortex	2	52	20	44	6	122	Secretory tissue
Pituitary	2	13	8	6	4	31	Secretory tissue
Mammary	2	24	29	20		73	Secretory tissue
Pancreas	2	52	44	48	24	168	Secretory tissue
Thyroid	2	16	7	9		32	Secretory tissue
Cartilage	2	52	44	47	27	170	Skeleton
Tongue muscle	2	8				8	Striated muscle
Diaphragm	2	8				8	Striated muscle
Dorsal muscle	1	40	40	40		120	Striated muscle
Pectoral muscle	1				32	32	Striated muscle
Sartorius muscle	1				36	36	Striated muscle
Bladder	2	8		2		10	Urinary tract
Kidney	2	16		4	16	36	Urinary tract
Renal cortex	2	40	36	40		116	Urinary tract
Renal medulla	2	44	40	44		128	Urinary tract
Ureter	2	4				4	Urinary tract
Urethra	2	8				8	Urinary tract
Grand total	1878	1204	1322	743		5137	

sustained brown color in the Alpine goat. In chickens, sections of 0.5-cm diameter were sampled from an unfeathered area on the internal face of the leg.

Subcutaneous fat and perirenal abdominal fat were collected for mammals. In addition, mesenteric adipose that lies with layers of the peritoneal mesothelium connecting the small and large intestine was collected at UCD. In chickens, abdominal fat was collected around the gizzard (INRAE) or in the abdominal cavity (UCD).

Muscle samples were taken from the longissimus lumborum for all mammals and from two different muscles for chickens: pectoralis major ("white fibers") from the breast and sartorius or semimembranosus ("red fibers") from the leg. A section from the center of each muscle was collected to avoid adipose tissue and major connective tissue structures. In addition, biceps femoris (bottom/outside round), gluteus medius (top sirloin), longissimus dorsi (ribeye/loin), and psoas major (tenderloin) were also collected in the Hereford cow.

Blood was sampled using EDTA as anticlotting agent. For mammals, whole blood was used to prepare peripheral lymphocytes (called hereafter PBMCs). In order to get a sufficient number of each type of cells, sampling was repeated from the jugular vein between two and five times according to age and species, in the weeks preceding slaughter. Blood was immediately handled to separate PBMCs, as described in FAANG protocols for cattle and goats (INRA_SOP_PBMC_purification_cattle_caprine_20160504.pdf), and for pigs (INRA_SOP_PBMC_seperation_swine_blood_20160504.pdf). As it was not possible to immediately perform cell sorting after each sampling, all PBMCs were frozen before sorting, in order to standardize the preparation of defined populations of lymphocytes.

Set 2: Tissues Requiring Specific Sampling Protocols

Brain Tissues

These tissues are extremely sensitive to degradation. A specific team needs to be in charge of sampling them in the shortest delay after death. In mammals, the brain was separated in four regions: cerebellum, frontal lobe, olfactory lobe, and hypothalamus (INRA_SOP_cattlebrain_sampling_20171104.pdf). The pituitary gland was sampled, and its posterior/anterior parts were separated for Holstein only. In Hereford, the cortex was separated in three subregions: frontal, parietal, and temporal. In addition, pigment epithelium eye, spinal cord, medulla, pons, and thalamus were collected too. In Yorkshire pigs, only cerebellum, cortex, and hypothalamus were collected. In chickens, olfactory bulbs were not dissected, and three parts were dissected: cerebellum, frontal lobe, and hypothalamus. Pituitary was also sampled at INRAE.

Digestive Tract

Mammals

In the Hereford cow and the Yorkshire pig, parotid salivary glands were collected and minced/homogenized using scalpel/scissors. Tongue muscle was collected from approximately halfway in the organ. Superficial tongue sample was collected from the papillary epithelium using a scalpel to separate it from the muscle.

The whole digestive tract was set on a table. The different gut sections were identified, and 10- to 15-cm portions of each region were isolated between two ligatures, after pushing the maximal

amount of the content on each side. Then, each portion was open and rinsed with PBS before sampling. Reticulum, rumen, and abomasum were collected at UCD. At INRAE, duodenum, jejunum, ileum, and colon were collected, keeping the mucosa and the muscular layer together before transfer into individual cryotubes, whereas at UCD, mucosa was scraped from the lumen of the tissue using a clean glass slide. Tissue remaining after mucosal scrapping was saved as smooth muscle sample. This conditioning was applied to abomasum, duodenum, jejunum, ileum, cecum, and colon for which three parts (whole, mucosa, and smooth muscle) were collected. The caudal mesenteric node was identified as the most distal lymph node of the mesenteric chain, and square sections of 0.5 × 1 cm were sampled.

Chicken

Caeca were sampled in place of colon, and gizzard was taken in addition to gut sections. Portions were isolated in a similar way as the one used in mammals, and sections were rinsed with PBS before transferring to individual cryotubes. In addition, mucosal scrapping was performed for half of the aliquots sampled for duodenum, jejunum, and ileum.

Secretory Organs: Mammary Gland, Pancreas, Kidney, Adrenal Glands, and Thyroid

A specific protocol was set up for mammary gland to sample the secretory parenchyma (INRA_SOP_mammarygland_sampling_20171104.pdf). All females were lactating, and the mammary gland was sampled for the cow, goat, and sow.

The entire kidney organs were extracted from the carcass. Capsule was first removed. Each kidney was separated by the middle in two pieces to observe differently colored cortex and medulla parts. Bands of 0.5 cm large were cut with scalpel through the cortex part, and cubes with 0.5-cm-long edges were isolated from the bands and frozen. Cubes of 0.5-cm-long edges were individually dissected from the medulla apparent pyramids. Ureter, bladder, and urethra were also collected in the Hereford cow.

As chicken kidney does not have a similar cortex/medulla structure, it was sampled as a homogenous tissue with several aliquots.

Pancreas and thyroid were collected as quickly as possible without being further dissected. Thyroid could not be found in some individuals.

The cortex of the adrenal glands, which produces cortisol and aldosterone, was dissected in mammals, whereas the whole gland was sampled in chickens.

Reproductive Tract

Mammals

Ovaries underwent a specific dissection protocol, which allowed separating ovocytes from granulosa cells (INRA_SOP_oocytes-granulosa_mammals_sampling_20160721.pdf). In addition, small cubes of 0.5-cm side were cut out from the ovarian cortex and the luteal body (corpus luteum), rinsed in PBS, and transferred individually to cryotubes. There was no luteal body available for sampling in sows.

Uterus samples were taken in the main body of the uterus and included both the mucosa and the muscular layer for all species. Further dissection was implemented for the Hereford cow: tissues from caruncular and intercaruncular regions were collected separately; endometrium was collected by scraping the inner layer of the uterus, and myometrium was isolated from serosa and endometrium by scraping and scalpel. Fourteen sections were then separated: ampula (contralateral to corpus luteum), ampula (ipsilateral to *corpus luteum*), infundibulum (contralateral to *corpus luteum*), infundibulum (ipsilateral to corpus luteum), isthmus of fallopian tube (contralateral), isthmus of fallopian tube (ipsilateral), uterine myometrium, ovarian section (without corpus luteum), uterine endometrium – caruncular (contralateral to corpus luteum), uterine endometrium – caruncular (ipsilateral to corpus luteum), uterine endometrium – intercaruncular (contralateral to corpus luteum), uterine endometrium – intercaruncular (ipsilateral to corpus luteum), Fornix vagina, and cervical lining.

Male reproductive tissues were dissected to separate testis from epididymis and seminal vesicle. Testis was sampled as small cubes or slices, and seminiferous tubules were also dissected and pretreated to implement Hi-C protocol (see *Specific preparations*). For boars and bucks, semen was obtained from epididymis after slaughter and was conditioned with a tris, citrate, and glucose solution supplemented with 15% (vol/vol) egg yolk and 5% (vol/vol) glycerol as described in Pini et al. (2018). For bulls, semen was collected in an accredited artificial insemination center, which provided semen straws that were transferred into a liquid nitrogen tank in order to preserve their fertilizing ability for future functional studies and/or production of progeny.

At UCD, three sections of the epididymis were further separated (caput, corpus, and tail), and prostate, vas deferens, and bulbourethral gland were collected in cow.

Chicken

In females, theca and granulosa cells were separately dissected from the largest follicles, ordered by decreasing size. In addition, very small follicles were preserved in five aliquots.

The presence of an egg in the shell gland was recorded. Several sections of the reproductive tract were sampled: infundibulum (also called oviductal ampula, the closest section from the ovary, where fecundation takes place), magnum (where the albumen is produced), the isthmus, and shell gland (equivalent to the uterus). In addition, glands located at the uterovaginal junction (between the shell gland and the cloaca) were also sampled, as they play a key role in the preservation of spermatozoa after insemination.

Six weeks before slaughter, adult males were trained for 2 weeks in order to collect semen by massage twice a week during 2 weeks. Semen volume, motility, and viability were recorded for each ejaculate. Semen was then diluted with a cryoprotectant agent and frozen in 0.5-mL straws, which were identified by a color code and the unique animal number (INRA_SOP_freezinggallussemen_20200401.pdf). Straws were stored in liquid nitrogen to preserve fertilizing ability of spermatozoon.

At slaughter, as testes are internal organs of homogenous structure, they were separated from vas deferens: small cubes (0.5 cm^3) were cut out either from testis or from vas deferens,

rinsed in PBS and transferred individually to cryotubes. There are no seminal vesicles in chickens.

Immune Tissues: Thymus, Bone Marrow, and Lymph Nodes

In mammals, three types of lymph node were sampled: the caudal mesenteric node, the node located at the trachea-bronchial bifurcation, and the neck lymph nodes. Thymus was collected. Yellow bone marrow cubes of 0.5-cm-long edges were individually collected in the hemimedullary cavity of the tibia. Peyer patches were sampled in cattle, except for Holstein females, in pigs, and in goats.

In chickens, there are no lymph nodes, as well as no Peyer patch. For this species, immune cells were separated from the spleen, as described in *Set 1: the standard set*. Thymus and bone marrow were collected.

Bone and Cartilage

In mammals, the tibia bone of the anterior right leg was extracted from the carcass and cut by the middle in two pieces with a butcher knife. In chicken, the whole tibia bone was cut in 10 small pieces, which were stored in individual cryotubes.

Cartilage was sampled from the femur in mammal species, as described in INRA_SOP_cartilage-sampling_20171117.pdf.

Set 3: Specific Cell Preparations

Immune Cells

$CD4^+$ and $CD8^+$ cells were sorted from PBMCs previously prepared for mammals, following protocols adapted to each species (INRA_SOP_sorting_cattle_CD_cells_20171201.pdf; INRA_SOP_sorting_caprine_CD_cells_20171201.pdf; INRA_SOP_sorting_swine_CD_cells_20160504.pdf). For chickens, $CD4^+$ and $CD8\beta^+$ cells were sorted from purified spleen cells, as described in INRA_SOP_sorting_chicken_CD_cells_20180213.pdf. Alveolar macrophages were separated according to a specific protocol that was applied to mammals only (INRA_SOP_alveolar-macrophages_mammals_sampling_20160721.pdf).

These cells have been stored in liquid nitrogen for future studies.

Epithelial Cells

Tracheal epithelium was dissected from the trachea for mammals only, (INRA_SOP_tracheal_epithelium_mammals_sampling_20160721.pdf) and then stored at −80°C, as other tissues.

TISSUE CONDITIONING

Standard Procedure

Aliquots for homogenous tissues were stored without any buffer into individual cryotubes and immediately snap frozen into liquid nitrogen (INRA_SOP_tissue_sampling_1a_20160721.pdf) on the collection site and then transported in dry ice and placed at −80°C for long-term conservation. Individual aliquots were placed in a single tube, and supernumerary aliquots for a given tissue were pooled, as described in INRA_SOP_tissue_aliquots_sampling_1b_20160721.pdf.

Handling Cellular Heterogeneity Within a Tissue

For very heterogeneous tissues, it is obvious that individual aliquots would not be comparable. Thus, aliquots were not stored individually but were pooled for future studies; this was the case for hypothalamus, pituitary, lymph nodes, spleen, thyroid, and thymus, as described in INRA_SOP_tissue_aliquots_sampling_1b_20160721.pdf.

An additional option was to save tissue morphology for further cellular dissection. For that aim, we used the Optimal Cutting Temperature (OCT) compound to perform embedding of a slice of tissue of 1 cm long, 0.5 cm wide, and 0.3 cm thick in a mold placed on dry ice, (INRA_SOP_tissue_sampling_protocol_6_20180426.pdf). This was done on one aliquot for most tissues, in order to make possible future analysis on identified cell types using either histology or laser microdissection. It was then stored at −80°C.

Specific Preparations

Two types of specific analyses were planned for FAANG: Hi-C and ATAC-Seq. Dedicated cell preparation was performed on fresh samples at the site of sampling in view of HiC, as described in INRA_SOP_liver_spleen_mammarygland_forHiC_sampling_20160721.pdf and INRA_SOP_testis_forHiC_sampling_20160721.pdf, or in view of ATAC-seq, as described in INRA_SOP_ATAC-seq_AG_v1_20160805.pdf for liver, spleen, and $CD4^+$/$CD8^+$ cells from pigs, goats, cattle, and chickens.

At present, ATAC-Seq analyses are known to be possible from snap-frozen tissues. Thus, we can consider that most tissues from this collection are now available for ATAC-seq analyses.

Altogether, 17 specific sampling or conditioning protocols and four general sampling protocols can be found in https://data.faang.org/protocol/samples.

USE OF THE SAMPLES

Sample Description

A total of 3,949 tissue aliquots from the FRAGENCODE project are currently identified with a Biosamples ID (1,184 for cattle, 1,148 for goats, 1,188 for pigs, and 429 for chickens; **Supplementary Table 2**), and a total of 462 samples, with two or three aliquots each, from the FarmENCODE project are available with Biosamples ID (SAMEA4454482-4455404 for chickens, SAMEA4454615-4455481 for cattle, and SAMEA4454570-4454614 for pigs; **Supplementary Table 3**). Ontologies such as UBERON or BRENDA have been used to describe the samples. Additional tissues are stored at INRAE, which require additional curation to get a final ontology, particularly for chicken female reproductive tract, as well as for pigmented or non-pigmented skin in all species. The total number of aliquots preserved is currently 5,137, representing 73 different tissues or tissue section, which can be grouped into 12

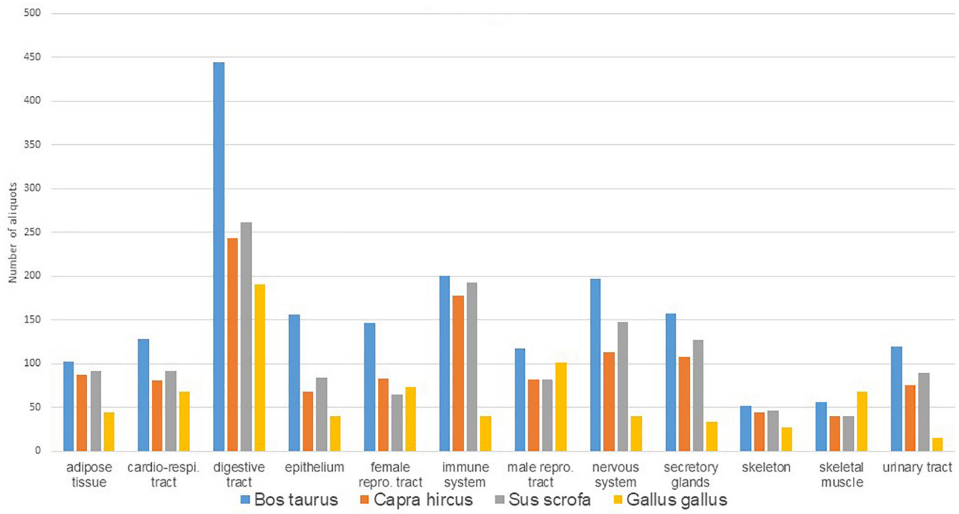


FIGURE 1 | Distribution of number of aliquots available in Biosamples according to species and main tissue types.

main functional categories (**Table 2**). The number of aliquots is shown in **Figure 1** according to species and functional category. There are 21 tissues collected for all four species, and 37 collected for the three mammals. Cattle showed the highest number (67) of tissues or tissue sections collected. Within a species and a sex, the number of tissues was the same per individual. Females had more subsections of the reproductive tract sampled than males, so that the total number of samples was higher for females. Chickens had a lower number of tissues sampled because of anatomical differences (i.e., no lymph nodes, no mammary gland, no subcutaneous fat in the White Leghorns) and also had a lower number of aliquots because of the smaller size of tissues, particularly for brain, but also for kidney, where cortex and medulla are not distinguished as in mammals.

There is no ontology commonly used for breed name and raising conditions; this remains to be validated and used at the international level. A list of breeds can be obtained from the Food and Agriculture Organization database[2], but the naming of breeds is not necessarily harmonized across countries. Consequently, this information has been described with some details in *Animals* of this article, which also provides the link to the sampling protocols.

Quality Control

To validate sampling protocols and verify the RNA integrity, RNA extraction was performed from an aliquot of different tissues or cell types for liver, muscle, mammary gland, lung, spleen, heart, and immune cells. Samples were homogenized in TRIzol Reagent (Life Technologies, Carlsbad, CA, United States) using an Ultra Turrax (IKA) set at 26,000 revolution/min. Total RNAs were extracted according to the manufacturer's instructions (Life Technologies) using optional instructions. Insoluble materials were removed after homogenization by centrifugation at 12,000 *g* for 10 min at 4°C before adding chloroform.

RNA yield and purity were monitored by spectrophotometry (NanoDrop ND-1000). RNA integrity was assessed using an Agilent (Santa Clara, CA, United States) 2100 Bioanalyzer and RNA 6000 nano kits. RNA quality was evaluated using the RNA integrity number (RIN) value introduced by Agilent (Schroeder et al., 2006). RIN values for liver RNA ranged from 7.8 to 8.8 for mammals and from 8.8 to 9.1 in chicken. For immune cells, higher RIN values were obtained in mammals (from 7.7 to 9.7, with a majority of samples with an RIN $>$ 9) than in chickens (from 5.1 to 8.1), which could be due to the more complex separation procedure from spleen cells.

Transcriptome Studies

The FRAGENCODE project aimed at improving the genomic annotation of four species (cattle, goat, chicken, and pig). This was achieved by performing molecular assays on tissue dissociated cells (liver) and on sorted primary cells ($CD4^+$ and $CD8^+$ T lymphocytes) from two males and two females of each species. These assays included RNA-seq, ATAC-seq, and Hi-C to characterize the transcriptome, the chromatin accessibility, and the genome 3D topology in these cells, respectively (Foissac et al., 2019). Additional work was carried out using these RNA-seq datasets for the annotation of long-non-coding RNAs (lnRNAs) (Jehl et al., 2020).

The collection is being used to complete the reference transcriptome of six tissues (cerebellum, lung, kidney, dorsal skin, skeletal muscle, small intestine/Ileum), in addition to the liver datasets reported by Jehl et al. (2020). This additional annotation is being conducted in the frame of the H2020 FAANG project GENE-SWitCH[3].

[2]http://www.fao.org/dad-is/en/

[3]https://www.gene-switch.eu/

The FarmENCODE project was initiated to functionally annotate farm animal genomes (chicken, pig, and cattle), particularly in the regulatory elements. The eight distinct tissues (adipose, cerebellum, cortex, hypothalamus, liver, lung, skeletal muscle, and spleen) from two males of each species have been used to identify tissue-specific expressed mRNAs and lnRNAs across three species (Kern et al., 2018). The ATAC-seq assay on these eight tissues from pig and cattle was performed to analyze chromatin accessibility conservation across mammals (Halstead et al., 2020). Furthermore, ChIP-seq (four histone modification marks and CTCF) assays across three species and DNase-seq in chickens were performed to annotate dynamic chromatin states across tissues and species (Kern et al., 2021).

PROCEDURE TO REQUEST SAMPLES

All tissue aliquots are available for any researcher, provided that the work planned is scientifically sound or will be useful to improve a methodology for FAANG, for example.

Because sanitary conditions have been recorded, health regulations should not be a limitation to access to samples. Since 2014, an EU regulation makes it compulsory to comply with the Nagoya protocol of the Convention on Biological Diversity, regarding access and benefit sharing from the use of genetic resources. In France, there is no access measure for genetic resources from domestic animals, so there is no limitation to access to these samples. United States is not party to the Nagoya protocol, and there is no condition for access either.

The only request for a user of an FAANG sample described in this article is to acknowledge the origin of samples by referring to the present article and to share the results obtained with all partners of the FAANG initiative. A moderate cost can be requested to cover preservation and shipment costs, in order to keep the tissue collection available in the long term.

For the INRAE collection, the procedure to request tissue samples is to create an account on the CRB-Anim web portal, https://crb-anim.fr/access-to-collection/#. The portal provides access to the 3,949 tissue samples also declared in Biosamples. Discovering the whole FRAGENCODE sample collections is possible by a simple browse that will provide information about species, breed, and sample type. You need to identify yourself by creating an account in order to get more precise information about the tissue type and to request samples of interest to you with the advanced search procedure. For any specific question, a contact address is available (contact-crb-anim@inrae.fr).

Access to the UCD collection is possible by contacting the corresponding author from UCD and will be made available from the CRB-Anim web portal in the course of 2021.

CONCLUSION

The FAANG tissue collection set up by INRAE and UCD illustrates the concept of biobank for research in genomics of domestic animals. Whereas the Biosamples database sets the reference identification for biological samples that can be used for research and makes possible to connect molecular data with these samples, additional procedures set up by a biobank are needed to manage the conservation and distribution of samples to the scientific community. To facilitate sample sharing, documenting sampling protocols as well as animal physiological status and raising conditions is needed and has been described in this protocol article. Combining different methods or types of analyses on a limited set of reference animals avoids the random noise due to variation among experiments and makes proposing a reference data set for genome structure and function possible. Furthermore, cryoconservation of spermatozoa enables the production of progeny from these males, for which gene expression profile will have been studied. Once a reference set is defined, targeted experiments with additional sampling will be able to identify deviations from the reference, as long as sampling protocols as well as animal physiological status and raising conditions are known. It is thus highly recommended to upload all sampling protocols in https://data.faang.org/protocol/samples. The preservation and distribution of reference samples, as well as of samples from well-defined experiments, are expected to decrease the number of animals included in future experiments. At present, biobanking stem cells is becoming the priority in order to facilitate the production of organoids, also an alternative to *in vivo* experiments.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: crb-anim.fr/access-to-collection/# for FRAGENCODE project. In addition, excel files extracted from Biosamples for both FRAGENCODE and FarmENCODE projects are provided in **Supplementary Data, Files 2, 3**.

ETHICS STATEMENT

The animal study was reviewed and approved by the INRA Val de Loire Ethical Committee for animal experimentation for blood sampling of live animals and received the approvals (APAFIS/project#): 334-2015031615255004_v4 and 333-2015031613482601_v4 (pigs), 3066- 201511301610897_v2 (cattle), 03936.02 and 8613-2017012013585646_v4 (goats). No approval is needed in France for tissue sampling after slaughter. For UCD, tissues were collected following Protocol for Animal Care and Use #18464, approved by the Institutional Animal Care and Use Committee (IACUC), University of California, Davis.

COLLABORATIVE WORKING GROUP

Catherine Taragnat, Cecile Berri, Déborah Jardet, Estelle Godet, Fabrice Laurent, Gilles Gomot, Hughes Dardente, Isabelle Grasseau, Jean-Philippe Dubois, Joel Gautron, Nadine Gérard, Pascale Quéré, Roger-Paul Lavocat, Rozenn Dalbies-Tran, Sonia Métayer, Sylvain Marthey, Vincent Coustham, and Xavier Druart.

AUTHOR CONTRIBUTIONS

MT-B, SF, SD-P, AG, HA, SV-N, and EGi designed and implemented the FRAGENCODE project. PR, YW, GC, HC, CE, VL, and HZ designed and implemented the FarmENCODE project. MT-B drafted the manuscript with HZ and EGi. Authors from the collaborative working group CB, VC, RD-T, HD, XD, J-PD, JG, EGo, GG, NG, IG, FL, SMé, PQ, and CT organized the sampling and brought a special expertise on a tissue or a particular species. DJ, R-PL, and SMa identified all samples and handled the data from collection to submission to Biosamples for FRAGENCODE project. All authors contributed to the article and approved the submitted version.

FUNDING

This work received funding for tissue collection and storage from the CRB-Anim infrastructure project, ANR-11-INBS-0003, funded by the French National Research Agency in the frame of the "Investing for the Future" program. The FR-AgEncode project was funded by the SELGEN program of INRAE. The FarmENCODE project was supported by United States Department of Agriculture, National Institute of Food and Agriculture (Competitive Grant Project No. 2015–67015-22940 to HZ). Further support was provided by United States Department of Agriculture, National Institute of Food and Agriculture, National Animal Genome Research Program, U.S. Poultry, Cattle, and Swine Genome Coordination Fund, National Pork Board, and Aviagen Ltd.

ACKNOWLEDGMENTS

The authors are grateful to Sasha Aimee Kaufman Dubrovsky, Joseph Ryan Owen, Justin Buchanan, Bryan Thomas Welly, James L. Chitwood, Jennifer Rose Hughes, Karen Tracy, Colin Kern, Huazhen Liu, Huai-Chen Chang, Khin Khine Zar Mon, Samgenie lee, Perot Saelao, Shuang Niu, Sihua Jin, Susan Margaret Waters, Tae Hyun Kim, Catherine Ann Lopez, Yanina Soledad Bogliotti, Vivian Yu, Michelle Halstead, Alison Van Eenennaam, at UCD, Scott Funkhouser and Deborah Velez-Irizarry at MSU, Laurie Molitor at USDA-ADOL, for their help during tissue collection of the FarmENCODE project. Damien Capo, Yoann Bailly, Stéphane Ferchaud, Tony Terrasson, Yves Gallard, Yannick Baumard, Harold Rigoreau, Frédéric Bouvier and Thierry Fassier are gratefully acknowledged for their help in the choice and preparation of animals at INRAE. The proper sampling of tissues for the FRAGENCODE project would not have been possible without the help of Thierry Bordeau, Aurélien Brionne, Magali Chessé, Paul Constantin, Nathalie Couroussé, Michel Couty, Sandrine Crochet, Maryse Lecomte, Nathalie Même, and Cindy Riou for chickens, and the help of Claire Chevaleyre, Françoise Drouet, Michel Olivier, Emmanuelle Rebours, Marie-Noëlle Rossignol, Laurence Dufourny, Pascal Froment, Yann Locatelli, Cecile Mansanet, Martine Migaud, and Danielle Monniaux for mammals.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.666265/full#supplementary-material

REFERENCES

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project. *Genome Biol.* 16:57. doi: 10.1186/s13059-015-0622-4

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., et al. (2019). Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol.* 17:108. doi: 10.1186/s12915-019-0726-5

Giuffra, E., Tuggle, C. K., and Faang Consortium (2019). Functional annotation of animal genomes (FAANG): current achievements and roadmap. *Annu. Rev. Anim. Biosci* 7, 65–88. doi: 10.1146/annurev-animal-020518-114913

Halstead, M. M., Kern, C., Saelao, P., Wang, Y., Chanthavixay, G., Medrano, J. F., et al. (2020). A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. *BMC Genomics* 21:698. doi: 10.1186/s12864-020-07078-9

Harrison, P. W., Fan, J., Richardson, D., Clarke, L., Zerbino, D., Cochrane, G., et al. (2018). FAANG, establishing metadata standards, validation and best practice for the farmed and companion animal community. *Anim. Genet.* 49, 520–526. doi: 10.1111/age.12736

Jehl, F., Muret, K., Bernard, M., Boutin, M., Lagoutte, L., Désert, C., et al. (2020). An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. *Sci. Rep.* 10:20457. doi: 10.1038/s41598-020-77586-x

Kern, C., Wang, Y., Chitwood, J., Korf, I., Delany, M., Cheng, H., et al. (2018). Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. *BMC Genomics* 19:684. doi: 10.1186/s12864-018-5037-7

Kern, C., Wang, Y., Xu, X., Pan, Z., Halstead, M., Chanthavixay, G., et al. (2021). Functional genome annotations of three domestic animal species provide a vital resource for comparative and agricultural research. *Nat. Commun.* 12:1821. doi: 10.1038/s41467-021-22100-8

National Research Council (2012). *Nutrient Requirements of Swine. 11th Revised Edn*. Washington, DC: National Academies Press.

Pini, T., Rickard, J. P., Leahy, T., Crossett, B., Druart, X., and de Graaf, S. P. (2018). Cryopreservation and egg yolk medium alter the proteome of ram spermatozoa. *J. Proteomics* 181, 73–82. doi: 10.1016/j.jprot.2018.04.001

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., et al. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* 7:3. doi: 10.1186/1471-2199-7-3

Tuggle, C. K., Giuffra, E., White, S. N., Clarke, L., Zhou, H., Ross, P. J., et al. (2016). GO-FAANG meeting: a gathering on functional annotation of animal genomes. *Anim. Genet.* 47, 528–533. doi: 10.1111/age.12466

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.

Copyright © 2021 Tixier-Boichard, Fabre, Dhorne-Pollet, Goubil, Acloque, Vincent-Naulleau, Ross, Wang, Chanthavixay, Cheng, Ernst, Leesburg, Giuffra, Zhou and the Collaborative working group. 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: 23 June 2021
doi: 10.3389/fgene.2021.664379

Putative Causal Variants Are Enriched in Annotated Functional Regions From Six Bovine Tissues

Claire P. Prowse-Wilkins[1,2], Jianghui Wang[2], Ruidong Xiang[1,2], Josie B. Garner[3], Michael E. Goddard[1,2] and Amanda J. Chamberlain[2]*

[1] Faculty of Veterinary and Agricultural Science, The University of Melbourne, Parkville, VIC, Australia, [2] Agriculture Victoria, AgriBio, Centre for AgriBiosciences, Bundoora, VIC, Australia, [3] Agriculture Victoria, Ellinbank Dairy Centre, Ellinbank, VIC, Australia

OPEN ACCESS

Edited by:
Eduard Murani, Leibniz Institute for Farm Animal Biology (FBN), Germany

Reviewed by:
Martijn Derks, Wageningen University and Research, Netherlands
David E. MacHugh, University College Dublin, Ireland

***Correspondence:**
Claire P. Prowse-Wilkins
claire.prowse-wilkins@agriculture.vic.gov.au

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

Received: *05 February 2021*
Accepted: *24 May 2021*
Published: *23 June 2021*

Citation:
Prowse-Wilkins CP, Wang J, Xiang R, Garner JB, Goddard ME and Chamberlain AJ (2021) Putative Causal Variants Are Enriched in Annotated Functional Regions From Six Bovine Tissues. Front. Genet. 12:664379. doi: 10.3389/fgene.2021.664379

Genetic variants which affect complex traits (causal variants) are thought to be found in functional regions of the genome. Identifying causal variants would be useful for predicting complex trait phenotypes in dairy cows, however, functional regions are poorly annotated in the bovine genome. Functional regions can be identified on a genome-wide scale by assaying for post-translational modifications to histone proteins (histone modifications) and proteins interacting with the genome (e.g., transcription factors) using a method called Chromatin immunoprecipitation followed by sequencing (ChIP-seq). In this study ChIP-seq was performed to find functional regions in the bovine genome by assaying for four histone modifications (H3K4Me1, H3K4Me3, H3K27ac, and H3K27Me3) and one transcription factor (CTCF) in 6 tissues (heart, kidney, liver, lung, mammary and spleen) from 2 to 3 lactating dairy cows. Eighty-six ChIP-seq samples were generated in this study, identifying millions of functional regions in the bovine genome. Combinations of histone modifications and CTCF were found using ChromHMM and annotated by comparing with active and inactive genes across the genome. Functional marks differed between tissues highlighting areas which might be particularly important to tissue-specific regulation. Supporting the cis-regulatory role of functional regions, the read counts in some ChIP peaks correlated with nearby gene expression. The functional regions identified in this study were enriched for putative causal variants as seen in other species. Interestingly, regions which correlated with gene expression were particularly enriched for potential causal variants. This supports the hypothesis that complex traits are regulated by variants that alter gene expression. This study provides one of the largest ChIP-seq annotation resources in cattle including, for the first time, in the mammary gland of lactating cows. By linking regulatory regions to expression QTL and trait QTL we demonstrate a new strategy for identifying causal variants in cattle.

Keywords: bovine, ChIP-seq, histone modifications, function, causal variants, differential binding, annotation, ChromHMM

INTRODUCTION

Finding the genetic variants which lead to different phenotypes has been the goal of geneticists for many years. Bridging the genotype to phenotype "gap" has linked genes to their functions and identified causes of disease. In the dairy industry, finding genetic variants which affect phenotypes would improve selective breeding using genomic selection (MacLeod et al., 2016). Genomic selection relies on associations between genotypes and phenotypes to predict the phenotypes of animals. But this association could be based on linkage disequilibrium (LD) between a SNP and the causal variant rather than a direct effect of the SNP itself. Therefore, identifying the causal variant would prevent breakdown of LD over time and extend genomic predictions to populations with different LD (Hayes et al., 2016).

Most traits of interest to the dairy industry are complex traits which are predicted to have many causal variants of small effect (Goddard et al., 2016). Because the individual effect of each causal variant is small, these are difficult to find using classical genetics methods, especially in a large long-lived mammal. Studies in humans and other species have shown that trait-associated variants from genome wide association studies (GWAS) are enriched in functional regions of the genome, such as regulatory or protein coding regions (Maurano et al., 2012; Schaub et al., 2012; Trynka et al., 2013; Ma et al., 2015). Apart from genes, most functional regions are not well annotated in the bovine genome which has hampered attempts to ask the same question in cattle (Koufariotis et al., 2014). However, two recent studies found that functional regions identified in cattle were more likely to contain QTL than other regions (Wang et al., 2017; Fang et al., 2019). This reveals the exciting proposition that if functional regions can be identified in the bovine genome, this can narrow down the space in which we search for causal variants.

While genes can be broadly identified using sequence homology, other types of functional elements do not have easily identifiable features, lack sequence conservation and can be located far from genes (Kellis et al., 2014). One study, using homology with human regulatory regions, indicated that these regions in cattle were enriched for QTL (Nguyen et al., 2018), however, recent evidence showed that the use of functional annotations predicted from humans in cattle is limited (Xiang et al., 2019; Raymond et al., 2020). This suggests that identifying functional regions directly in cattle is optimal. Accordingly, the FAANG consortium for the Functional Annotation of ANimal Genomes has been set up to jointly annotate functional regions in livestock genomes by assaying directly for them in relevant species and tissues (Andersson et al., 2015).

Some functional regions are marked by histone modifications – post translational modifications to the histone proteins which DNA is wrapped around in the cell (Zhou et al., 2011). For example, the histone protein H3 has a tail which can be modified by mono (H3K4Me1) or tri-methylation (H3K4Me3) at its 4^{th} lysine (Kimura, 2013). Numerous studies (Bernstein et al., 2005; Roh et al., 2006; Heintzman et al., 2007) have found that H3K4Me3 is found at promoters of genes, with one study (Guenther et al., 2007) showing that up to 75% of genes in human embryonic stem cells were marked by H3K4Me3 at their promoter. Other studies have found that H3K4Me1 also marks promoters (Barski et al., 2007; Robertson et al., 2008) and that the regulatory DNA sequences called enhancers, which enhance the transcription of genes (Pennacchio et al., 2013), are marked by H3K4Me1 and sometimes H3K4Me3 (Barski et al., 2007; Robertson et al., 2008; Spicuglia and Vanhille, 2012). The tail of histone H3 can also be modified by acetylation (H3K27ac) or tri-methylation (H3K27Me3) at its 27^{th} lysine (Kimura, 2013). Studies have found that active genes and enhancers tended to be marked by H3K27ac (Creyghton et al., 2010) while repressed regions were marked by H3K27Me3 (Zhao et al., 2007; Tie et al., 2009). Potential functional regions can also be marked by other factors. The zinc finger protein CTCF (CCCTC-binding factor) has many functions in the genome. CTCF acts as a transcription factor which can block and activate gene expression, an insulator by blocking interactions between enhancers and promoters, and is involved in the machinery that regulates chromatin conformation (Ong and Corces, 2014; Kim et al., 2015). Assaying the location of these five marks should identify the location of many functional regions in the bovine genome.

The locations of histone modifications and transcription factors can be assayed across the genome using Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) (Park, 2009). Chromatin is fixed so that DNA is bound to the proteins it is interacting with and antibodies are used to isolate the protein of interest, such as a histone modification or transcription factor. The DNA bound to these proteins is then sequenced and aligned to the genome with reads forming "peaks" at the location where the protein was bound. These peaks can be used to annotate putative functional regions in the genome singularly, or by combining data from several proteins (Park, 2009; Ernst and Kellis, 2010). The height of the peaks (characterised by read counts) is also useful. Peak height has been used to predict the expression of nearby genes (Karlić et al., 2010) and variants which associate with peak height have been shown to overlap with variants associated with gene expression (McVicker et al., 2013).

Functional regions have been identified in the bovine genome in liver (Villar et al., 2015), rumen epithelial cells (Fang et al., 2019) and other tissues (Kern et al., 2021). However, functional regions can vary between tissues (Kellis et al., 2014). This study aimed to increase the catalog of functional regions in the bovine genome by using ChIP-seq to assay the genomic locations of one transcription factor and 4 histone modifications in 6 tissues in 2–3 lactating Holstein dairy cows (**Figure 1**). This data was used to annotate putative functional regions in the bovine genome and identify tissue specific functional regions. We showed that peak height in some regions correlated with the level of gene expression in nearby genes. Lastly, we confirm that QTL and eQTL are enriched in these putative functional regions.

MATERIALS AND METHODS

Chromatin Immunoprecipitation and RNA Sequencing

Tissue from liver, lung, mammary gland, kidney, heart, and spleen from three lactating Holstein dairy cows were sampled

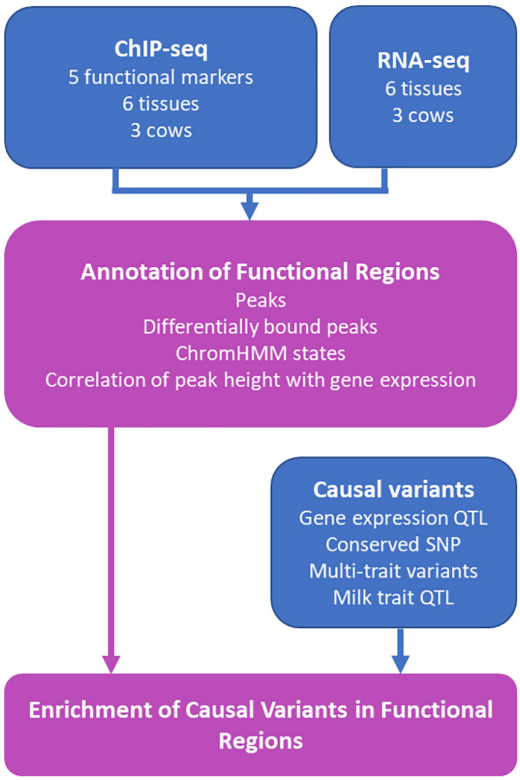


FIGURE 1 | Schematic of Study. Results from the study are outlined in purple boxes while data sources contributing to the results are represented in blue. RNA-seq and ChIP-seq data was generated from heart, liver, lung, kidney, mammary gland, and spleen from the same three cows. The five functional markers assayed were H3K4Me1, H3K4Me3, H3K27Me3, H3K27ac, and CTCF. Details of causal variants are outlined in **Table 1**.

after euthanasia (Chamberlain et al., 2015; Dorji et al., 2020). Ethics approval for the euthanasia and sampling of two of the cows were obtained from Department of Jobs, Precincts and Regions Ethics Committee (Application No. 2014-23). These animals were euthanised out of line of sight of other animals and sedated with 600mg of xylazine IV and 300mg of ketamine before 1l of 25% magnesium sulphate was injected intravenously until animal was deceased. The Third cow was not euthanised for the purposes of this study but because she injured her leg, for this reason the local Animal Ethics Commitee (DEPI Agricultural Research and Extension Animal Ethics Commitee) advised ethics approval was not required. This animal was euthanised by captive bolt.

Tissues were dissected, snap frozen in liquid nitrogen and stored at −80°C until use. Frozen tissue was ground for 3 min in the Geno/Grinder (SPEX SamplePrep). Ground, frozen tissue was fixed for 10 min with 10% formaldehyde and chromatin prepared using the Magnify Chromatin Immunoprecipitation kit (ThermoFisher) as per the manufacturer's instructions. Fixed chromatin was sheared to 200–500 bp using the Covaris S2 (Covaris). Mammary and liver chromatin was sheared for 3 min, duty cycle five, %intensity four and 200 cycles per burst 200. The remaining tissue were sheared at duty cycle two, %intensity three, 200 cycles per burst for 5–15 min.

Chromatin immunoprecipitation was performed using the Magnify Chromatin immunoprecipitation kit (ThermoFisher) with some modifications. For the mammary and liver each sample was immunoprecipitated in three separate reactions with 10 ul of chromatin and 0.25 ug or 0.5 ug of antibody for the histone modifications and 10 µl of antibody for CTCF. Triplicate samples were combined after de-crosslinking using MinElute PCR purification kit (QIAGEN). For spleen, heart, kidney and lung, 1.5–10 µg chromatin was used for immunoprecipitation with 0.5–0.15 µg antibody (H3K4Me1, H3K4Me3, H3K27ac, and H3K27Me3) per reaction or 10 µl of CTCF antibody. The DNA obtained from ChIP was purified and concentrated using Monarch Genomic DNA Purification Kit (New England Biolabs).

Sequence libraries for ChIP and a corresponding input sample were prepared with the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) each with unique barcodes as per the manufacturer's instructions and run on the HiSeq 3000 (Illumina) in a 150 cycle paired end run.

Each library was sequenced to between 20 and 300 million reads. Raw sequence reads were trimmed of adapters and poor-quality ends using Trimmomatic (Bolger et al., 2014) removing base pairs from the 3′ and 5′ ends of the sequence if their quality was less than 20 and excluding trimmed reads with length less than 50 bp. Trimmed reads were mapped to the bosTau8/UMD-3.1.1 bovine genome with BWA mem using default settings (Li, 2013). Poor quality reads were removed with Samtools (Li et al., 2009) using $q > 15$ and duplicate reads removed. ChIP and input reads were used to call peaks with MACS2 default settings (Zhang et al., 2008). Quality checks of peaks was performed with deepTools plotFingerprint (Ramírez et al., 2016) and SPP (Kharchenko et al., 2008).

RNA extraction and sequencing on the same six tissues in the three cows was as described in Chamberlain et al. (2015) and Dorji et al. (2020).

Profile Plots and Replicate Comparisons

Plots for the profile of ChIP-seq reads for each mark were generated using deepTools (Ramírez et al., 2016). First bigWig files were created from mapped ChIP and input reads using the command bamCompare with bin size 10 and using the RPKM option (Reads Per Kilobase of transcript, per Million mapped reads) to normalise the number of reads between samples. Output bigWig scores were the $\log_2$ ratio of ChIP to input. The command computeMatrix was used to calculate scores at each mark around transcription start sites (TSS) taken from Ensembl [Release 94 (Hunt et al., 2018)]. Active and inactive TSS were determined using the RNAseq data described in Chamberlain et al. (2015). Active TSS were defined when all samples had a count more than 200 at that gene and inactive when all animals had a count less than 10. Matrices were visualised using the command plotProfile.

Similarity between samples was also calculated using deepTools plotCorrelation (Ramírez et al., 2016). The command multiBigWigSummary was used to summarise bigwig files across each mark in bins of 100 bp. The command plotCorrelation

was used to generate a heatmap of these values using Pearson correlations.

Differential Binding Between Tissues

Differential binding (DB) between tissues was calculated using edgeR (Robinson et al., 2010) for consensus peaks. To define a set of consensus peaks for each mark each position in the genome was defined as under a peak or not. Positions which were under a peak in two or more samples were included in the consensus peak dataset. Reads were normalised using edgeR and DB tested by defining a design matrix for which the intercept was the mean in all tissues. Peaks were considered significantly differentially bound with a P value less than 0.05 and binding greater than two-fold different to the average binding of that peak in all other tissues.

TABLE 1 | Putative causal SNPs.

Dataset	Number of SNPs	Description	References
Allele specific eQTL	1,100,446	Allele specific expression QTL from white blood cells and milk cells in 112 holstein cows ($P < 0.0001$)	Chamberlain et al. (2018)
Exon eQTL	945,832	Exon expression QTL from white blood cells, milk cells, liver, and muscle in 209 holstein cows ($P < 0.0001$)	Xiang et al. (2018); Xiang et al. (2019)
Gene eQTL	110,200	Gene expression QTL from white blood cells, milk cells, liver, and muscle in 209 holstein cows ($P < 0.0001$)	Xiang et al. (2018); Xiang et al. (2019)
Conserved regions	378,472	SNPs conserved in 100 species lifted over from human to bovine genome	Xiang et al. (2019)
SNP 80k	83,454	Top 80,000 sequence variants ranked for their contributions to 34 traits	Xiang et al. (2019); Xiang et al. (2021)
Splice QTL	1,112,324	Splice QTL from blood, milk cells, liver, and muscle in 209 holstein cows ($P < 0.0001$)	Xiang et al. (2018); Xiang et al. (2019)
Protein Yield QTL	3,317	QTL from GWAS in 32,347 cows for protein yield with $P < 1 \times 10^{-7}$	Xiang et al. (2020)
Fat yield QTL	4,815	QTL from GWAS in 32,347 cows for fat yield with $P < 1 \times 10^{-7}$	Xiang et al. (2020)
Milk Yield QTL	6,883	QTL from GWAS in 32,347 cows for milk yield with $P < 1 \times 10^{-7}$	Xiang et al. (2020)
Fat percentage QTL	12,373	QTL from GWAS in 32,347 cows for fat percentage with $P < 1 \times 10^{-7}$	Xiang et al. (2020)
Protein percentage QTL	17,012	QTL from GWAS in 32,347 cows for protein percentage with $P < 1 \times 10^{-7}$	Xiang et al. (2020)

Details of each SNP dataset used in enrichment analysis.

Correlation Between Peak Height and Gene Expression Level

The correlation between consensus peak counts and gene count for every peak within 100 kb of a gene TSS across 16 samples (4 tissues X 3 cows and 2 tissues X 2 cows) for H3K27Me3 and H3K27ac or 18 samples (6 tissues X 3 cows) for the other marks was calculated.

Counts for each set of consensus peaks (described above) were calculated with DiffBind (Stark and Brown, 2011) and normalised using Trimmed Mean of M-values (TMM) and full library size. Normalised RNA-seq counts for each gene were used from Chamberlain et al. (2015) and Dorji et al. (2020). Correlations were calculated using corr.test in R.

Annotation With ChromHMM

Chromatin states were defined using ChromHMM (Ernst and Kellis, 2012). Filtered and deduplicated input and ChIP-seq bam files were binarized using the binarizeBam option with default settings. Between five and 24 states were learned using the LearnModel function of ChromHMM. A model with 7 states was chosen for further analysis as in this model each state was unique.

To annotate, chromatin states were compared to known regions in the genome using the OverlapEnrichment function in ChromHMM. Locations of TSS and genes were downloaded from Ensembl [release 94 (Hunt et al., 2018)]. The promoter region was characterised as 2kb upstream of the TSS and proximal regions as 8 kb upstream from the promoter start site. Normalised read counts from RNA-seq data (Chamberlain et al., 2015; Dorji et al., 2020) were used to define genes, and their associated promoters and proximal regions, as active or inactive. Active genes were defined when all animals had a count over 200 in that tissue and for inactive genes all animals had a count less than 10 in that tissue. We also looked for tissue-specific active genes, these were when all animals had counts above 200 in that tissue and below 10 in all other tissues.

Enrichment of Putative Causal SNPs in Functional Regions

Enrichment of putative causal SNPs in peaks, peaks which correlated with gene expression, differentially bound peaks and ChromHMM states was calculated. Enrichment was calculated using the formula outlined in ChromHMM (Ernst and Kellis, 2012) and described below.

Enrichment = (C/A)/(B/D) where: A is the number of positions under peaks, B is the number of positions that were putative causal SNPs, C is the number of positions under peaks and also a putative causal SNP and D is the number of positions in the genome.

The significance of enrichment or depletion was calculated using a hypergeometric test in R. A variety of putative causal SNP datasets were used for enrichment analysis (**Table 1**). There were no SNP on the X chromosome in the datasets so only autosomal regions were tested. Total genome size was calculated as the sum of chromosomes 1–29.

Enrichment Based on Location in the Genome

We tested whether enrichment of putative causal SNPs within peaks changed depending on location in the genome. Enrichment was calculated with only peaks and SNPs located within 100 kb from a TSS, 100–200 kb from a TSS and so on up to a million base pairs from a transcription start site. Transcription start sites for each gene were taken from Ensembl [Release 94 (Hunt et al., 2018)]. The distance to the nearest TSS for each peak was calculated from the summit (as defined by MACS2) of narrow peaks and the midpoint of broad peaks (halfway between the start and end of the peak).

Filtering SNPs for Linkage Disequilibrium

To account for SNPs in high linkage disequilibrium (LD), the GWAS SNP datasets were filtered for LD using plink (Purcell et al., 2007). Within each 1mb window, with a moving step of 50 SNPs, SNPs were retained if they had pairwise $r^2 < 0.5$. Enrichment was calculated with these filtered SNP datasets as described above.

RESULTS

Description of ChIP-seq Peaks

86 ChIP-seq datasets were generated as shown in **Supplementary Table 1**.

Quality of the ChIP-seq assay was assessed by calculating the Jensen-Shannon Distance (JSD) between the sample and input and cross strand correlation metrics. All but two samples exceeded ENCODE standards for cross correlation metrics (NSC > 1.05 and RSC > 0.9, **Supplementary Table 2**). However, we found JSD to be a more reliable metric, which was less sensitive to read depth. All ChIP-seq data had a JSD between 0.23 and 0.5 (**Supplementary Table 2**).

After filtering of poor-quality reads and removal of duplicates, between 24,847,326 and 316,216,350 mapped ChIP-seq reads remained for each sample (**Supplementary Table 2**). To account for bias in shearing, library preparation and mapping, an input control using the same batch of sheared chromatin was also sequenced for each sample. Mapped ChIP-seq and input control bam files were used in MACS2 (Zhang et al., 2008) to call between 31,303 and 871,452 peaks for each sample. The average size of peaks was 400 bp to 600 bp for peaks designated as "narrow" in MACS2 (H3K27ac, H3K4Me3, and CTCF) and 1000 bp to 1100 bp for peaks called as "broad" (H3K4Me1 and H3K27Me3, **Table 2**). An average of 13% of the genome was under a peak in any one dataset. When considering narrow peaks, the percentage of genome covered in each dataset was strongly correlated with number of mapped reads ($r = 0.715$, $P < 0.001$) but was less strongly correlated in broad peaks ($r = 0.363$, $P = 0.036$). Number of peaks per dataset was also strongly correlated with the number of mapped reads ($r = 0.656$, $P < 0.001$) (**Supplementary Figure 1**).

TABLE 2 | Summary of peaks.

Mark	Average number of peaks	Average mapped reads	Average% of genome	Average size of peaks
H3K4Me3	419,386	132,285,209	7	560
H3K4Me1	513,830	136,545,806	21	1,163
H3K27ac	493,905	154,628,688	10	583
H3K27Me3	555,459	127,247,856	21	1,049
CTCF	456,881	90,288,075	7	410

For each mark, the average number of peaks, mapped reads, percentage of genome under peaks, and average size of peaks are reported.

Profile Plots and Replicate Comparisons

Profile plots display the normalised ChIP-seq signal above input signal. As expected H3K4Me3 displayed the highest signal around the TSS, followed by H3K27ac. The profiles for H3K4Me3, H3K4Me1, and H3K27ac displayed a slight bimodal shape at the TSS (**Figure 2**). When comparing the profiles around the TSS of 475 active genes and 8,398 inactive genes, H3K27ac samples showed high signal near the TSS of active genes while H3K27Me3 samples showed lower signal (**Supplementary Figure 2**).

We compared replicates within marks using Pearson correlations. In general, all samples were strongly correlated regardless of tissue or animal, however, there were some batch effects (**Supplementary Figure 3**).

Differential Binding Between Tissues

Normalised ChIP-seq counts at consensus peaks in each tissue were compared for differential binding. Peaks were defined as differentially bound where the mean of the counts in one tissue was significantly different ($P < 0.05$) and two-fold higher or lower than the mean of the counts in all other tissues.

The largest number of differentially bound peaks were in H3K27ac where almost 24% of peaks were different between tissues (**Table 3**). Of the H3K27ac DB peaks, a large proportion were higher in heart (**Supplementary Figure 4**).

Correlation Between Peak Height and Gene Expression Level

We examined whether variation in peak height correlated with variation in gene expression. Normalised ChIP-seq counts at consensus peaks in each mark were compared to normalised gene counts from RNA-seq in the same sample. Every peak within 100 kb of a gene was tested for an association between peak height and gene expression. This resulted in more than 1 million tests for each mark. It was not expected that there would be a correlation between all these peak-gene pairs, however, it was unknown which peaks were interacting with which genes, so it was necessary to test them all. Although the sample size was low (16 or 18 depending on the mark), there was a significant ($P < 0.05$) correlation in 6–11% of peak-gene pairs, which is more than the 5% expected by chance. Similarly, most correlations were negative for H3K27Me3 and positive for all other marks which would not be expected if the correlations were random chance (**Figure 3**).

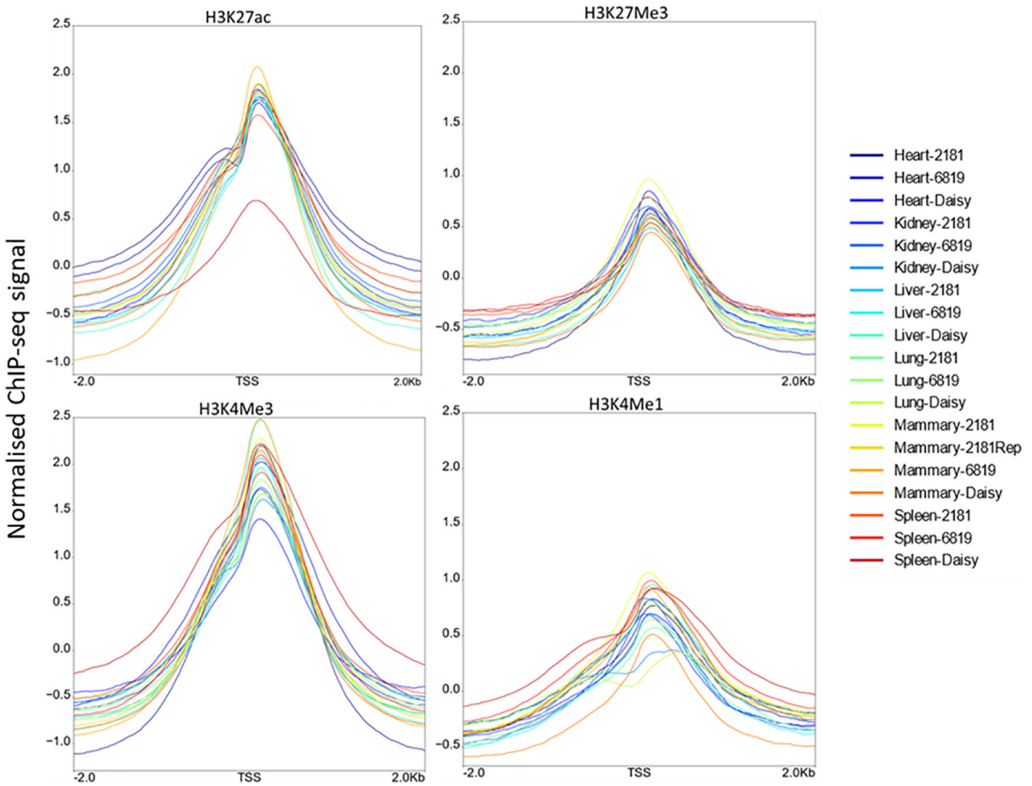


FIGURE 2 | Distribution of ChIP-seq reads around the transcription start site (TSS). For each histone modification these plots show the normalised ChIP-seq signal above input signal within 2 kb of the TSS.

TABLE 3 | Differentially bound peaks.

Mark	Peaks tested	Peaks DB	Percentage peaks DB
H3K27ac	885,919	213,293	24.1
H3K4Me3	853,788	30,965	3.6
H3K4Me1	649,485	21,230	3.3
H3K27Me3	740,188	17,347	2.3
CTCF	961,881	93,473	9.7

Number of peaks tested as well as number and percentage of peaks differentially bound (DB) between tissues ($P < 0.05$, fold change > 2).

Annotation With ChromHMM

Using the four histone modifications and CTCF seven chromatin states were defined across the genome (**Table 4**) using ChromHMM (47). When higher numbers of states were tested ChromHMM defined states with similar functional mark profiles at similar probabilities. To annotate the seven states, we looked for enrichment within these states within active and inactive genes, promoters and proximal regions and used evidence from other studies of what these marks represent. Between 1,169 and 1,826 genes were defined as active in any one tissue and between 10,653 and 11,815 defined as inactive. There were between six and 134 genes which were only active in one tissue (**Supplementary Table 4**).

The seven states were annotated as inactive promoters, repressed, no signal, hyperChIPable region, active enhancer, permissive, and active promoter (**Table 4**). State 2 was annotated as "repressed" as it had a high chance of observing the inactive mark H3K27Me3 and was enriched in inactive genes (**Figure 4**). State 1 was defined as "inactive promoter" because it was slightly more enriched upstream of inactive genes (except for in liver) consistent with inactive promoters and had a high probability of observing H3K27Me3 but also H3K4Me1, H3K4Me3, and CTCF. Two states (3 and 6) displayed low probability of any mark. State 3 was defined as "no signal" as it covered most of the genome but was not highly enriched in any annotated regions suggesting regions which were not functional at that time in these tissues. State 6 was annotated as "permissive" as it was highly enriched in active genes and regions proximal to active genes indicating a more open, permissive state. State 5 was termed "active enhancer" as it had a high probability of H3K27ac and H3K4Me1 (**Table 4**) which is a combination thought to denote active enhancers (26). This state was slightly enriched in active genes in all tissues (**Figure 4**). State 7 had a high probability of observing H3K4Me1, H3K4Me3, CTCF, and the activating mark H3K27ac (**Table 4**) and was defined as "active promoter." This state was enriched at all regions 2 kb upstream of the TSS but was particularly enriched upstream of active TSS (**Figure 4**). State 4 had a high probability of observing all 5 marks at once and was enriched in both active and inactive regions but did not show a consistent pattern in

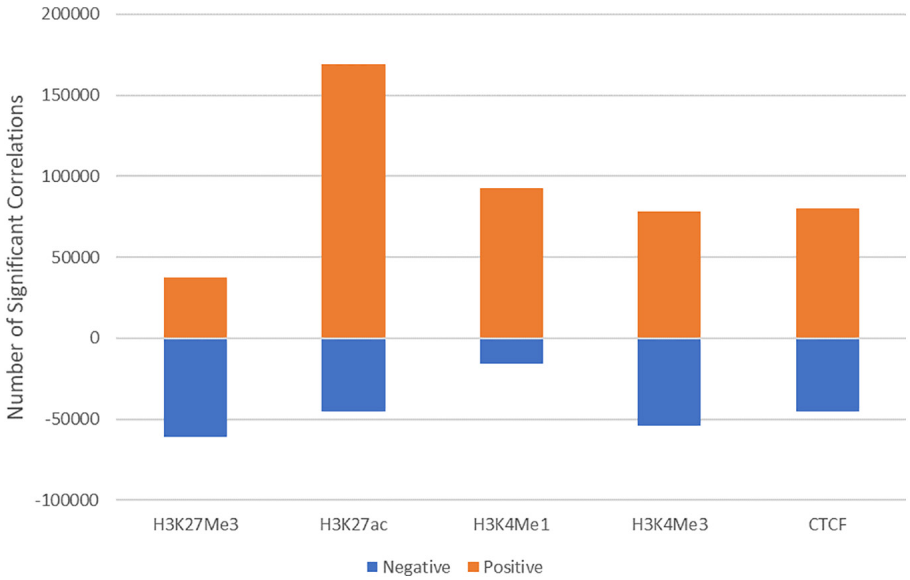


FIGURE 3 | Direction of significant correlations. The direction of correlations between ChIP-seq and RNA-seq counts for H3K27Me3, H3K27ac, H3K4Me1, H3K4Me3, and CTCF.

TABLE 4 | Emission probabilities for seven states from ChromHMM.

State	H3K27ac	H3K4Me1	CTCF	H3K4Me3	H3K27Me3	
1	0.57	0.94	0.94	0.90	0.98	Inactive promoter
2	0.04	0.29	0.43	0.22	0.67	Repressed
3	0.00	0.00	0.00	0.00	0.00	No signal
4	0.99	0.99	0.94	0.92	0.98	HyperChIPable region
5	0.72	0.85	0.45	0.16	0.34	Active enhancer
6	0.03	0.04	0.01	0.00	0.00	Permissive
7	0.99	0.90	0.69	0.69	0.22	Active promoter

The emission probabilities for each mark in each state provided by ChromHMM. Darker colour indicates higher probability. Annotations (last column) for each state was based on location in the genome and known associations from literature.

any of the regions tested. We defined State 4 as "hyperChIPable regions" as coined in a recent study (Massa et al., 2021).

Analysis of tissue specific genes was hampered by the small numbers involved (**Supplementary Table 4**); Kidney, lung and spleen only had 14, 11, and six active genes identified as tissue specific. However, in mammary and liver (26 and 134 tissue specific genes, respectively) active tissue specific genes, promoters, and proximal regions were enriched in "active promoters" similar to other active genes, promoters, and proximal regions (**Supplementary Table 5**). However, in heart, tissue specific active genes, promoters and proximal regions were more enriched for "hyperChIPable regions" than "active promoters" which was not consistent with the other active genes, promoters, and proximal regions.

Inspection of ChromHMM states near known genes showed clear delineation between active and inactive genes in appropriate tissues (e.g., **Figure 5**). Alpha S1 casein (*CSN1*) is a gene known to be highly expressed in the mammary gland (Ibeagha-Awemu et al., 2016). The gene expression data showed that it is highly expressed in mammary tissue but not in the other tissues studied here. **Figure 5** shows mammary tissue is marked by hyperChIpable (yellow), active enhancer (red) and permissive (dark green) states, all active states, while the remaining tissues are marked by inactive promoter (blue), repressed (purple), and no signal (light green) states, all inactive states. State 7 is not present, which is meant to be an active promoter state.

Enrichment of Putative Causal SNPs in Functional Regions

Putative functional regions described above were tested for enrichment of potential causative SNPs from 11 datasets that included expression QTL, milk production trait QTL and sites conserved across numerous species (see materials and methods **Table 1** for descriptions of the datasets).

Enrichment in Peaks

All putative causal variants were significantly overrepresented in peaks (enrichment > 1) with average enrichment between 1.17 (for SNP80k) and 3.82 (for QTL Protein Yield) for each SNP dataset (**Figure 6** and **Supplementary Table 6**). The narrow peaks (H3K4Me3, H3K27ac, and CTCF) had the highest enrichment across all SNP datasets although the differences were small. The QTL for protein yield had the highest enrichment across all marks.

There was little variation in enrichment between tissues, except in the milk production QTL results where H3K4Me3, H3K27ac, and CTCF peaks in mammary gland consistently had higher enrichment (e.g., **Figure 6**, **Supplementary Table 6**). All enrichment was highly significant with $P < 0.001$ (**Supplementary Table 6**).

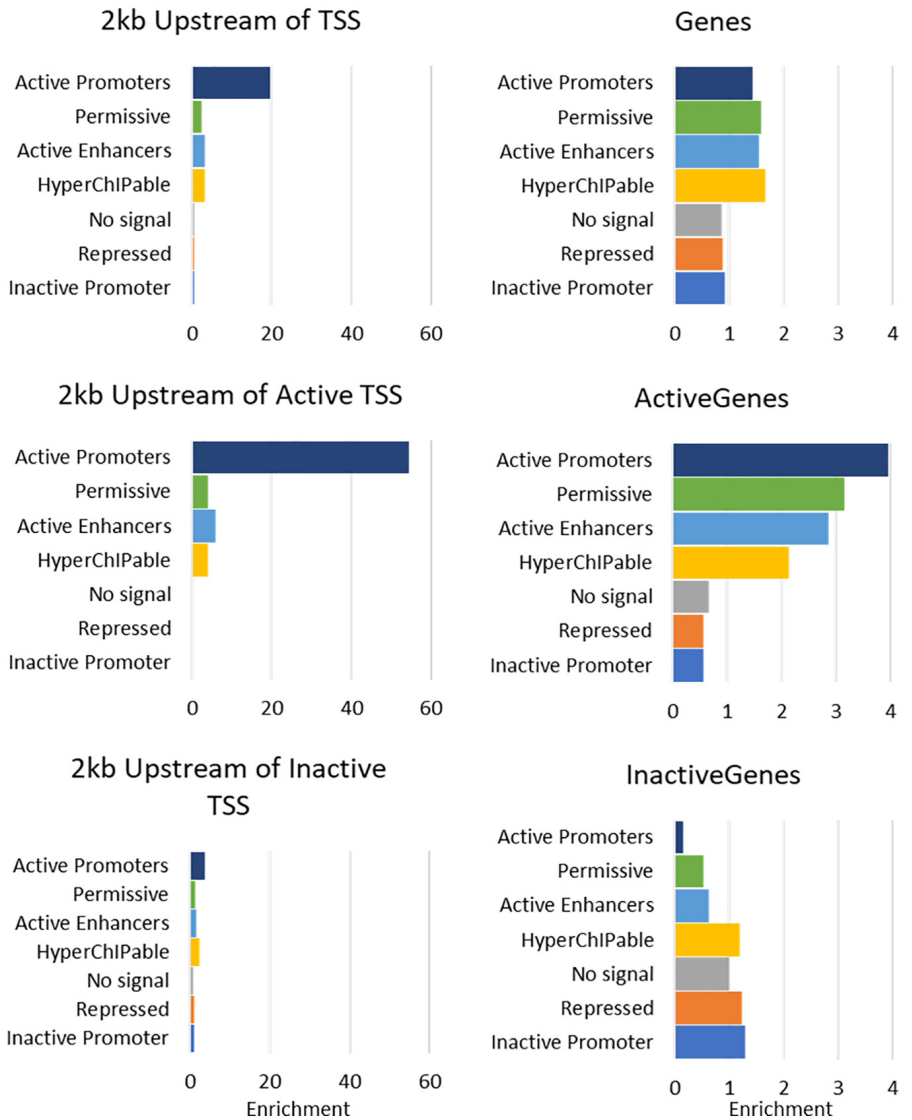


FIGURE 4 | Enrichment of 7 Chromatin states in Mammary Gland at annotated regions of the bovine genome. For more details of enrichment at annotated regions see **Supplementary Figure 5**.

Enrichment in Differentially Bound Peaks

Some differentially bound peaks were enriched for causal variants and some were depleted (**Supplementary Table 7**). This was not consistent across marks or across tissues. This likely reflects the small numbers of peaks which were differentially bound in some cases (**Table 3**) which increased the noise in the data.

Enrichment in Peaks Which Correlate With Gene Expression

Peaks which correlate with gene expression may be affecting gene expression and so are strong candidate regions for causal variants. Filtering peaks for those correlated with gene expression (**Supplementary Table 8**) improved enrichment for all SNP sets except SNP 80 k and conserved regions (**Figure 7** and **Supplementary Table 9**). Enrichment was significant in all cases with $P < 0.001$.

Enrichment in ChromHMM States

All the states defined by ChromHMM except state 3 ("no signal") were enriched for some causal SNP datasets (**Figure 8**). The states with the highest enrichment were State 4 ("hyperChIPable region") and State 7 ("active promoters") except for the 80 k SNPs which were most enriched in State 1 ("inactive promoters"). The highest enrichment was for QTL for protein yield in State 4 ("hyperChIPable region"), these QTL also had the lowest enrichment in State 3 ("no signal"), where they were strongly depleted. Most of the SNP datasets showed highest enrichment in states 4 and 7, moderate enrichment in states 1,2,5, and 6 and depletion in state 3. However, the 80 k SNP dataset showed low

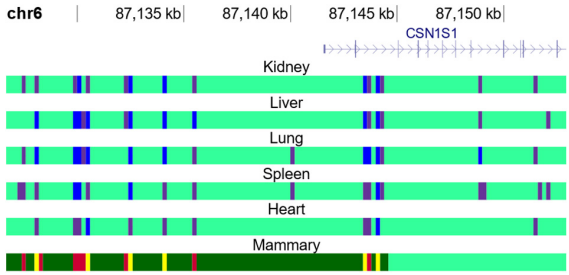


FIGURE 5 | ChromHMM states vary between tissues. Representation of ChromHMM states in heart, kidney, liver, lung, spleen, and mammary gland tissue near the transcription start site of CSN1, a gene highly expressed in the mammary gland. "Inactive promoter" (blue), "Repressed" (purple) and "No Signal" (light green) states are all inactive states, while "Active Enhancer" (red), "HyperChIPable Region" (yellow), and "Permissive" (dark green) are all active states.

to no enrichment in all states except State 1 and conserved SNPs showed low to no enrichment in all states except 4 and 7. The highest enrichment in State 4 was only slighter better than the highest enrichment when considering peak regions defined by the histone mark and was worse than when using peaks correlated with gene expression. Depletion and enrichment were statistically significant in most cases (**Supplementary Table 10**).

Consistent with results in peaks there was little variation in enrichment between tissues in most of the SNP datasets except for conserved regions and milk production QTL (**Supplementary Table 10**). Conserved regions were largely similar between tissues except for heart which had higher enrichment in State 1 than the other tissues. State 1 in heart was also consistently more enriched in all five milk production QTL SNP datasets than the other tissues. State 7 in liver and in some cases mammary displayed little to no enrichment for the milk production QTL while the other tissues displayed high enrichment for the milk production QTL in this state.

Confounding Factors

Enrichment Based on Location in the Genome

One confounding factor of the enrichment analysis is that potential causal SNP may be found within peaks because both peaks and causal SNP may tend to be found close to genes. We

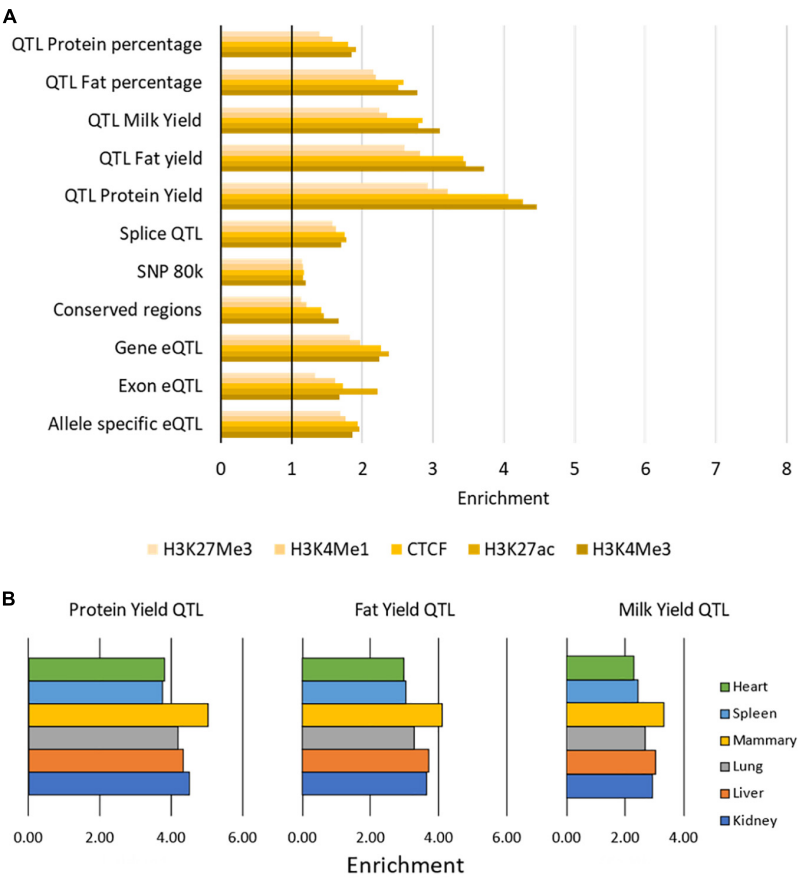


FIGURE 6 | Enrichment of causal variants within functional regions. **(A)** Enrichment of causal variants within peaks. Enrichment of each SNP dataset within each histone modification or CTCF averaged across tissues. All values above 1 (indicated by the vertical black bar) are enriched for causal variants. Enrichment was significant with $P < 0.001$ for all tests. **(B)** Enrichment of three sets of milk trait QTL within H3K27ac peaks. Peaks in mammary gland have the highest enrichment for these milk trait QTL.

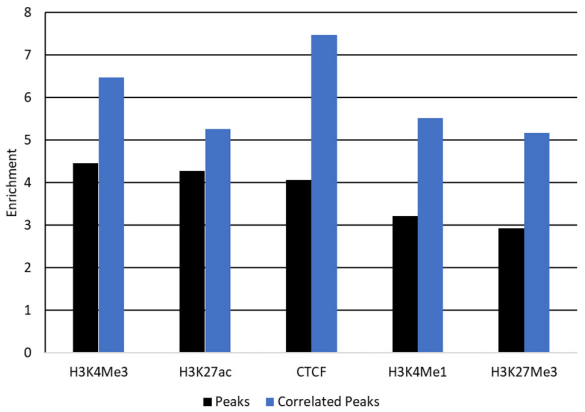


FIGURE 7 | Enrichment of QTL for protein yield in peaks correlated with gene expression. Enrichment of QTL for protein yield within peaks (black) and peaks correlated with gene expression (blue) for each mark. Enrichment was significant with $P < 0.001$ for all tests.

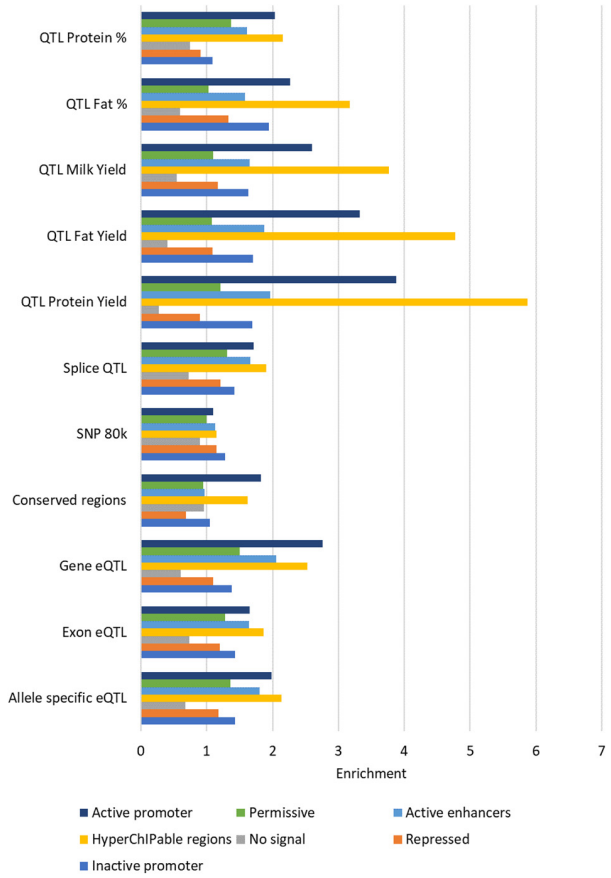


FIGURE 8 | Average enrichment of chromatin states for each SNP dataset. Enrichment for each state averaged across tissues.

tested whether enrichment of potential causal SNPs in peaks changed depending on the distance of the peaks from the nearest gene and whether this explained the high enrichment seen within peaks. Peaks and SNPs were split up into 10 groups of 100 kb increments (**Supplementary Tables 11**, **12**) depending on their distance to the nearest TSS and enrichment was tested within these groups.

As distance from TSS increased the number of SNPs and peaks in these regions decreased (**Supplementary Tables 11**, **12**). In general enrichment of most of the SNP datasets within peaks remained constant regardless of location (**Figure 9**). However, conserved SNPs were not enriched in any tissues or marks more than 100 kb from TSS and Exon eQTL and Splice QTL were not enriched further than 800 kb from TSS. Allele specific eQTL remained enriched within peaks in all regions but the strongest enrichment was between 0 and 100 kb from TSS (**Figure 9**). Milk production QTL SNPs were not found more than 300 kb from a TSS (**Supplementary Table 11**) so enrichment could not be assessed beyond this, but peaks were enriched for SNPs up to this point (**Figure 9**).

Filtering SNPs

Another confounding factor of this analysis is that high enrichment might be driven by a group of SNPs all in high LD with each other. To avoid this possibility significant milk production QTL SNPs ($P < 1 \times 10^{-7}$) were filtered to account for linkage disequilibrium (LD pruning) where SNPs were retained with $r^2 < 0.5$ (**Supplementary Table 13**). Pruning improved enrichment in peaks but not in peaks whose height was correlated with gene expression (**Figure 10**). Pruning consistently improved enrichment in State 1 and 4 but had mixed effects in other states.

DISCUSSION

This study presents the results of ChIP-seq for 4 histone modifications and one transcription factor in six tissues from 3 lactating dairy cows. The ChIP-seq data was used to annotate functional regions in the bovine genome and establish whether functional regions are enriched for causal variants.

To confirm the quality of the ChIP-seq data we looked at the intensity of ChIP signal for each histone modification within 2 kb of the TSS (**Figure 2**). All 4 histone modifications had an increase at the TSS. H3K4Me3 had the highest intensity consistent with expectations (Guenther et al., 2007) followed by H3K27ac. Some studies show a bimodal distribution of H3K4Me1 and/or H3K4Me3 signal at promoters (Barski et al., 2007; Kingsley et al., 2020) but this is not always observed (Hoffman et al., 2010; Bae and Lesch, 2020). In this study only a very slight bimodal distribution in the profile of H3K4Me3, H3K4Me1, and H3K27ac was observed. We also plotted the profile of the active and inactive marks H3K27ac and H3K27Me3 at active and inactive genes (**Supplementary Figure 2**). H3K27ac had much higher signal than H3K27Me3 at active genes as expected (Barski et al., 2007) but there was only a small difference between active and inactive gene profiles in H3K27Me3. This modest difference has also been found in other studies (Barski et al., 2007) with Mikkelsen et al. (2007) finding that not all repressed promoters are marked by H3K27Me3.

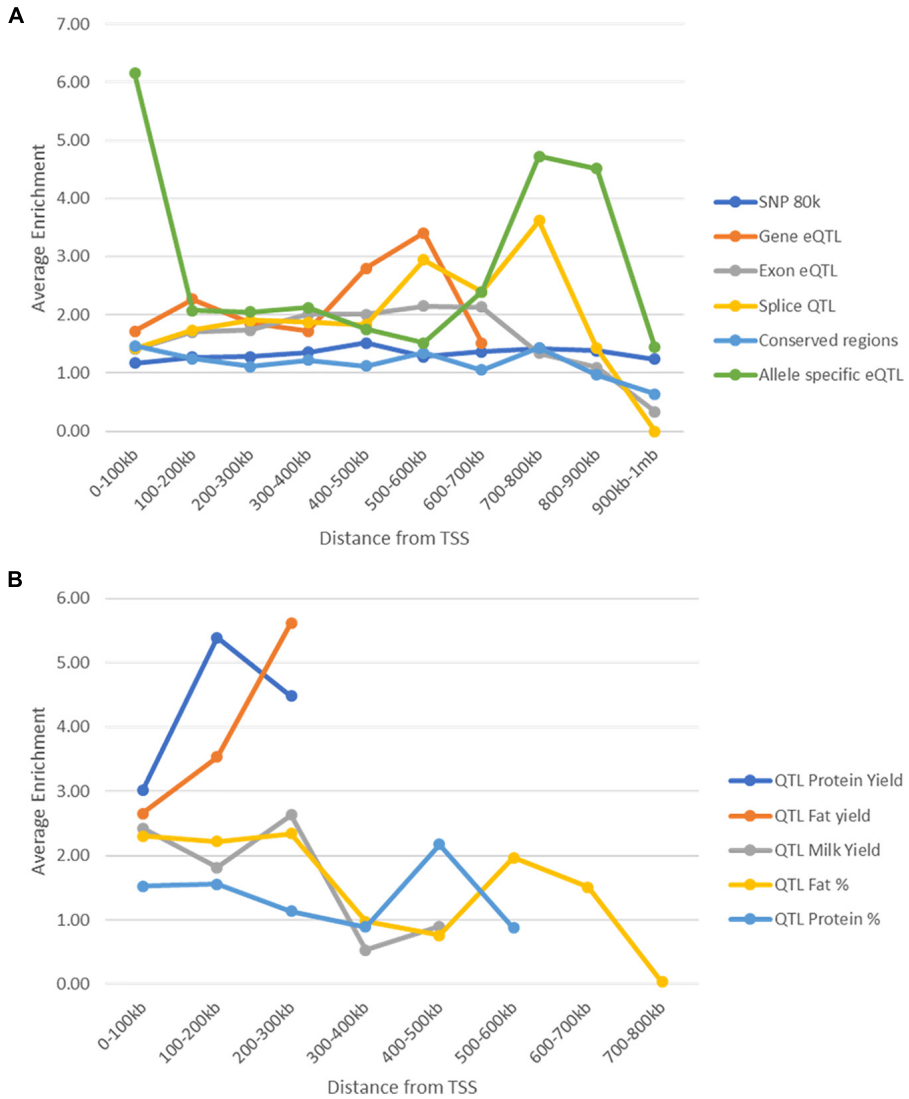


FIGURE 9 | Enrichment of SNPs within peaks by distance to the nearest transcription start site. **(A)** Enrichment (averaged across all animals, tissues, and marks) of SNPs within peaks at different distances from transcription start sites. Enrichment stays above 1 in all cases until 800 kb from the nearest transcription start site. **(B)** Enrichment (averaged over all tissues and animals) of milk production QTL within peaks at different distances from transcription start site. Enrichment was not calculated for protein and fat yield more than 300 kb from a TSS as there were less than 10 SNPs.

For each functional mark, peaks were called from ChIP DNA sequence using MACS2. Up to 500,000 peaks were found for each mark representing millions of functional regions in the genome of dairy cows. Differentially bound peaks were also annotated. Only very low numbers of peaks were different between tissues in most of the marks except for H3K27ac (**Table 3**). This is consistent with high correlation among all samples as observed in **Supplementary Figure 3**. It is not surprising that H3K27ac differed between tissues more than other marks because it is thought to be associated with active promoters and enhancers (Barski et al., 2007). However, work in other species suggests that the enhancer associated mark H3K4Me1 should also display some tissue specificity while marks such as H3K4Me3 are more uniform (Heintzman et al., 2009; Shen et al., 2012). This study did not find this. There were notable differences in the number of differentially bound peaks between tissues. For example, 24% of the H3K27ac regions tested were different in at least one tissue but more than half of these differentiated regions were specific to heart (**Supplementary Figure 4**). In addition, CTCF in mammary and liver had extremely high numbers of down-regulated peaks and very few up-regulated peaks, while kidney, lung and spleen had the opposite. The addition of more samples in future will likely improve this result.

To annotate peaks which are correlated with gene expression, all peaks within 100 kb of a transcription start site were tested for a correlation between read counts under the peak and in the gene. As expected, most peak-gene pairs were not correlated, however, there were more significant correlations than would

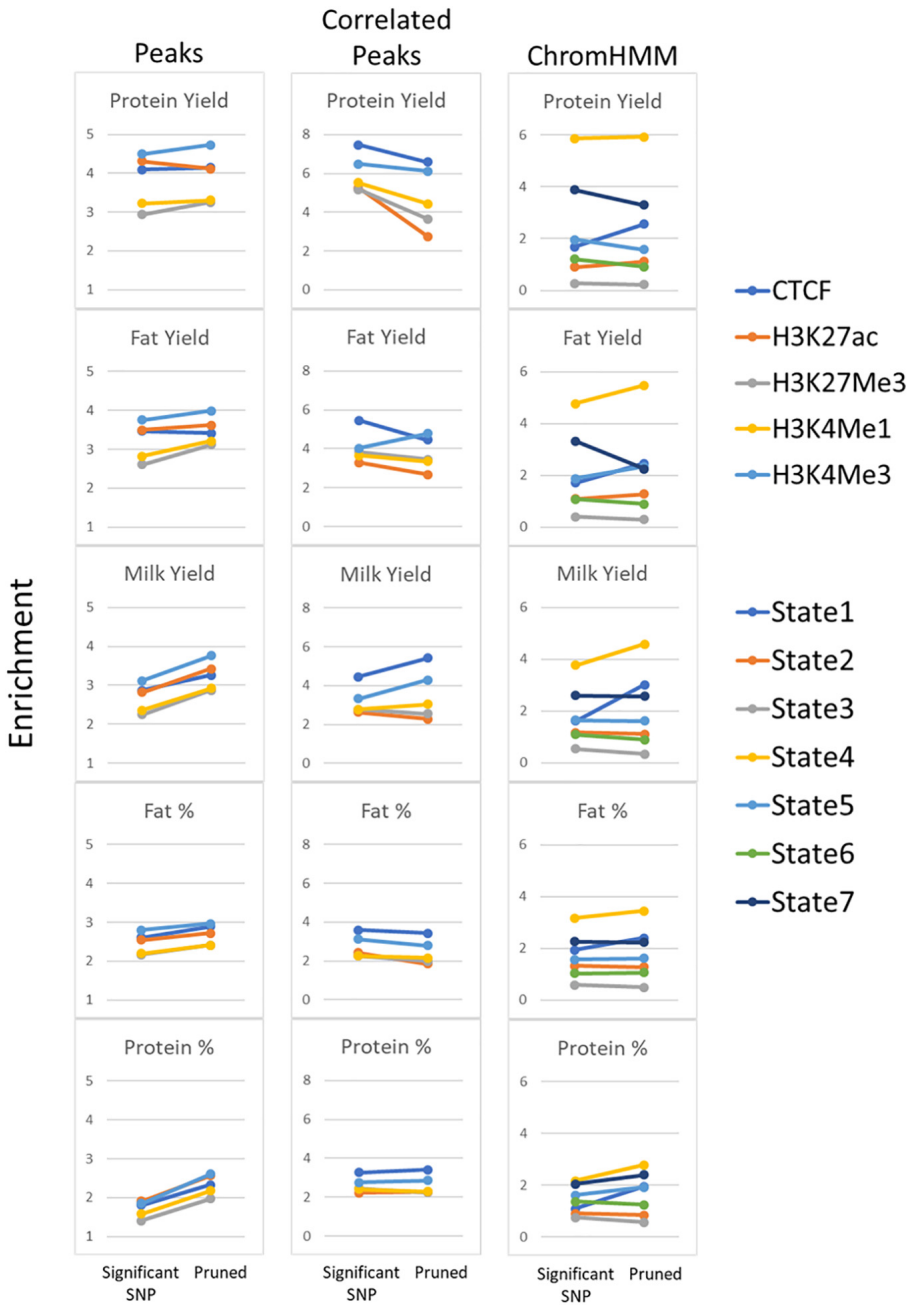


FIGURE 10 | Enrichment of pruned SNPs within functional regions. Comparison of enrichment for all significant milk production QTL ($P < 1 \times 10^{-7}$) compared with pruning SNPs at $r^2 < 0.5$. All peaks and peaks correlated with gene expression (Correlated Peaks) for all 5 traits are shown as well as ChromHMM states.

be expected by chance. For most marks, a high proportion of correlations were positive which would also not occur by chance. The highest number of correlated peaks were from histone modification H3K27ac (**Figure 3**), this is logical as this mark has been found to correlate with active regions of the genome (Wang et al., 2008; Cotney et al., 2012). Similarly, the majority of H3K27Me3 correlations were negative (**Figure 3**), which is also expected as this mark has been found to represent repression of transcriptional activity (Barski et al., 2007; Cotney et al., 2012). These peaks are important functional regions to annotate as they indicate a potential functional link between significantly correlated peaks and nearby gene expression.

ChromHMM (Ernst and Kellis, 2012) was used to combine data from multiple marks and call seven chromatin states across the genome. RNA-seq data from the same tissues were used to annotate these states. The results show a demarcation between active and inactive states consistent across multiple tissues. For example, ChromHMM was able to differentiate State 3

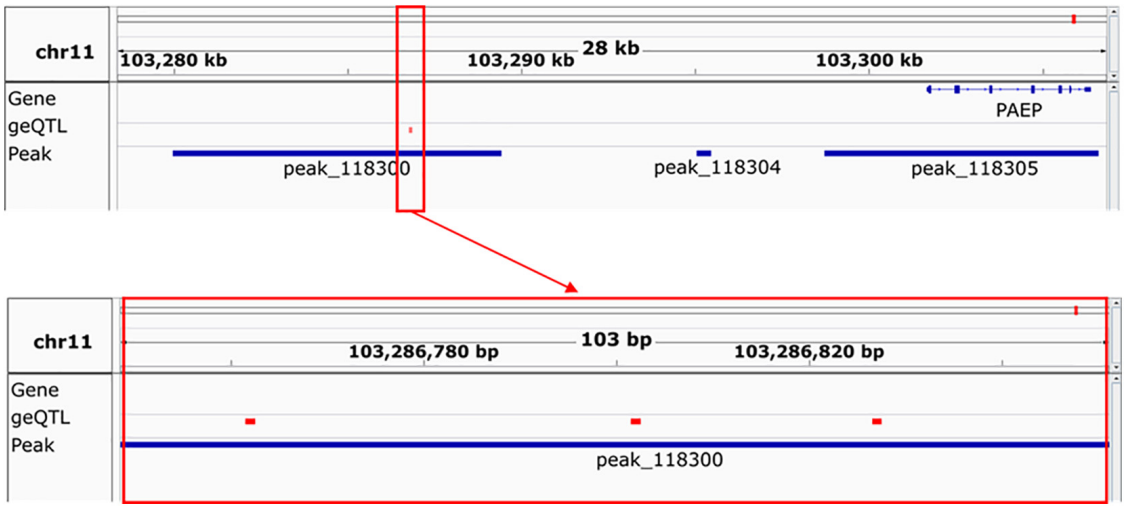


FIGURE 11 | An example of the utility of this data. Peak 118,300, whose location is shown in the upper panel, is significantly 2-fold higher in mammary gland than other tissues and its height (not shown) correlates (r = 0.71, P = 0.002) with the expression of the gene PAEP. Three geQTL, shown in the red box in the upper and lower panels, which correlate with the expression of PAEP in milk cells, can be found in this peak.

("no signal"), which covered most of the genome, from State 6 ("permissive"), which was highly enriched at active genes, even though they had very similar histone modification and transcription factor occupancy. However, this was not always consistent. State 7 ("active promoters") was highly enriched at active promoters but there was also some enrichment at inactive promoters. Similarly State 2 ("repressive") was depleted in active regions but was only slightly enriched in inactive regions. Lastly, State 1 which we called "Inactive promoters" was only slightly enriched upstream of inactive genes. Genes were defined as inactive when there was no RNA-seq counts from that gene which is very stringent, so it is possible some of these genes were poised or were not accurately annotated.

The remaining states were annotated based on information from other studies. State 5 was labelled an "active enhancer" because the marks present in this state reflect enhancer conditions in other species (Heintzman et al., 2007; Creyghton et al., 2010), however, we were unable to provide any genomic evidence for this as enhancers are poorly annotated in the bovine genome. State 4 ("hyperChIPable regions") had all five marks occurring in the same place and was found to varying degrees in all genomic regions looked at. This phenomenon was described in a recent study and was termed a "hyperChIPable" region (Massa et al., 2021). However, this is not consistent with most other published literature which suggests H3K27ac and H3K27Me3, active and inactive marks, respectively, should not occur at the same place in the genome (Tie et al., 2009). We hypothesise this could occur for multiple, not mutually exclusive reasons. (1) These regions are "poised" between active and inactive states in the mass of heterogeneous cells in the tissue. (2) One or more of the peaks are false positives found when the antibody binds off-target (Jain et al., 2015). (3) High sequence depth is picking up weak peaks which are not found at "normal" read depths (20 million reads are recommended by ENCODE (Landt et al., 2012), our data is between 20 and 300 million reads). (4) These regions represent an important functional region. It seems likely that all 4 options are contributing to our observations including the fact that some of these peaks are false positives. Despite this we found these regions were highly enriched for putative causal SNPs which suggests some function.

It is hypothesised that causal variants are found in functional regions of the genome. To test whether this is true for this study we tested multiple potential causal variant datasets for enrichment in the functional regions described in this study. The enrichment of ChIP-seq peaks for potential causal variants in all datasets was significantly higher than random, although the degree varied across datasets, tissues, and marks. ChIP-seq peaks in mammary gland were particularly enriched for SNPs identified in GWAS for milk production traits. This is consistent with other studies showing trait-associated variants are enriched in peaks from tissues associated with the trait (Ernst et al., 2011; Kundaje et al., 2015). The ChromHMM state with a high probability of all five marks occurring together (State 4-" hyperChIPable regions") was also enriched for putative causal SNPs while the state with no signal from any functional marks was depleted for all SNP datasets tested. The highest enrichment was observed in peaks which correlated with gene expression. Except for SNPs from conserved regions and the 80 k SNPs, all SNP datasets were highly enriched in these peaks. A potential causal mechanism for these variants would be that they affect the functional mark binding which in turn affects gene expression making these good candidate SNPs for further investigation.

A limitation of the enrichment study is that none of the SNPs tested were confirmed causal variants. Most were SNPs which associate with a phenotype so may be in linkage disequilibrium with a causal variant but are not the causal variant themselves. This means that the enrichment of causal variants (the proportion of which will be different in different datasets)

will be diluted by the non-causal variants in the dataset, but it could also mean that our functional regions were just enriched for SNPs in linkage disequilibrium with causal variants. However, the data shows that across multiple SNP datasets using different methodologies there was consistent enrichment for these SNPs within functional regions. This is consistent with studies in human and cattle (Maurano et al., 2012; Schaub et al., 2012; Trynka et al., 2013; Ma et al., 2015; Wang et al., 2017; Fang et al., 2019).

Another possible limitation is that the peaks were enriched because they and the SNPs tested were both near genes. This would mean they are more likely to intersect because of similar distributions in the genome rather than due to a causal variant affecting functional mark binding (Cano-Gamez and Trynka, 2020). To test this, we split the genome into regions based on how far each was from the nearest TSS and tested enrichment with peaks and SNPs just within these regions. In most cases SNPs were enriched in ChIP-seq peaks regardless of the distance of the region being tested from a TSS. This suggests that the enrichment observed was not just a function of proximity to genes. Unfortunately, this could only be tested up to 300 kb in the milk trait GWAS dataset as there were too few SNPs more than 300 kb from a TSS. Due to the small number of SNPs in the GWAS datasets and the fact that they mostly cluster together we were concerned that multiple SNPs from few locations were enriched in few peaks and these SNPs were all tagging one causal variant. To account for this, SNPs were filtered for LD and only the most significant was included from each LD region. This reduced the number of SNPs dramatically but in most cases filtering in this way either improved enrichment or did not change it suggesting successful pruning of SNPs which were in LD with the one causal variant and peaks were still enriched for these variants.

To highlight the utility of the data generated, an example of three SNPs which we hypothesise are good candidate causal variants for important milk traits is shown (**Figure 11**). In a gene expression QTL study in milk cells, Xiang et al. (2020) found 531 SNPs which significantly associate with expression of the progestagen-associated endometrial protein gene *PAEP* (Ibeagha-Awemu et al., 2016). In this study, we filtered these 531 SNPs down to three (rs208362116, rs210272536, and rs136737193) that were found in a H3K27ac peak which was higher in mammary gland than the other five tissues. The height of this peak also correlated with *PAEP* gene expression ($r = 0.71$, $P = 0.002$). A genetic variant in this peak which altered its height therefore may also alter *PAEP* gene expression. Further study would be needed to verify this, but ChIP-seq data enabled us to filter 531 SNPs down to three genetic variants with a potential causal mechanism for altering gene expression.

Genomic selection will be more accurate over time and across breeds if causal variants are included in predictive modelling (MacLeod et al., 2016). It is thought that causal variants are found in functional regions of the genome but until recently (Villar et al., 2015; Fang et al., 2019; Kern et al., 2021) these were not well annotated in cattle (Koufariotis et al., 2014; Nguyen et al., 2018). This study annotated functional regions in six tissues in 2–3 Holstein dairy cows using ChIP-seq for four histone modifications and one transcription factor. This is the first time this has been done in the mammary gland of a lactating dairy cow. Although many histone modifications overlapped between tissues, some regions showed a difference in binding across tissues and some peaks were correlated with differences in gene expression. Lastly, we confirmed that putative causal variants were enriched in the functional regions discovered. This confirms that future work should consider using these regions when selecting SNPs for genomic selection (MacLeod et al., 2016; Xiang et al., 2019).

DATA AVAILABILITY STATEMENT

The sequence data for this study is publicly available in the European Nucleotide Archive at http://www.ebi.ac.uk/ena/data/view/PRJEB41939.

ETHICS STATEMENT

The animal study was reviewed and approved by the Department of Jobs, Precincts, and Regions Ethics Committee (Application No. 2014-23).

AUTHOR CONTRIBUTIONS

CP-W contributed to conception and design, acquisition of data, data analysis, and drafting of manuscript. JW contributed to acquisition of data. RX contributed to data analysis and critical revision. JG contributed to data acquisition. MG contributed to conception and design, data analysis, and critical revision. AC contributed to conception and design, data acquisition, data analysis, and critical revision. All authors contributed to the article and approved the submitted version.

FUNDING

This project was funded by the DairyBio (a joint venture project between Agriculture Victoria and Dairy Australia). Australian Research Council's Discovery Projects (DP160101056) supported RX and MG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the Agriculture Victoria staff at Ellinbank and Bundoora research institutes who helped with sample collection. The authors also thank DataGene for access to data used in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.664379/full#supplementary-material

Supplementary Figure 1 | Pearson correlations of mapped reads with other parameters. **(A)**. The number of peaks found significantly increases as mapped reads increases ($r = 0.656$, $P < 0.001$). **(B)**. The percent of the genome covered by narrow peaks significantly increases as mapped reads increase ($r = 0.715$, $P < 0.001$). **(C)**. The percent of the genome covered by broad peaks significantly increases as mapped reads increase ($r = 0.363$, $P = 0.036$).

Supplementary Figure 2 | Comparison of profiles of H3K27Me3 and H3K27ac at active and inactive genes. Normalised ChIP-seq signal above input signal of the active (H3K27ac) and inactive (H3K27Me3) histone modifications within 2kb of transcription start sites (TSS) for active and inactive genes in all six tissues.

Supplementary Figure 3 | Comparison of replicates in each mark. Pearson correlations of replicates in each mark.

Supplementary Figure 4 | Direction of differential binding. Number of differentially bound peaks with either higher (Up) or lower (Down) binding in heart, liver, kidney, lung, mammary gland (MG), and spleen.

Supplementary Figure 5 | Enrichment of 7 Chromatin states in different tissues at annotated regions of the bovine genome. Where state1 is "inactive promoter," state 2 is "repressed," state 3 is "no signal," state 4 is "hyperChIPable," state 5 is "active enhancer," state 6 is "permissive," and state 7 is "active promoter." Darker blue indicates higher enrichment.

Supplementary Table 1 | Summary of 86 ChIP-seq datasets. The number of biological replicates is shown for each tissue-mark combination.

Supplementary Table 2 | Summary data for each ChIP-seq library. Information on the number of mapped reads, quality and coverage of the genome for each ChIP-seq dataset.

Supplementary Table 3 | Correlation between ChIP-seq and RNA-seq counts. The number of tests for correlation between gene expression and peak height as well as the number and percentage significantly correlated ($P < 0.05$) for each mark. All peaks within 100 kb of a gene were tested.

Supplementary Table 4 | Active and inactive genes used for ChromHMM annotation. The number of genes defined as active, tissue-specific active and inactive based on normalised count data from RNA-seq in siz tissues across three animals.

Supplementary Table 5 | Enrichment of 7 Chromatin states in different tissues at annotated regions of the bovine genome. Where state1 is "inactive promoter," state 2 is "repressed," state 3 is "no signal," state 4 is "hyperChIPable regions," state 5 is "active enhancer," state 6 is "permissive," and state 7 is "active promoter." Darker green indicates higher enrichment.

Supplementary Table 6 | Enrichment of SNP datasets within peaks. The enrichment of each SNP dataset within peaks for each sample. Significance of enrichment or depletion was determined with a hypergeometric test. A summary table is included to the right with average enrichment for each tissue. Darker green indicates higher enrichment.

Supplementary Table 7 | Enrichment of SNP datasets within differentially bound peaks. The enrichment of each SNP dataset within differentially bound peaks for each sample with either higher (Up) or lower (Down) binding. Significance of enrichment or depletion was determined with a hypergeometric test. A summary table is included to the right with average enrichment for each tissue. Darker green indicates higher enrichment.

Supplementary Table 8 | Number of peaks correlated with gene expression. Number of peaks for each mark whose height correlated with gene expression.

Supplementary Table 9 | Enrichment of putative causal variants of peaks and peaks correlated with gene expression. Enrichment of each SNP dataset within peaks and peaks correlated with gene expression. Filtering for correlation with gene expression increased enrichment for causal variants. Enrichment was significant with $P < 0.001$ for all tests.

Supplementary Table 10 | Enrichment of SNP datasets within ChromHMM states. The enrichment of each SNP dataset within each state in each tissue. Significance of enrichment or depletion was determined with a hypergeometric test. A summary table is included to the right with average enrichment for each tissue.

Supplementary Table 11 | The number of SNPs in each dataset grouped by distance to nearest transcription start site. For each SNP dataset we grouped SNPs based on distance to nearest transcription start site in bins of 100 kb.

Supplementary Table 12 | The number of peaks grouped by distance to transcription start site, averaged over samples. The average number of peaks, grouped by distance to the nearest transcription start site, tested for enrichment.

Supplementary Table 13 | Number of SNPs in milk production QTL datasets after filtering using r^2.

REFERENCES

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project. *Genome Biol.* 16, 1–6.

Bae, S., and Lesch, B. J. (2020). H3K4me1 distribution predicts transcription state and poising at promoters. *Front. Cell Dev. Biol.* 8:289. doi: 10.3389/fcell.2020.00289

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837. doi: 10.1016/j.cell.2007.05.009

Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120, 169–181. doi: 10.1016/j.cell.2005.01.001

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Cano-Gamez, E., and Trynka, G. (2020). From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.* 11:424. doi: 10.3389/fgene.2020.00424

Chamberlain, A., Hayes, B., Xiang, R., Jagt, C. V., Reich, C., MacLeod, I., et al. (2018). "Identification of regulatory variation in dairy cattle with RNA sequence data," in *Proceedings of the World Congress on Genetics Applied to Livestock Production Molecular Genetics*, Vol. 1 (Palmerston North: Massey University), 254.

Chamberlain, A. J., Vander Jagt, C. J., Hayes, B. J., Khansefid, M., Marett, L. C., Millen, C. A., et al. (2015). Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics* 16:993. doi: 10.1186/s12864-015-2174-0

Cotney, J., Leng, J., Oh, S., DeMare, L. E., Reilly, S. K., Gerstein, M. B., et al. (2012). Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res.* 22, 1069–1080. doi: 10.1101/gr.129817.111

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* 107, 21931–21936. doi: 10.1073/pnas.1016071107

Dorji, J., Vander Jagt, C. J., Garner, J. B., Marett, L. C., Mason, B., Reich, C. M., et al. (2020). Expression of mitochondrial protein genes encoded by nuclear and mitochondrial genomes correlate with energy metabolism in dairy cattle. *BMC Genomics* 21:720. doi: 10.1186/s12864-020-07018-7

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* 28, 817–825. doi: 10.1038/nbt.1662

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–216. doi: 10.1038/nmeth.1906

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49. doi: 10.1038/nature09906

Fang, L., Liu, S., Liu, M., Kang, X., Lin, S., Li, B., et al. (2019). Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biol.* 17:68. doi: 10.1186/s12915-019-0687-8

Goddard, M., Kemper, K., MacLeod, I., Chamberlain, A., and Hayes, B. (2016). Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc. Royal Soc. B* 283:20160569. doi: 10.1098/rspb.2016.0569

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., and Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 130, 77–88. doi: 10.1016/j.cell.2007.05.042

Hayes, B., Chamberlain, A., Daetwyler, H., Vander Jagt, C., and Goddard, M. (2016). Improving genomic selection across breeds and across generations with functional annotation. *J. Anim. Sci.* 94, 3–4. doi: 10.2527/jas2016.94supplement43a

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459, 108–112. doi: 10.1038/nature07829

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318. doi: 10.1038/ng1966

Hoffman, B. G., Robertson, G., Zavaglia, B., Beach, M., Cullum, R., Lee, S., et al. (2010). Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. *Genome Res.* 20, 1037–1051. doi: 10.1101/gr.104356.109

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., et al. (2018). Ensembl variation resources. *Database* 2018:bay119. doi: 10.1093/database/bay119

Ibeagha-Awemu, E. M., Li, R., Ammah, A. A., Dudemaine, P.-L., Bissonnette, N., Benchaar, C., et al. (2016). Transcriptome adaptation of the bovine mammary gland to diets rich in unsaturated fatty acids shows greater impact of linseed oil over safflower oil on gene expression and metabolic pathways. *BMC Genomics* 17:104–104. doi: 10.1186/s12864-016-2423-x

Jain, D., Baldi, S., Zabel, A., Straub, T., and Becker, P. B. (2015). Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Res.* 43, 6959–6968. doi: 10.1093/nar/gkv637

Karlić, R., Chung, H.-R., Lasserre, J., Vlahovièek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 107, 2926–2931. doi: 10.1073/pnas.0909344107

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., et al. (2014). Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 111, 6131–6138.

Kern, C., Wang, Y., Xu, X., Pan, Z., Halstead, M., Chanthavixay, G., et al. (2021). Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nat. Commun.* 12:1821.

Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* 26, 1351–1359. doi: 10.1038/nbt.1508

Kim, S., Yu, N.-K., and Kaang, B.-K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. *Exp. Mol. Med.* 47:e166. doi: 10.1038/emm.2015.33

Kimura, H. (2013). Histone modifications for human epigenome analysis. *J. Hum. Genet.* 58, 439–445. doi: 10.1038/jhg.2013.66

Kingsley, N., Kern, C., Creppe, C., Hales, E. N., Zhou, H., Kalbfleisch, T., et al. (2020). Functionally annotating regulatory elements in the equine genome using histone mark chip-seq. *Genes* 11:3. doi: 10.3390/genes11010003

Koufariotis, L., Chen, Y.-P. P., Bolormaa, S., and Hayes, B. J. (2014). Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. *BMC Genomics* 15:436. doi: 10.1186/1471-2164-15-436

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22, 1813–1831.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [Preprint]. https://arxiv.org/abs/1303.3997v2

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Ma, M., Ru, Y., Chuang, L.-S., Hsu, N.-Y., Shi, L.-S., Hakenberg, J., et al. (2015). Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC Genomics* 16:S3. doi: 10.1186/1471-2164-16-S8-S3

MacLeod, I., Bowman, P., Vander Jagt, C., Haile-Mariam, M., Kemper, K., Chamberlain, A., et al. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* 17:144. doi: 10.1186/s12864-016-2443-6

Massa, A. T., Mousel, M. R., Herndon, M. K., Herndon, D. R., Murdoch, B. M., and White, S. N. (2021). Genome-wide histone modifications and CTCF enrichment predict gene expression in sheep macrophages. *Front. Genet.* 11:612031. doi: 10.3389/fgene.2020.612031

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

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., et al. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science* 342, 747–749. doi: 10.1126/science.1242429

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560. doi: 10.1038/nature06008

Nguyen, Q. H., Tellam, R. L., Naval-Sanchez, M., Porto-Neto, L. R., Barendse, W., Reverter, A., et al. (2018). Mammalian genomic regulatory regions predicted by utilizing human genomics, transcriptomics, and epigenetics data. *GigaScience* 7:gix136.

Ong, C.-T., and Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* 15, 234–246. doi: 10.1038/nrg3663

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680. doi: 10.1038/nrg2641

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., and Bejerano, G. (2013). Enhancers: five essential questions. *Nat. Rev. Genet.* 14, 288–295. doi: 10.1038/nrg3458

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., et al. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44, W160–W165.

Raymond, B., Yengo, L., Costilla, R., Schrooten, C., Bouwman, A. C., Hayes, B. J., et al. (2020). Using prior information from humans to prioritize genes and gene-associated variants for complex traits in livestock. *PLoS Genet.* 16:e1008780. doi: 10.1371/journal.pgen.1008780

Robertson, A. G., Bilenky, M., Tam, A., Zhao, Y., Zeng, T., Thiessen, N., et al. (2008). Genome-wide relationship between histone H3 lysine 4 mono-and tri-methylation and transcription factor binding. *Genome Res.* 18, 1906–1917. doi: 10.1101/gr.078519.108

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Roh, T.-Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. *Proc. Natl. Acad. Sci. U. S. A.* 103, 15782–15787. doi: 10.1073/pnas.0607617103

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Res.* 22, 1748–1759.

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature* 488, 116–120. doi: 10.1038/nature11243

Spicuglia, S., and Vanhille, L. (2012). Chromatin signatures of active enhancers. *Nucleus* 3, 126–131. doi: 10.4161/nucl.19232

Stark, R., and Brown, G. (2011). *"DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data"*. Available online at: http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf

Tie, F., Banerjee, R., Stratton, C. A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., et al. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development* 136, 3131–3141. doi: 10.1242/dev.037127

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., et al. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* 45, 124–130. doi: 10.1038/ng.2504

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., et al. (2015). Enhancer evolution across 20 mammalian species. *Cell* 160, 554–566. doi: 10.1016/j.cell.2015.01.006

Wang, M., Hancock, T. P., MacLeod, I. M., Pryce, J. E., Cocks, B. G., and Hayes, B. J. (2017). Putative enhancer sites in the bovine genome are enriched with variants affecting complex traits. *Genet. Sel. Evol.* 49:56.

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* 40, 897–903. doi: 10.1038/ng.154

Xiang, R., Hayes, B. J., Vander Jagt, C. J., MacLeod, I. M., Khansefid, M., Bowman, P. J., et al. (2018). Genome variants associated with RNA splicing variations in bovine are extensively shared between tissues. *BMC Genomics* 19:521. doi: 10.1186/s12864-018-4902-8

Xiang, R., MacLeod, I. M., Daetwyler, H. D., de Jong, G., O'Connor, E., Schrooten, C., et al. (2021). Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. *Nat. Commun.* 12:860. doi: 10.1038/s41467-021-21001-0

Xiang, R., van den Berg, I., MacLeod, I. M., Daetwyler, H. D., and Goddard, M. E. (2020). Effect direction meta-analysis of GWAS identifies extreme, prevalent and shared pleiotropy in a large mammal. *Commun. Biol.* 3:88.

Xiang, R., Van Den Berg, I., MacLeod, I. M., Hayes, B. J., Prowse-Wilkins, C. P., Wang, M., et al. (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl. Acad. Sci. U. S. A.* 116, 19398–19408. doi: 10.1073/pnas.1904159116

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137.

Zhao, X. D., Han, X., Chew, J. L., Liu, J., Chiu, K. P., Choo, A., et al. (2007). Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* 1, 286–298. doi: 10.1016/j.stem.2007.08.004

Zhou, V. W., Goren, A., and Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* 12, 7–18. doi: 10.1038/nrg2905

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.

Copyright © 2021 Prowse-Wilkins, Wang, Xiang, Garner, Goddard and Chamberlain. 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: 23 June 2021
doi: 10.3389/fgene.2021.689406

Reference Transcriptomes of Porcine Peripheral Immune Cells Created Through Bulk and Single-Cell RNA Sequencing

Juber Herrera-Uribe[1†], Jayne E. Wiarda[2,3,4†], Sathesh K. Sivasankaran[2,5], Lance Daharsh[1], Haibo Liu[1], Kristen A. Byrne[2], Timothy P. L. Smith[6], Joan K. Lunney[7], Crystal L. Loving[2‡] and Christopher K. Tuggle[1*‡]*

[1] Department of Animal Science, Iowa State University, Ames, IA, United States, [2] Food Safety and Enteric Pathogens Research Unit, National Animal Disease Center, Agricultural Research Service, United States Department of Agriculture, Ames, IA, United States, [3] Immunobiology Graduate Program, Iowa State University, Ames, IA, United States, [4] Oak Ridge Institute for Science and Education, Agricultural Research Service Participation Program, Oak Ridge, TN, United States, [5] Genome Informatics Facility, Iowa State University, Ames, IA, United States, [6] USDA, ARS, U.S. Meat Animal Research Center, Clay Center, NE, United States, [7] USDA-ARS, Beltsville Agricultural Research Center, Animal Parasitic Diseases Laboratory, Beltsville, MD, United States

OPEN ACCESS

Edited by:
Angela Cánovas, University of Guelph, Canada

Reviewed by:
Paolo Zambonelli, University of Bologna, Italy
Sergey V. Kozyrev, Uppsala University, Sweden

***Correspondence:**
Crystal L. Loving
crystal.loving@usda.gov
Christopher K. Tuggle
cktuggle@iastate.edu

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

[‡] These authors have contributed equally to this work and share senior and last authorship

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

Received: *31 March 2021*
Accepted: *18 May 2021*
Published: *23 June 2021*

Citation:
Herrera-Uribe J, Wiarda JE, Sivasankaran SK, Daharsh L, Liu H, Byrne KA, Smith TPL, Lunney JK, Loving CL and Tuggle CK (2021) Reference Transcriptomes of Porcine Peripheral Immune Cells Created Through Bulk and Single-Cell RNA Sequencing. Front. Genet. 12:689406. doi: 10.3389/fgene.2021.689406

Pigs are a valuable human biomedical model and an important protein source supporting global food security. The transcriptomes of peripheral blood immune cells in pigs were defined at the bulk cell-type and single cell levels. First, eight cell types were isolated in bulk from peripheral blood mononuclear cells (PBMCs) by cell sorting, representing Myeloid, NK cells and specific populations of T and B-cells. Transcriptomes for each bulk population of cells were generated by RNA-seq with 10,974 expressed genes detected. Pairwise comparisons between cell types revealed specific expression, while enrichment analysis identified 1,885 to 3,591 significantly enriched genes across all 8 cell types. Gene Ontology analysis for the top 25% of significantly enriched genes (SEG) showed high enrichment of biological processes related to the nature of each cell type. Comparison of gene expression indicated highly significant correlations between pig cells and corresponding human PBMC bulk RNA-seq data available in Haemopedia. Second, higher resolution of distinct cell populations was obtained by single-cell RNA-sequencing (scRNA-seq) of PBMC. Seven PBMC samples were partitioned and sequenced that produced 28,810 single cell transcriptomes distributed across 36 clusters and classified into 13 general cell types including plasmacytoid dendritic cells (DC), conventional DCs, monocytes, B-cell, conventional CD4 and CD8 αβ T-cells, NK cells, and γδ T-cells. Signature gene sets from the human Haemopedia data were assessed for relative enrichment in genes expressed in pig cells and integration of pig scRNA-seq with a public human scRNA-seq dataset provided further validation for similarity between human and pig data. The sorted porcine bulk RNAseq dataset informed classification of scRNA-seq PBMC populations; specifically, an integration of the datasets showed that the pig bulk RNAseq data helped define the CD4CD8 double-positive T-cell populations in the scRNA-seq data. Overall, the data provides deep and well-validated transcriptomic data from sorted PBMC populations and the first single-cell

transcriptomic data for porcine PBMCs. This resource will be invaluable for annotation of pig genes controlling immunogenetic traits as part of the porcine Functional Annotation of Animal Genomes (FAANG) project, as well as further study of, and development of new reagents for, porcine immunology.

Keywords: pig, immune cells, transcriptome, single-cell RNA-seq, bulkRNA-seq, FAANG

INTRODUCTION

A major goal of biological research is using genomic information to predict complex phenotypes of individuals or individual cells with specific genotypes. Predicting complex phenotypes is an important component of broad Genome-to-Phenome (G2P) understanding (Koltes et al., 2019), and investing in sequencing of multiple animal genomes, including pigs (*Sus scrofa*), for improved genome and cell functional annotation is key in solving the G2P question (Andersson et al., 2015; Giuffra et al., 2019). In addition to their major role in the world supply of dietary protein, pigs have anatomic, physiologic, and genetic similarities to humans and serve as biomedical models for human disease and regenerative medicine (reviewed in Swindle et al., 2012; Kobayashi et al., 2018). Thus, deep annotation of porcine genome function would be a major milestone for addressing the G2P question. A highly contiguous porcine genome assembly with gene model-level annotation was recently published (Warr et al., 2020). However, this annotation is based primarily on RNA sequencing (RNA-seq) data from solid tissues, with few sample types representative of immune cells, with the exception of alveolar macrophages and dendritic cells (Auray et al., 2016). Given the interaction of animal health and growth, any functional annotation of the porcine genome will be incomplete without deep analysis of expression patterns and regulatory elements controlling the immune system.

The transcriptomes of circulating immune cells serve as a window into porcine immune physiology and traits (Chaussabel et al., 2010; Mach et al., 2013; Schroyen and Tuggle, 2015; Auray et al., 2020). Blood RNA profiling has been used to understand variation in porcine immune responses (Huang et al., 2011; Arceo et al., 2013; Knetter et al., 2015; Munyaka et al., 2019) and genetic control of gene expression (Maroilley et al., 2017). One goal of such research is to develop gene signatures predictive of disease states (Berry et al., 2010) and predict responses to immunizations and/or infections (Chaussabel and Baldwin, 2014; Tsang et al., 2014), as has been demonstrated in humans. Whole blood is easily collected from live animals, but represents an extremely complex mixture of cell types. Estimates of gene expression in mixed samples are inherently inaccurate as cell composition differences are difficult to adjust for, complicating the interpretation of RNA differences across samples and treatments. Thus, starting from whole blood transcriptomic data, it is nearly impossible to link gene expression and regulation to a specific cell or cell type. To determine direct regulatory interactions, we must analyze specific cell populations and even individual cells. A cell type-specific understanding of peripheral immune cell gene expression patterns will thus enhance biological understanding of porcine immunity, reveal targets for phenotyping, and provide a comparison to other species.

Predominant immune cell populations in porcine peripheral blood mononuclear cell (PBMC) preparations are comprised mainly of monocytes, B-cells, and T-cells, with minor fractions of dendritic cells (DCs), natural killer (NK) cells, and NKT-cells also present. Porcine peripheral T-cell populations (reviewed in Gerner et al., 2009, 2015) and DCs (Summerfield et al., 2015; Auray et al., 2016) are readily described based on phenotype, though deeper characterization of porcine immune cells could improve identification of valuable reagent targets and biological understanding of porcine immunity. T-cell populations are commonly grouped as αβ or γδ T-cells according to T-cell receptor (TCR) chain expression and further divided based on CD2, CD4, CD8α, and/or CD8β expression. Pigs have a unique $CD2^-$ γδ T-cell lineage contributing to higher percentages of circulating γδ T-cells (Takamatsu et al., 2006) and unique αβ T-cells expressing both CD4 and CD8α (Zuckermann, 1999). Relatively little is known about different circulating B-cell populations in pigs, as reagents for phenotyping are limited.

Various technical approaches can be used to enrich or isolate specific cell populations, improving resolution of cell types for deeper interrogation of gene expression. Flow cytometry is used to characterize cells based on expression of cell type-specific protein markers, and live cells can be sorted by magnetic- and/or fluorescence-activated cell sorting (MACS/FACS) for use in subsequent assays. MACS/FACS enrichment followed by transcriptomic analysis can provide additional insight of gene expression in specific cell types, but cells expressing the same combination of markers are often still a heterogeneous mixture (Sutermaster and Darling, 2019). Some major subtypes of porcine immune cell populations can be labeled for cell sorting by existing antibody reagents (Gerner et al., 2009), but some subtypes such as B-cells lack these resources.

An exciting alternative to sorting specific cell types for transcriptomic analysis is single-cell RNA-seq (scRNA-seq).

Abbreviations: AUC, area under the curve; ASC, antibody-secreting cell; B, B-cell; bulkRNA-seq, bulk RNA sequencing; cDC, conventional dendritic cell; DC, dendritic cell; DEGs, differentially expressed genes; DGE, differential gene expression; Exp, experiment; FAANG, Functional Annotation of Animal Genomes; FACS, Fluorescent activated cell sorting; G2P, Genome-to-Phenome; GO, gene ontology; GSEA, gene set enrichment analysis/analyses; HBSS, Hank's balanced salt solution; HEGs, highly enriched genes; MACS, Magnetic activated cell sorting; mDC/myDC, myeloid dendritic cell; n, negative; NK, natural killer; p, positive; PBMC, peripheral blood mononuclear cell; PC, principal component; PCA, principal component analysis; pDC, plasmacytoid dendritic cell; RF, random forest; RIN, RNA integrity number; RNA-seq, RNA sequencing; scRNA-seq, single-cell RNA sequencing; scREF-matrix, single-cell reference matrix; SEG, significantly enriched genes; sPCA, supervised principal component analysis; SWC6, swine workshop cluster 6.; T, T-cell.; TCR, T-cell receptor; TPM, transcripts per million; t-SNE, t-distributed stochastic neighbor embedding; UMAP, uniform manifold approximation and projection; UMI, unique molecular identifier; γδ, Gamma-delta; αβ, alpha beta.

Many scRNA-seq approaches do not require prior phenotypic/functional information or antibody reagents but instead rely on physical partitioning of cells to uniquely tag transcripts from individual cells and sharpen resolution of subsequent transcriptomic analysis to single cells (Liu and Trapnell, 2016; Vieira Braga et al., 2016; Zheng et al., 2017). scRNA-seq methods have been applied to human PBMCs (Zheng et al., 2017) and provide more accurate and detailed analyses of transcriptional landscapes that can identify new cell types (Villani et al., 2017) when compared to other transcriptomic approaches. There are limitations to scRNA-seq, with tradeoffs in total genes detected per cell vs. total cells captured for analysis, depending on the approach used (Wilson and Göttgens, 2018).

To deeply annotate the porcine genome for peripheral mononuclear immune cell gene expression and further inform phenotype and function of the heterogenous pool of immune cells in PBMC preparations, two approaches were used to isolate peripheral immune cells for RNA-seq. MACS followed by FACS was used to enrich for eight PBMC populations using population-specific cell surface markers, and RNA isolated from enriched populations was used for bulk RNA-seq (bulkRNA-seq) or a NanoString assay to evaluate gene expression. PBMCs were also subjected to droplet-based partitioning for scRNA-seq. Gene expression patterns of porcine immune cells using different approaches were compared to each other and to multiple human datasets. Complementary methods provided an improved annotation and deeper understanding of porcine PBMCs, as well as explicated datasets for further query by the research community.

MATERIALS AND METHODS

Animals and PBMC Isolation

Four separate PBMC isolations were performed, with different animals used in each experiment. Cells were used for bulkRNA-seq, targeted RNA detection (NanoString), or scRNA-seq. PBMCs from experiments were used as follows: Experiment A (ExpA) for bulkRNA-seq of sorted populations from two ∼6-month-old pigs (A1, A2); Experiment B (Exp B) for NanoString and scRNA-seq from three ∼12-month-old pigs (B1, B2, B3); Experiment C (ExpC) for scRNA-seq from three ∼12-month-old pigs (C1, C2, C3); Experiment D (ExpD) for scRNA-seq from two ∼7-week-old pigs (D1, D2). All pigs were crossbred, predominantly Large White and Landrace heritage. All pigs from experiments A, B, and C were male. In experiment D, animal D1 was female, and animal D2 was male. All animal procedures were performed in compliance with and approval by NADC Animal Care and Use Committee. PBMCs were isolated, enumerated, and viability assessed as previously described (Byrne et al., 2020).

Enrichment and Sorting Eight Leukocyte Populations by MACS/FACS

Peripheral blood mononuclear cells were labeled with biotin labeled anti-porcine CD3ε (PPT3, Washington State University Monoclonal Antibody Center) for 15 min at 4°C, mixing continuously. Cells were washed with Hank's Balanced Salt Solution (HBSS), incubated with anti-biotin microbeads (Miltenyi Biotec), placed on LS columns, and separated into $CD3\varepsilon^+$ and $CD3\varepsilon^-$ fractions according to manufacturer's directions (Miltenyi Biotec). $CD3\varepsilon^+$ and $CD3\varepsilon^-$ fractions were each fluorescently-sorted into four subpopulations based on surface marker expression shown in **Figure 1** and **Table 1**. For NanoString assays, B-cells were sorted as $CD3\varepsilon^-CD172\alpha^-CD8\alpha^-$; CD21 was not used for sorting. Each fraction for FACS was confirmed $CD3\varepsilon^+$ or $CD3\varepsilon^-$ by labeling with anti-mouse IgG1-PE-Cy7 to detect anti-CD3ε antibody used for MACS. Cells were sorted into supplemented HBSS using a BD FACSAria II with 70 mm nozzle. After sorting, cells were pelleted and enumerated as described above. Sorted cell purity was >85% for each population. Cells were stained, sorted, and further processed within 10 h of collection keeping cells on ice between processing steps.

RNA Isolation for BulkRNA-Seq/NanoString

BulkRNA-seq: after FACS, cells were pelleted, enumerated, and immediately lysed in RLT Plus buffer. RNA extractions were performed using the AllPrep DNA/RNA MiniKit (QIAGEN) following manufacturer's instructions. Eluted RNA was treated with RNase-free DNase (QIAGEN). RNA quantity/integrity were assessed with an Agilent 2200 TapeStation system (Agilent Technologies). Samples used had RNA integrity numbers (RINs) $\geq$ 7.9. From ExpA, only one RNA sample for NK cells was used.

For NanoString assay: after FACS, cells were pelleted, enumerated, and immediately stored in Trizol. RNA extraction was performed using the Direct-zol RNA MicroPrep Kit (Zymo) with on-column DNase treatment following manufacturer's instructions. RNA quantity and integrity were assessed as described above, with RINs $\geq$ 6.9. RNA was preserved at −80°C until further use.

BulkRNA-Seq Library Preparation and Data Analysis

RNA was fragmented and 15 libraries prepared using the TruSeq Stranded Total RNA Sample Preparation Kit (Illumina). Libraries were diluted and pooled in approximately equimolar amounts. Pooled libraries were sequenced in paired-end mode (2 × 150-bp reads) using an Illumina NextSeq 500 (300 cycle kit).

Preprocessing, Mapping, Alignment, Quality Control

Data processing was performed as previously reported (Herrera-Uribe et al., 2020) using R v4.0.3. Sscrofa 11.1 genome and annotation v11.1.97 were used. Counts per gene of each sample in the two count tables were added together to get the final count table. Given that different types of immune cells have different transcriptome profiles (Hicks and Irizarry, 2015), YARN (Paulson et al., 2017), a tissue type-aware RNA-seq data normalization tool, was used to filter and normalize the count table. Genes with extremely low expression levels (<4 counts in at least one cell type) were filtered out using *filterLowGenes()*. The final count table contained 12,261 genes across 15 samples, which was then normalized using *normalizeTissueAware()*, which leverages the smooth quantile normalization method (Hicks et al., 2018).

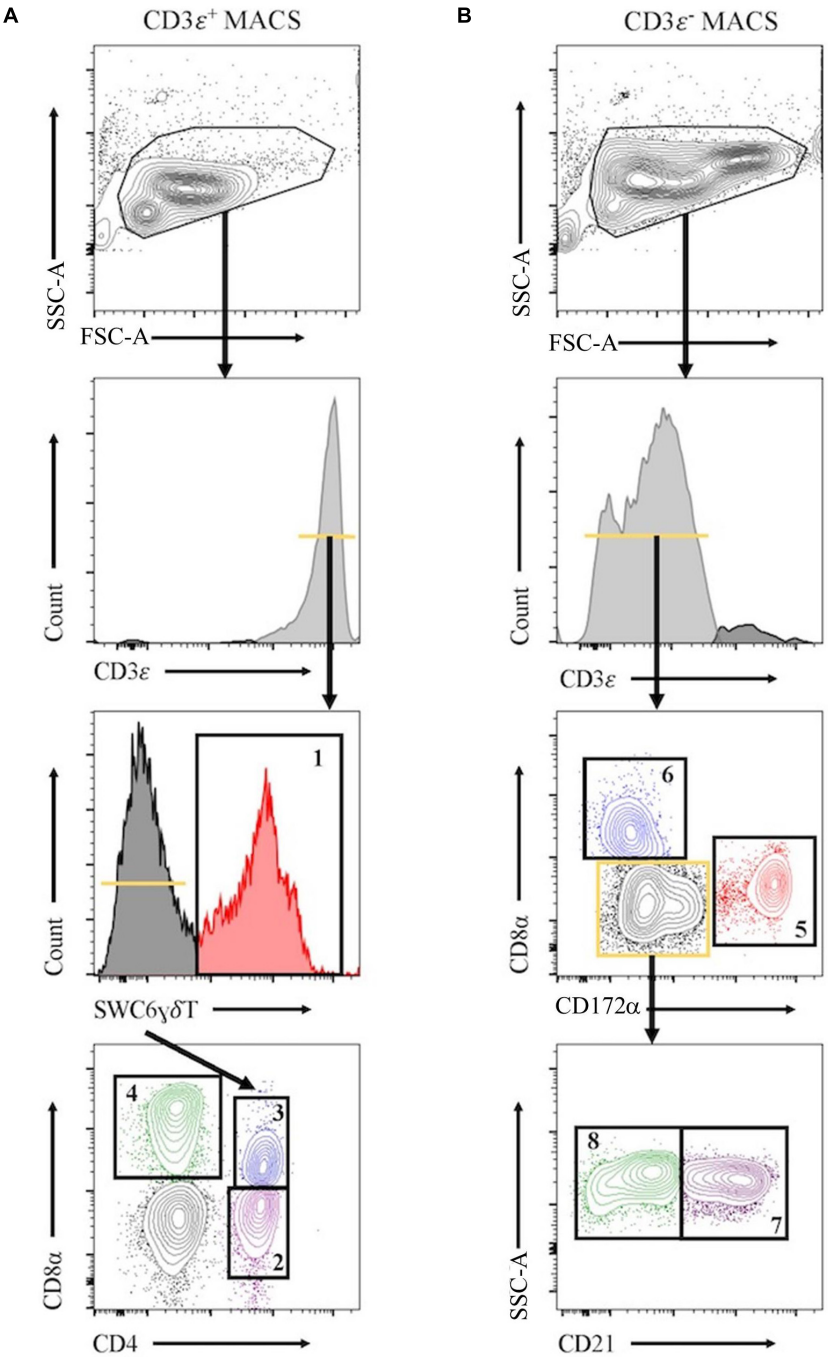


FIGURE 1 | Representative plots for fluorescence-activated cell sorting (FACS) isolation of 8 leukocyte populations from pig peripheral blood mononuclear cells (PBMCs). Porcine PBMCs were first subjected to magnetic-activated cell sorting (MACS) to enrich for CD3ε + and CD3ε– fractions. **(A)** Cells in $CD3\varepsilon^+$ MACS fraction were FACS gated on FSC vs. SSC, doublets removed (not shown), and $CD3\varepsilon^+$ cells were isolated into 4 population: $SWC6^+$ γδ T-cells (gate 1), and the $SWC6^-$ cells sorted as $CD4^+CD8\alpha^-$ (gate 2), $CD4^+CD8\alpha^+$ (gate 3), $CD4^-CD8\alpha^+$ (gate 4) T-cells. **(B)** Cells in $CD3\varepsilon^-$ MACS fraction were FACS gated on FSC vs. SSC, doublets removed (not shown), and $CD3\varepsilon^-$ cells were isolated into 4 populations: $CD172\alpha^+$ myeloid lineage leukocytes (gate 5), $CD8\alpha^+CD172^-$ NK cells (gate 6), and the remaining $CD8\alpha^-$ $CD172\alpha^-$, cells were isolated as $CD21^+$ (gate 7) and $CD21^-$ (gate 8) B-cells. **Table 1** outlines abbreviations and sort criteria for each population.

Data quality control was performed using DESeq2 (v1.24.0) (Love et al., 2014) within RStudio s (v1.2.1335). Regularized log-transformation was applied to the normalized count table with the rld function. Then principal component analysis (PCA) and sample similarity analyses were carried out and visualized using *plotPCA()* and *distancePlot()*, respectively.

Heatmaps to display enriched genes were created using pheatmap (v1.0.12) within RStudio.

Cell Type-Enriched and Cell Type-Specific Gene Identification

The normalized count table was used for differential gene expression (DGE) analysis with DESeq2 by setting the size factor for each sample to 1. A generalized linear model was fitted for each gene in the count table, with negative binomial response and log link function of the effect of cell types and pig subjects. *nbinomWaldTest()* was used to estimate and test the significance of regression coefficients with the following explicit parameter settings: betaPrior = FALSE,maxit = 50000,useOptim = TRUE,useT = FALSE,useQR = TRUE. Cell type-enriched genes and cell type-specific genes were identified using the results function separately. A gene was labeled as cell-type enriched if the expression level (averaged across replicates) in one cell type was at least 2× higher than the average across all remaining cell types and adjusted p-value < 0.05. A gene was labeled as cell type-specific if the averaged expression level in one cell type was at least 2× higher in pairwise comparison to the average in each other cell type and adjusted p-value < 0.05 (Benjamini and Hochberg, 1995). Heatmaps to display specific genes were created as mentioned above.

For cross-species comparison, human hematopoietic cell (Haemopedia) RNA-seq expression data (Hilton Laboratory at the Walter and Eliza Hall Institute[1]) was used. Only orthologous genes with one-to-one matches between human and pig [orthologous gene list obtained from BioMart (Durinck et al., 2009)] were compared. Orthologous gene transcript per million (TPM) values from naive and memory B-cells, myeloid dendritic cells (myDC), myeloid dendritic cells CD123 + (CD123PmDC), plasmacytoid DC (pDC), monocytes, NK cells, CD4T and CD8T cells from healthy donors were used (Choi et al., 2019). Spearman rank correlation analyses was performed to identify correlation between orthologous gene expression levels (absolute TPM) in pig and human sorted populations. Significance level was set at $p < 0.05$ and level of Spearman's rank correlation coefficient (rho) was defined as low (<0.29), moderate (0.3–0.49), and strong (0.5–1) correlation.

[1]https://www.haemosphere.org/datasets/show

Gene Ontology (GO) Enrichment Analysis

Metascape analysis (Zhou et al., 2019) was performed for GO analysis of the top 25% enriched genes and specific genes identified as described above, with threshold p-value < 0.01. Several terms were clustered into the most enriched GO term. Term pairs with Kappa similarity score > 0.3 were displayed as a network to show relationship among enriched terms. Terms associated with more genes tended to have lower p-values. All networks displayed were visualized using Cytoscape. All Ensembl Gene IDs with detectable expression level in each cell type were used as the background reference.

NanoString Assay and Data Analysis

A total of 230 test genes with nine housekeeping genes, eight positive and nine negative control genes were chosen for gene expression quantification on the NanoString nCounter analysis system (NanoString Technologies) using custom-made probes. The custom designed CodeSet was selected from genes and pathways associated with porcine blood, lung, lymph node, endometrium, placenta or macrophage response to infection with a porcine virus (Van Goor et al., 2020). RNA samples were diluted to 25–100 ng/ul in RNase-free water, and 5 ul of each sample was used in the assay using manufacturer's instructions with the nCounter Master kit.

The nCounter analysis system produces discrete count data for each gene assayed within each sample. We used the NanoString software nSolver Analysis Software (v3.0, NanoString Technologies), following manufacturer's instructions. The nSolver corrected for background based on negative control samples, performed within-sample normalization based on positive control probes, and performed normalization across samples using the median expression values of housekeeping genes (*GAPDH, HMBS, HPRT1, RPL32, RPL4, SDHA, TBP, TOP2B, YWHAZ*), providing confidence in our normalization method.

TABLE 1 | Abbreviations and phenotype information of pig sorted immune cells.

Gate[a]	Population Abbreviation	Marker	Clone	Fluorophore	Company (Catalog #)[b]	FACS Sort Criteria
CD3ε + MACS Fraction						
–	–	Anti mouse IgG1	RMG1-1	PE-Cy7	BioLegend(406614)	–
1	SWC6gdT	SWC6gdT	MAC320	APC	BD(561482)	$CD3\varepsilon^+SWC6^+$
2	CD4T	CD4	74-12-4	FITC	BD(559585)	$CD3\varepsilon^+SWC6^-CD4^+CD8\alpha^-$
3	CD4CD8T	CD4CD8α	74-12-4/76-2-11	FITC/PE	BD(559585)/BD(559584)	$CD3\varepsilon^+SWC6^-CD4^+CD8\alpha^+$
4	CD8T	CD8α	76-2-11	PE	BD(559584)	$CD3\varepsilon^+SWC6^-CD4^-CD8\alpha^+$
CD3ε − MACS Fraction						
5	Myeloid	CD172	74-22-15A	FITC	BD(561498)	$CD3\varepsilon^-CD172\alpha^+CD8\alpha^-$
6	NK	CD8α	76-2-11	PE	BD(559584)	$CD3\varepsilon^-CD172\alpha^-CD8\alpha^+$
7	CD21pB	CD21	BB6-11C9.6	AF647	Southern Biotech(4530-31)	$CD3\varepsilon^-CD172\alpha^-CD8\alpha^-CD21^+$
8	CD21nB	–	–	–	–	$CD3\varepsilon^-CD172\alpha^-CD8\alpha^-CD21^-$

*[a]Refers to gate in **Figure 1**. [b]Reagents listed in Materials and Methods.*

All statistical analyses were performed using the statistical programming language R v3.5. Raw count data were normalized using *normalizationFactors()* and *NanoStringDataNormalization()* from NanoStringDiff (v1.1.2.0) (Wang et al., 2016). One gene (*ISG20*) without detected expression in any samples was removed. Hierarchical clustering and PCA suggested there were substantial hidden variations among the expression data. Surrogate variable analysis has been shown to be a powerful method to detect and adjust for hidden variations in high throughput gene expression data (Li et al., 2014; Qian Liu, 2016), so surrogate variable analysis was applied to remove further hidden variations in the gene expression data using *svaseq()* from sva (v3.30.1) (Leek et al., 2012). A full model with cell subpopulations and RINs as independent variables, and a reduced model with RINs as the only independent variable were used. Three surrogate variables were estimated and used to adjust for the hidden variations.

Gene expression values were transformed to $\log_2$(TPM) using *voom()* from limma (Law et al., 2014). Linear mixed effect models were used to fit the transformed gene expression data by using *lmer()*in lme4 (Bates et al., 2015). The model included fixed effect for cell subpopulation, RIN, the three surrogate variables, and random effect for each animal. One minus Spearman correlation coefficient was used as distance measure for gene clustering, and Euclidian distances was used for sample clustering.

Additionally, Spearman correlation analysis was performed to assess the correlation between bulkRNA-seq and NanoString results. The significant level was set at $p < 0.05$, and the level of Spearman's rank correlation coefficient (rho) was defined as described above.

scRNA-Seq Library Preparation

Peripheral blood mononuclear cell isolation experiments were performed at different times and samples sequenced in different runs. For ExpB, 1×10^7 viable PBMCs per animal were cryopreserved according to 10× Genomics Sample Preparation Demonstrated Protocol, shipped on dry ice to University of Minnesota's Core Sequencing Facility, and thawed, partitioned, and scRNA-seq libraries prepared. For ExpC/ExpD, freshly isolated PBMCs were transported on ice to Iowa State University Core Sequencing Facility for partitioning and library preparation. Partitioning and library preparation were performed according to Chromium Single Cell 3′ Reagent Kits v2 User Guide (10× Genomics). For all experiments, 100 base paired-end reads were sequenced on an Illumina HiSeq3000 at ISU Core Sequencing Facility. One sample from ExpB was omitted from further analyses due to poor sequence performance.

scRNA-Seq Data Analysis

Read Alignment/Gene Quantification

Raw read quality was checked with FASTQC[2]. Reads 2 (R2) were corrected for errors using Rcorrector (Song and Florea, 2015), and 3′ polyA tails > 10 bases were trimmed. After trimming, R2 > 25 bases were re-paired using BBMap[3]. *Sus scrofa* genome Sscrofa 11.1 and annotation GTF (v11.1.97) from Ensembl were used to build the reference genome index (Yates et al., 2020). The annotation file was modified to include both gene symbol (if available) and Ensembl ID as gene reference (e.g., GZMA_ENSSSCG00000016903) using custom Perl scripts. Processed paired-end reads were aligned and gene expression count matrices generated using CellRanger (v4.0; 10× Genomics) with default parameters. Only reads that were confidently mapped (MAPQ = 255), non-PCR duplicates with valid barcodes, and unique molecular identifiers (UMIs) were used to generate gene expression count matrices. Reads with same cell barcodes, same UMIs, and/or mapped to the same gene feature were collapsed into a single read.

Quality Control/Filtering

All quality control/filtering steps were performed using R v3.6.2. CellRanger output files were used to remove ambient RNA from each sample with SoupX (Young and Behjati, 2020) function *autoEstCont()*. Corrected non-integer gene count matrices were outputted in CellRanger file format using DropletUtils (Lun et al., 2019) function *write10xCounts()* and used for further analyses. Non-expressed genes (sum zero across all samples) and poor quality cells (>10% mitochondrial genes, < 500 genes, or < 1,000 UMIs per cell) were removed using custom R scripts and Seurat (Stuart et al., 2019). Filtered count matrices were generated using *write10xCounts()* and used for further analyses. High probability doublets were removed using Scrublet (Wolock et al., 2019), specifying 0.07 expected doublet rate and doublet score threshold of 0.25.

Integration, Visualization, and Clustering

Integration, visualization, and clustering were performed with R v3.6.2 and Seurat v3.2.0. Post-quality control/filtering gene counts/cells from each sample were loaded into a Seurat object and transformed individually using *SCTransform()*. Data were integrated with *SelectIntegrationFeatures()*, *PrepSCTIntegration()*, *FindIntegrationAnchors()*, and *IntegrateData()* with default parameters. PCA was conducted with *RunPCA()*, and the first 14 principal components (PCs) were selected as significant based on <0.1% variation of successive PCs. Significant PCs were used to generate two-dimensional t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) coordinates for visualization with *RunTSNE()* and *RunUMAP()*, respectively, identify nearest neighbors and clusters with *FindNeighbors()* and *FindClusters()* (clustering resolution = 1.85), respectively, and perform hierarchical clustering with *BuildClusterTree()*. Counts in the RNA assay were further normalized and scaled using *NormalizeData()* and *ScaleData()*.

Differential Gene Expression (DGE) Analyses

Differential gene expression analyses were performed with R v3.6.2 and Seurat v3.2.0. Normalized counts from the RNA assay were used for DGE analyses. Differentially-expressed genes (DEGs) between pairwise cluster combinations were calculated using *FindMarkers()*. DEGs in one cluster relative to the

[2]https://www.bioinformatics.babraham.ac.uk/projects/fastqc

[3]https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/

average of all other cells in the dataset were calculated using *FindAllMarkers()*. The default Wilcoxon Rank Sum test was used for DGE analyses. Genes expressed in >20% of cells within one of the cell populations being compared, with | logFC| > 0.25, and adjusted *p*-value < 0.05 were considered DEGs.

Gene Set Enrichment Analyses (GSEA)

Enrichment of gene sets within our porcine scRNA-seq dataset were performed using AUCell (v1.10.0) (Aibar et al., 2017). Enriched genes in sorted porcine bulkRNA-seq populations were identified as described in preceding methods. Log2FC values were used to curate gene sets of genes enriched in the top 25, 20, 15, 10, 5, or 1% of bulkRNA-seq populations. Gene sets from human bulkRNA-seq cell populations (Choi et al., 2019) were recovered by performing a High Expression Search on the Haemosphere website[4], setting Dataset = Haemopedia-Human-RNASeq and Sample group = celltype. Gene sets for CD4: + T-cell; CD8: + T-cell; Memory B-cell; Monocyte; Myeloid Dendritic Cell; Myeloid Dendritic Cell CD123 + ; Naïve B-cell; Natural Killer Cell; and Plasmacytoid Dendritic Cell options corresponded to CD4T, CD8T, MemoryB, Monocyte, mDC, CD123PmDC, NaïveB, NK, and pDC designations, respectively. Genes with high expression scores >0.5 (lower enrichment level) or >1.0 (higher enrichment level) were selected and filtered to include only one-to-one gene orthologs as described in preceding methods. Human gene identifiers were converted to corresponding porcine gene identifiers or gene names used for scRNA-seq analyses.

Within each cell of the finalized scRNA-seq dataset, gene expression was ranked from raw gene counts. Area under the curve (AUC) scores were calculated from the top 5% of expressed genes in a cell and the generated gene sets. Higher AUC scores indicated a higher percentage of genes from a gene set were found amongst the top expressed genes for a cell. For overlay of AUC scores onto UMAP coordinates of the scRNA-seq dataset, a threshold value was manually set for each gene set based on AUC score distributions. For visualization by heatmap, AUC scores were calculated for each cell, scaled relative to all other cells in the dataset, and average scaled AUC scores were calculated for each cluster. R v4.0.2 was used.

Deconvolution Analysis (CIBERSORTx)

To deconvolve cluster-specific cell subsets from bulkRNA-seq of sorted populations, CIBERSORTx (Newman et al., 2019) was used to derive a signature matrix from scRNA-seq data. 114 cells were taken from each cluster using the Seurat *subset()* function and labeled with corresponding cluster identities. Cluster-labeled cells were used to obtain a single-cell reference matrix (scREF-matrix) that was used as input and run on CIBERSORTx online server using "Custom" option. Default values for replicates (5), sampling (0.5), and fraction (0.0) were used. Additional options for kappa (999), *q*-value (0.01), and No. Barcode Genes (300–500) were kept at default values. CIBERSORTx scREF-matrix was used to impute cell fractions from the bulkRNA-seq of sorted cell population "mixtures." The mixture file (TPM values) was used as an input and run on CIBERSORTx online server using the "Impute Cell Fractions" analysis with the "Custom" option selected, and S-mode batch-correction was applied. Cell fractions were run in relative mode to normalize results to 100%. The number of permutations to test for significance were kept at default (100). Resulting output provided estimated percentages of what scRNA-seq clusters defined each bulkRNA-seq sorted cell population.

Reference-Based Label Transfer/Mapping and *de novo* Integration/Visualization

R v3.6.3 and Seurat v3.9.9.9010 were used for the analyses described in this section. A CITE-seq dataset of human PBMCs (Hao et al., 2020) was used to transfer cell type annotations onto our porcine scRNA-seq dataset. Due to the cross-species comparison, we distilled human reference and pig query datasets to only include 1:1 orthologous gene, and human reference dataset was re-normalized and integrated mirroring previous methods (Hao et al., 2020). Each sample of the porcine query dataset was separately normalized using SCTransform. Anchors were found between the human reference and each pig query sample using FindTransferAnchors. Identified anchors were used to calculate mapping scores for each cell using MappingScore. The mapping scores provided a 0–1 confidence value of how well a porcine cell was represented by the human reference dataset. Prediction scores were calculated using available level 2 cell types from the human reference dataset. Prediction scores provided a 0–1 percentage value for an individual cell type prediction, based on how many nearby human cells shared the same cell type annotation that was predicted. Predicted cell annotations were projected back onto original UMAP of the porcine dataset. Cluster-averaged prediction and mapping scores were also calculated.

In order to identify cells from the porcine dataset that were not well represented by the human reference dataset the two datasets were integrated to perform *de novo* visualization by merging the two datasets and their respective sPCAs to create a new UMAP. From two-dimensional *de novo* UMAP, porcine cells that did not overlap with human cells were identified.

Cluster Subsetting

For deeper analyses of only subsets of clusters in the scRNA-seq dataset, cells belonging to only selected clusters were place in a new Seurat object using *subset()*. Genes with zero overall expression in the new data subset were removed using *DietSeurat()*, and counts were re-scaled with *ScaleData()*. Original cluster designations and PCs were left intact. UMAP/t-SNE visualization, hierarchical clustering, and DGE analyses were re-performed as described in the original analyses. Pairwise DGE analyses were not re-performed. R v4.0.2 was used.

Random Forest (RF) Modeling

Random forest modeling was performed with R version 4.0.4. The RF models provided an estimate of cluster similarity based on error rates. The R packages caret[5] and ranger[6] were used to create

[4]https://www.haemosphere.org/searches

[5]https://cran.r-project.org/web/packages/caret/caret.pdf

[6]https://cran.r-project.org/web/packages/ranger/index.html

RF models trained on cluster identities of cells. A normalized count matrix was used as input data for RF models. Each cell was labeled by its previously defined cluster. Two different types of models were created: (1) pairwise models where training data included only cells from two different clusters (ex. Clusters 0 and 3); (2) models where training data included cells from all clusters of a specified dataset (ex. all γδ T-cell clusters). Each model was trained on the cluster identity of each cell, with trees created = 500, target node size = 1, variables = 14,386, variables to sample at each split (Mtry) = 119. Each tree in the model is grown from a bootstrap resampling process that calculates an out-of-bag (OOB) error that provides an efficient and reasonable approximation of the test error. Variable importance was used to find genes or sets of genes that can be used to identify certain types of cells or discriminate groups of cells from one another. RF models are advantageous because they can provide ranked lists of genes most important for discriminating cells between different clusters. This method was used to identify groups of important genes to supplement single DGE analyses. Variable importance was assigned by measuring node impurity (Impurity) and using permutations (Permutation). Features that reduced error in predictive accuracy are ranked as more important. High error rate in the model suggests cells from the groups being compared are more similar to each other, whereas low error rate suggests cells from each cluster are unique.

Gene Name Replacement

Several gene names/Ensembl IDs used for data analysis were replaced in main text/figures for the following reasons: gene symbol was not available in the annotation file but was available under the gene description on Ensembl, gene symbol was updated in future Ensembl releases, or multiple Ensembl IDs corresponded to a single gene symbol. Affected genes included: *ABI3 = ENSSSCG00000035224, ABRACL = ENSSSCG00000004145, ANP32E = ENSSSCG00000035209, AP3S1 = ENSSSCG00000037595, AURKA = ENSSSCG00000007493, BANF1 = ENSSSCG00000012961, BUB1B = ENSSSCG00000004782, CBX3 = ENSSSCG00000016711, CCDC12 = ENSSSCG00000011329, CCL23 = ENSSSCG00000033457, CD163L1 = ENSSSCG00000034914, CDC2 = ENSSSCG00000010214, CDNF = ENSSSCG00000039658, CEP57 = ENSSSCG00000014969, CLIC1 = ENSSSCG00000039071, CMC2 = ENSSSCG00000032060, CR2 = ENSSSCG00000028674, CRIP1 = ENSSSCG00000037142, CRK = ENSSSCG00000038989, DBF4 = ENSSSCG00000020870, DEK = ENSSSCG00000001075, DHFR = ENSSSCG00000031117, EEF1A1 = ENSSSCG00000004489, KIF23 = ENSSSCG00000004969, FAM72A = ENSSSCG00000039370, FCGR3A = ENSSSCG00000036618, GBP1 = ENSSSCG00000024973, GBP7 = ENSSSCG00000006919, GFER = ENSSSCG00000008035, GGCT = ENSSSCG00000016679, GIMAP4 = ENSSSCG00000027826, GINS1 = ENSSSCG00000034913, GTSE1 = ENSSSCG00000000002, GZMA = ENSSSCG00000016903, H1-2 = ENSSSCG00000037565, HMGB1 = ENSSSCG00000009327, HMGB3 = ENSSSCG00000035908, HMGN1 = ENSSSCG00000033733, HMGN5 = ENSSSCG00000032946, HNRNPA2B1 = ENSSSCG00000036350, HNRNPAB = ENSSSCG00000014031, HOPX = ENSSSCG00000008898, IDI1 = ENSSSCG00000029066, IFITM1 = ENSSSCG00000014565, IGLL5 = ENSSSCG00000010077, JPT1 = ENSSSCG00000017213, KLRB1B = ENSSSCG00000034555, KLRC1 = ENSSSCG00000000640, KLRD1 = ENSSSCG00000026217, KNL1 = ENSSSCG00000039107, LSM4 = ENSSSCG00000034314, LSM5 = ENSSSCG00000026064, MAGOHB = ENSSSCG00000000635, MAL = ENSSSCG00000040098, MAN2B1 = ENSSSCG00000013720, MDK = ENSSSCG00000013260, MKI67 = ENSSSCG00000026302, MYL12A = ENSSSCG00000003691, NT5C3A = ENSSSCG00000022912, NUSAP1 = ENSSSCG00000035544, NUTF2 = ENSSSCG00000030295, PPIA = ENSSSCG00000016737, PRIM1 = ENSSSCG00000026055, PRKCH = ENSSSCG00000005095, PTTG1 = ENSSSCG00000017032, RPL14 = ENSSSCG00000011272, RPL22L1 = ENSSSCG00000036114, RPL23A = ENSSSCG00000035080, RPL35A = ENSSSCG00000040273, RPS15A = ENSSSCG00000035768, RPS19 = ENSSSCG00000003042, RPS27A = ENSSSCG00000034617, RPS3 = ENSSSCG00000014855, RPS8 = ENSSSCG00000003930, RRM2 = ENSSSCG00000031741, S100B = ENSSSCG00000026140, SEPHS1 = ENSSSCG00000031659, SIRPA = ENSSSCG00000028461, SKA1 = ENSSSCG00000004518, SLA-DQA1 = ENSSSCG00000001456, SLA-DRA = ENSSSCG00000001453* (listed as HLA-DRA in the gene annotation used), *SLA-DRB1 = ENSSSCG00000001455, SLPI = ENSSSCG00000022258, SNRPG = ENSSSCG00000024776, SPIB = ENSSSCG00000034211, TACC3 = ENSSSCG00000008677, TMSB15B = ENSSSCG00000012517, TMSB4X = ENSSSCG00000012119, TUBA1C = ENSSSCG00000000194, TXN = ENSSSCG00000005453, WEE1 = ENSSSCG00000013411, WIPF1 = ENSSSCG00000027348, YBX1 = ENSSSCG00000028485.*

RESULTS

BulkRNA-Seq Revealed Common and Distinct Transcriptomes in Circulating Immune Cells

Eight immune cell populations (**Table 1**) were sorted by cell-surface marker phenotypes for transcriptomic profiling by bulkRNA-seq (**Figure 1**) using primarily criteria previously outlined (Gerner et al., 2009), with some modifications. Our protocol utilized an antibody reactive to swine workshop cluster 6 (SWC6) protein to identify γδ T-cells, but the antibody only labels $CD2^-$ γδ T-cells (Yang and Parkhouse, 1996; Davis et al., 1998; Stepanova and Sinkora, 2013; Sedlak et al., 2014). $CD2^+$ γδ T-cells were likely sorted into the $CD3\varepsilon^+CD4^-CD8\alpha^-$ fraction that was not retained or the CD8T ($CD3\varepsilon^+CD4^-CD8\alpha^+$) population (Davis et al., 1998; Stepanova and Sinkora, 2013; Sedlak et al., 2014). A pan-B-cell marker for pigs is not currently available, so B-cells are often characterized through a series of negative gates. Cells in the $CD3\varepsilon^-$ fraction were considered B-cells if they also lacked expression of CD172α and CD8α. B-cells characterized in this manner were further terminally sorted into B-cell populations with or without CD21 (complement receptor 2) expression (CD21pB and CD21nB, respectively; **Figure 1**, gates 7 and 8 respectively). We acknowledge that the CD21nB gate likely contained other circulating cell types that were not sorted through positive gating approaches.

Transcriptomic profiles of sorted cell populations were constructed by bulk RNA-seq, and relationships among porcine immune cell transcriptomes were assessed and visualized through dimensionality reduction and hierarchical clustering (**Figures 2A,B** and **Supplementary File 1**). Specifically, T-cell populations (SWC6gdT, CD4T, CD4CD8T, CD8T), B-cell populations (CD21pB, CD21nB), myeloid leukocyte populations (Myeloid), and a single NK cell population (NK) were well separated from each other (**Figure 2A**) by PCA. Replicates of specific sorted cell populations clustered most closely together, while within T-cell populations or B-cell populations, considerable transcriptional similarity was observed (**Figure 2B**).

The total number of expressed genes in each sorted population was similar (**Supplementary File 1**). Significantly enriched genes (SEGs) with expression significantly different and at least 2× greater than the average of all other cell populations (see Methods) were identified for each sorted population (**Supplementary File 2**). Notably, around 11–23% of SEGs are not fully annotated (no symbol/gene name) in the Sscrofa 11.1 genome and annotation v11.1.97. However, there is evidence of 1:1 orthology within Ensembl for 5–14% of these unannotated genes (**Table 2**). The SWC6gdT population had the highest number of SEGs (3,591), while the NK population had the fewest (1,885) (**Table 2**). SEG lists were queried for corresponding protein targets used to sort cells, if known, to confirm enrichment of expression of genes corresponding to protein phenotypes (**Figure 2C**). Expression of *SIRPA** (encoding CD172α) had the highest fold-change in the Myeloid population, and *CR2* (encoding CD21, *ENSSSCG00000028674*), was highest in the CD21pB population, as would be predicted based on protein phenotypes. The two $CD4^+$ T-cell populations (CD4T and CD4CD8T) had the highest fold-change for *CD4*. The CD8T population had the highest fold-change for *CD8A*, with CD4CD8T and NK populations also having near a $\log_2$FC enrichment value of 5, in line with these populations also expressing CD8α. The SWC6gdT population had the highest fold-change for *TRDC*, though CD8T and CD21nB populations also had enrichment for *TRDC*. As noted previously, it's unlikely our sorting for γδ T-cells based on SWC6 captured all γδ T-cells, thus some γδ T-cells may be represented in other sorted populations. Thus, the CD8T population is likely comprised not only $CD8\alpha^+$ αβ T-cells, but also potentially $SWC6^-$ γδ T-cells expressing CD8α.

A subset of SEGs (25% highest log2FC values) for each sorted population, referred to as highly enriched genes (HEGs) that distinguish different circulating pig immune cell populations, were used for data visualization and GO analysis. The $\log_2$FC values for HEGs were clustered and visualized in **Figure 3** (four $CD3\varepsilon^-$ populations) and **Supplementary Figure 2** (four $CD3\varepsilon^+$ populations). GO analyses using HEG lists for each cell population indicated enrichment for biological processes characteristic of each respective cell population, depicted as networks of similar terms (**Figures 3E–H**, **Supplementary File 3**, and **Supplementary Figures 2E–H**). Terms for Myeloid HEGs included Myeloid leukocyte activation and response to bacterium (**Figure 3E**), and terms for NK HEGs included positive regulation of cell killing and natural killer cell mediated cytotoxicity (**Figure 3F**). Many terms enriched for CD21pB HEGs overlapped with those for CD21nB HEGs, as 38% of HEGs were shared between these populations (**Figures 3C,D**). Thus, top GO terms for B-cells, including adaptive immune response and B-cell proliferation were present in both populations (**Figures 3D,G**). However, some GO terms were unique to either B-cell population. GO related to B-cell activation, such as positive regulation of B-cell activation/proliferation processes associated with B-cell receptor

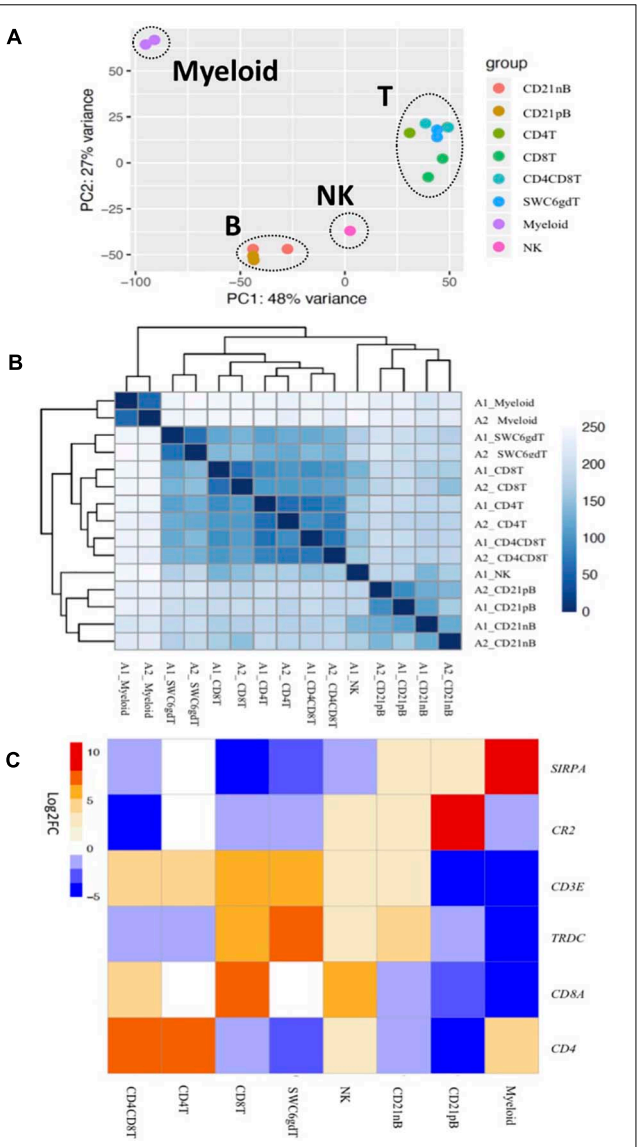


FIGURE 2 | Transcriptional expression patterns of immune cells are distinct and cluster more by progenitors. **(A)** Principal component analysis of transformed RNA-seq reads counts for whole transcriptomes. Axis indicate component scores. **(B)** Heat map depicting hierarchical clustering of sample-to-sample distance. Gene expression for whole transcriptomes were used to calculate sample to sample Euclidean distance (color scale) for hierarchical clustering. **(C)** Heatmap showing cell-type enriched gene values (Log2FC) between sorted immune cells. Gene coding proteins that were used for cell sorting were display.

signaling, were identified exclusively for CD21pB HEGs. For CD21nB HEGs, processes associated with humoral immunity and red blood cell processes such as coagulation or platelet activation were noted, which could indicate contamination of different cell types given the non-specific cell sorting approach used for CD21nB cells (**Figure 1**). For all sorted T-cell populations (CD8T, CD4T, CD4CD8T and SWC6gdT), HEG lists showed overlap (**Supplementary Figures 2A–D**). GO terms included T-cell activation, T-cell receptor signaling pathway, cytokine-cytokine receptor interaction and biological processes related to cytotoxicity activity (**Supplementary Figures 2E,F**). Overall, GO exploration of HEGs for sorted populations provided evidence that sorted immune cells represented expected immune cell functions.

The TPM values of expressed genes in sorted porcine cells were compared with orthologous human genes expressed in sorted human naïve hematopoietic cells from the Haemopedia (Choi et al., 2019) in order to identify cell-specific transcriptome similarities across species. Gene expression correlations assessed by Spearman's rank correlation indicated highly significant and moderately strong correlations (rho = 0.30–0.43, $p < 2.2e{-}16$) between porcine and anticipated corresponding human immune cell populations (**Supplementary Figure 3** and **Supplementary File 4**). A closer evaluation of genes reported as canonical cell markers for different mouse and human peripheral immune cell populations and expression of those genes in each of the sorted porcine populations revealed several commonalities. Specifically, genes such as *EBF1*, *CD19*, *MS4A1, CD79B*, *PAX5*, *HLA-DOB* (in CD21nB, CD21pB); *CD28* (in CD8T, CD4T, CD4CD8T); *CD5* (in CD8T, CD4T, CD4CD8T, SWC6gdT); *GZMA*, *GNLY*, *CCL5*, *KLRK1*, *KLRB1*, *CD244* (in NK, CD8T); and *VLDLR*, *NLRP3*, *CD14*, *STEAP4*, *CD163*, *DEFB1* (in Myeloid) for human cells showed specific enrichment in respective porcine populations (**Supplementary Figure 4**). Thus, additional query confirmed sorted porcine immune cell populations were equivalent to human counterparts in many ways.

TABLE 2 | Cell type-enriched and cell type specific genes identified in pig sorted immune cells.

Cell type	Enriched genes (SEG)	Top 25% SEG	Number of genes without gene name in the top 25% SEG list	Number of top 25% SEG without gene name that have orthologs in human	Specific genes
SWC6gdT	3591	898	141	15	8
CD8T	3318	830	150	19	2
CD4CD8T	2271	568	85	7	0
CD4T	2606	533	95	13	0
NK	1855	464	100	9	29
Myeloid	3440	860	102	15	397
CD21pB	2383	596	124	9	5
CD21nB	2456	614	146	7	0

High Homogeneity Amongst Sorted T-Cell and B-Cell Populations and Transcriptomic Distinctions in Myeloid and NK Populations

Pairwise DGE analyses between the cell populations identified genes with transcript abundance at least 2× higher in one population than in all other populations (adjusted p-value < 0.05, see Methods) which we define as cell type-specific. Consistent with PCA (**Figure 2A**), more cell type-specific genes were identified in the Myeloid population than in NK, T or B-cells. In total, we identified 2, 5, 8, 29, and 397 cell type-specific genes for CD8T, CD21pB, SWC6gdT, NK, and Myeloid populations, respectively (**Table 2** and **Supplementary Figure 5**). GO analyses using cell type-specific genes for the Myeloid population resulted in enrichment of terms such as Myeloid leukocyte, cytokine-cytokine receptor interaction, and pattern recognition receptor activity (**Supplementary Figure 5** and **Supplementary File 3**). Next, we determined if the cell type-specific genes identified were present in the list of HEGs for each population. In total, 2, 2, 5, 14, and 271 cell type-specific genes were identified in respective HEG lists for CD8T, CD21pB, SWC6gdT, NK, and Myeloid populations, respectively (**Table 3**), indicating the most highly-enriched cell type-specific genes were present in NK and Myeloid populations. Cell type-specific genes could not be identified for the remaining three sorted populations (CD4T, CD4CD8T, and CD21nB) using the criteria described above, indicating between-population transcriptional heterogeneity even for these enriched populations.

We then explored immune cell transcriptomic patterns to identify genes that could expand our knowledge of pathways active in specific cell populations, as well as predict new genes suitable to use for molecular analyses in immunology studies. Of interest, we found a remarkably high number of HEGs in our Myeloid population (**Table 3**), including immune-related genes involved in TLR signaling (*CD14*, *CD36*, *TLR2/3/4/8/9*, *NOD2*) and cytokine activity (*CSF1R*, *CSF2RA*, *CSF3R*, *IFNGR1*, *IL1B*, *IL1RAP*, *CXCR2*, *CCL21*, *CCL23*, *TNFRSF1B*, *IL1R2*, *TNFSF13*, *TNFSF13B*, *TNFRSF21*, *CXCL16*, *CCR2*). In NK cells fewer specific genes were detected than the Myeloid population (**Table 3**), with genes such as *OTOP2*, *OTOP3*, *OSPBL3*, *LY6D*, *RET* related to cytotoxic activity, a typical characteristic of NK cells (Rusmini et al., 2013; Rusmini et al., 2014; Belizário et al., 2018; Costanzo et al., 2018; Tu et al., 2018; Upadhyay, 2019), although their function in porcine NK cells is unexplored. In CD21pB cells, the gene for CD21 (*CR2*[*]) used for sorting the B-cell populations was predicted to be a HEG. The SWC6gdT population showed specific expression of *AVCR2A*, which is a Th17 cell specific gene in mice (Ihn et al., 2011) and regulates the proliferation of γδ T-cells in murine skin (Antsiferova et al., 2011). The CD8T population specifically expressed *TMIGD2* (a CD28 family member) and *JAML*, which encode T-cell transmembrane proteins (Zhu et al., 2013; Alvarez et al., 2015; Krueger et al., 2017).

[*] Refer to gene name replacement in Materials and Methods section

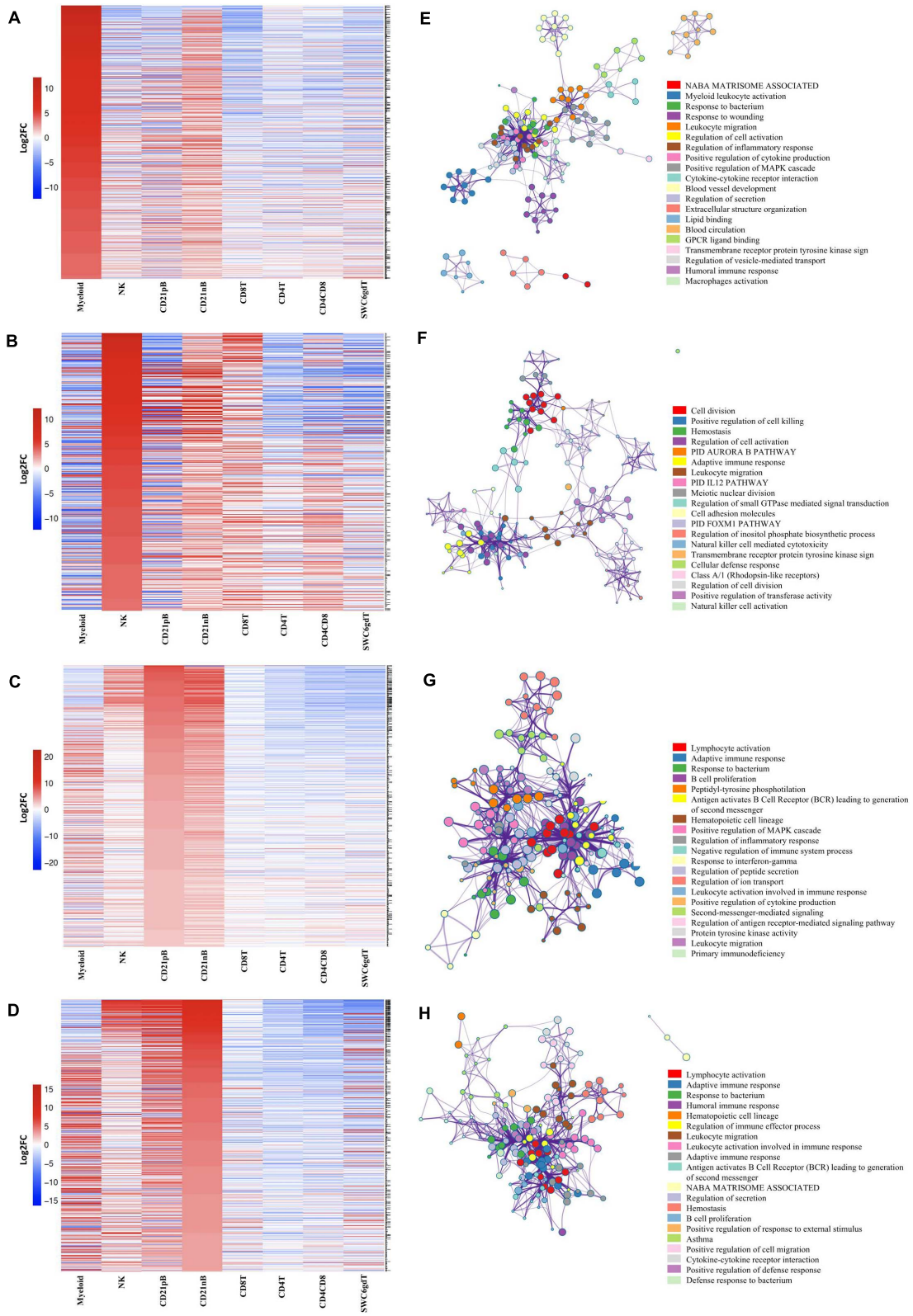


FIGURE 3 | Top 25% highly enriched genes in CD3^{-} sorted cells. Heatmap showing in decreasing order the top 25% of highly enriched genes in **(A)** myeloid, **(B)** NK, **(C)** CD21pB, and **(D)** CD21nB-cells. Ontology enrichment clusters of the top 25% highly enriched genes of **(E)** myeloid, **(F)** NK, **(G)** CD21pB, and **(H)** CD21nB cells. The most statistically significant term within similar term cluster was chosen to represent the cluster. Term color is given by cluster ID and the size of the terms is given by –log10 *p*-value. The stronger the similarity among terms, the thicker the edges between them.

TABLE 3 | Specific highly enriched genes in myeloid, NK, CD21pB, SWC6gdT, and CD4CD8T-cells.

Group	Total genes	Gene names
Specific Myeloid + Top 25% myeloid	271	*SLC18A1, ENSSSCG00000025687, KLHL13, PAK1, C1RL, MITF, SIRPB2, ENSSSCG00000014997, HNMT, C5AR1, A2M, TEK, SEL1L3, TSPAN13, ENSSSCG00000035960, ENSSSCG00000039214, ENSSSCG00000003226, APOE, CHST15, DNM1, GAS2L1, SERPING1, COL18A1, CDS1, ENSSSCG00000016184, CRHBP, KCNE3, NCAM1, ABHD12, ENSSSCG00000001850, ENSSSCG00000023479, ASAH1, FN1, ENSSSCG00000003554, ENSSSCG00000038429, GAA, ECE1, SLC46A2, UBTD1,CEBPD, CTSB, ENSSSCG00000031640, ENSSSCG00000037466, PLAC9, CCDC60, DOPEY2, TALDO1, ADAMTSL4, ENSSSCG00000034555, STK3, ENSSSCG00000021675, FAM129B, SIGLEC1, SULF2, TRPM2, MGP, CMKLR1, TNFRSF19, DOCK4, ENSSSCG00000027991, ULBP1, SLC11A1, SFXN3, TNFSF13, ENSSSCG00000013380, CD68, KCNQ1, RPS6KA2, CD14, MCF2L, ENSSSCG00000037541, ENSSSCG00000015839, PAM, SERPINB8, TSPAN12, F13A1, SASH1, C9orf72, PLCB4, SH3PXD2B, BLVRA, CXCL2, ADAM28, GPBAR1, CHI3L2, SNX9, LGALS3, SLC2A6, ENSSSCG00000035675, EHD4, ENSSSCG00000039758, UNC13A, ENSSSCG00000038418, C2, PLA2G7, FUCA1, ENSSSCG00000037426, ENSSSCG00000025271, ABCA9, RASGRP4, SLC7A7, VCAN, SLC39A8, ADAP2, SMIM5, DAGLA, RAB11FIP5, ZNF768, ENSSSCG00000007644, CTNND1, ENSSSCG00000022258, ENSSSCG00000017754, STXBP1, ENSSSCG00000027665, MANSC1, RND2, IGSF6, BMX, NLRP12, TPST1, NOD2, TREM1, SEMA6B, JDP2, FAM111B, CIDEB, ENSSSCG00000033457, MMP19, SGK3, CTTNBP2NL, MAPK4, PLAUR, INSIG1, RNASE4, FLVCR2, SCARF1, BCL2L14, ENSSSCG00000026196, MCTP1, WLS, ENSSSCG00000017920, PLOD1, CHPT1, PRCP, ENSSSCG00000013842, SH2D6, CA13, PLCB2, CAPN3, PRAM1, ENSSSCG00000038616, ALOX5, GPNMB, ACVRL1, SMIM3, GPR137B, LAMP1, NR1H3, ARL11, ITGB4, CYSLTR2, CCSER1, NCF2, GPCPD1, PDXK, NACC2, FOLR1, ADGRL2, MERTK, OLFM1, PLXNC1, ECM1, LRRC25, IFIT2, CORO1B, ASAP3, SLC43A3, STEAP4, CAMKK1, CTSS, TMEM47, TTLL7, AKR7A2, ENSSSCG00000036342, VIM, TLR8, LIN7A, MPP1, TBXAS1, LIPA, DRAM1, MRC2, TGM3, HEXB, GALM, EREG, JPH4, ANG, QPCT, PPT1, ARRDC4, RAB31, ABHD17C, NFAM1, TLR3, LTB4R, HSD3B7, VDR, ENSSSCG00000010497, CD163, OSCAR, DSC3, LRP6, ENSSSCG00000031951, ENSSSCG00000028635, PSAP, SCPEP1, EPB41L2, ZDHHC9, IL1R2, EXPH5, ENSSSCG00000023264, IFIT5, AGPAT2, NKD2, GUCY1B1, GLUL, COL14A1, TNFRSF1B, SLC16A3, GRN, ENSSSCG00000013100, CEBPA, OLFML2B, TLR4, XG, CCL21, ATF6, SLC49A3, HFE, ACVR1B, IFNGR1, ENSSSCG00000022925, SERPINB10, TCF7L2, ENSSSCG00000008769, ENSSSCG00000016093, UNC93B1, TIMP2, RAMP2, F11R, LGALS8, ENSSSCG00000032723, CFP, ZNF385A, CLIC2, TDRD1, HIP1, ENSSSCG00000026653, GSDMD, CSF1R, NAGK, GAB1, PGD, ENSSSCG00000034639, LRPAP1, DAPK1, ENSSSCG00000039956, GPAT3, GALNTL5, ENSSSCG00000029414*
Specific NK + Top 25% NK	14	*OTOP2, B3GNT7, OSBPL3, NR4A3, IGF2BP2, OTOP3, ENSSSCG00000010703, LY6D, RET, TUBB6, ENSSSCG00000033385, ENSSSCG00000036743, PTH1R, SUSD1*
Specific CD21pB + Top 25% CD21pB	2	*GP2, CR2*
Specific SWC6gdT + Top 25% SWC6gdT	5	*TMEM87B, ACVR2A, ENSSSCG00000028443, SLC4A4, CASS4*
Specific CD8T + Top 25% CD8T	2	*TMIGD2, JAML*

Finally, we compared pair-wise transcriptome differences between our porcine sorted CD4T and CD8T populations (**Supplementary File 2**) with the comparable populations from a previous study (Foissac et al., 2019). Even though the sorting approaches were different, 85% of the genes more highly expressed in CD4T compared to CD8T, respectively, were detected by Foissac and colleagues in their respective $CD4^+$ to $CD8^+$ comparison. Similar overlap was found (87%) for the genes more abundant in the "CD8 + high" list, while little overlap was found in the inverse comparisons (2.5 and 1%), strongly indicating these cell type gene expression patterns were similar between studies. However, given the lack of identification of cell-type specific genes for CD4T and CD8T populations, shared gene expression patterns may not be surprising.

NanoString Assay Validated BulkRNA-Seq

RNA abundance of each gene target (**Supplementary File 5**) in each sample was used to perform a hierarchical clustering analysis (**Supplementary Figure 6**). Similar to relationships observed in the bulkRNA-seq dataset, biological replicates clustered most closely together. T-cell populations (SWC6gdT, CD4T, CD4CD8T, CD8T) were more similar to each other than to other populations, with the exception of NK cells. RNA abundance for the genes encoding the marker proteins used for sorting cell populations confirmed cell identity in NanoString assays (**Supplementary Figure 7**). RNA abundance for each tested gene and cell population is included in **Supplementary File 5**. To validate gene expression levels calculated by bulkRNA-seq, a Spearman rank correlation analysis was performed between expression values determined by bulkRNA-seq and NanoString (**Supplementary Figure 8**). Highly significant and strong correlation (rho = 0.62–0.88, p-value $< 2.2e{-}16$) was observed for all sorted cell types (**Supplementary File 4**). Overall, gene expression estimates in the bulkRNA-seq dataset were confirmed by using the NanoString assay.

Defining the Transcriptomic Landscape of Porcine PBMCs at Single-Cell Resolution

Single-cells from PBMCs of seven conventional pigs were partitioned, sequenced, clustered, and visualized

(**Supplementary File 6**). In total, the final dataset included 28,810 cells, and each cell was assigned to one of 36 transcriptionally distinct clusters, with 9,176–12,683 genes detected within each cluster (**Figure 4A**, **Supplementary Figures 9A–C**; and **Supplementary File 6**). For identification of general cell types in each cluster, expression levels of genes known to be active in distinct porcine immune cell populations were mapped across single-cell clusters (**Figures 4B,C**). The 36 clusters were deduced to 13 general cell types (**Figure 4D**) as described below.

Monocyte clusters (13, 19, 20, 25, 27) expressed *CSF1R* and genes associated with microbial recognition (*CD14, CD163, NLRP3, TLR4*), reported as highly expressed by porcine monocytes (Auray et al., 2016). DC clusters (30, 32) expressed porcine pan-DC marker *FLT3* and were further classified as conventional DCs (cDCs; cluster 30) by elevated expression of *FCER1A* and MHCII-encoding genes (*SLA-DRB1*, SLA-DRA**) and pDCs (cluster 32) by elevated expression of *TCF4, XBP1, CLEC12A, CD93, IRF8, CD4,* and *CD8B* (Auray et al., 2016). Co-stimulatory gene *CD86* was expressed by all monocyte and DC clusters as reported (Auray et al., 2016). *SIRPA**, encoding CD172α is expressed by porcine monocytes/DCs (Piriou-Guzylack and Salmon, 2008; Auray et al., 2016) and used to sort myeloid leukocytes for bulkRNA-seq above, was minimally expressed in DC clusters.

B-cell clusters (2, 7, 8, 10, 11, 15, 16, 23, 26, 33) expressed *CD79A, CD19,* and *PAX5* (Faldyna et al., 2007; Piriou-Guzylack and Salmon, 2008; Bordet et al., 2019). Antibody-secreting cells (ASCs; cluster 29) expressed *IRF4* and *PRDM,* genes ascribed to immunoglobulin secretion (Shi et al., 2015; Liu et al., 2020). Detection of *CR2**, the gene encoding CD21 protein, was very low in any cluster.

Expression of *CD3E*, which encodes pan-T-cell CD3ε protein, identified T-cell clusters (0, 3, 4, 5, 6, 9, 12, 14, 17, 18, 21, 22, 24, 28, 31) (Gerner et al., 2009). Cluster 1 cells largely lacked *CD3E, CD5,* and *CD6* expression, while expressing *CD2, CD8A*, *PRF1,* NK receptor-encoding genes *KLRB1* (CD161) and *KLRK1* (NKG2D), and NK receptor signaling adaptor molecules *HCST* (DAP10) and *TYROBP* (DAP12), corresponding to a NK cell designation (Denyer et al., 2006; Piriou-Guzylack and Salmon, 2008; Gerner et al., 2009; Toka et al., 2009). γδ T-cells were identified by *TRDC* expression, encoding the γδTCR δ chain, and were subdivided into two major subtypes based on presence/absence of *CD2* expression (Piriou-Guzylack and Salmon, 2008; Gerner et al., 2009; Stepanova and Sinkora, 2013; Sedlak et al., 2014). Clusters 6 and 21 were identified as $CD2^-$ γδ T-cells and clusters 24 and 31 as $CD2^+$ γδ T-cells. Clusters expressing *CD3E* but not *TRDC* were considered αβ T-cells and were further subdivided based on *CD4* expression (0, 3, 4, 28 classified as $CD4^+$ αβ T-cells) or *CD8A* and *CD8B* expression (9, 12, 14, 18, 22 classified as $CD8\alpha\beta^+$ αβ T-cells) (Piriou-Guzylack and Salmon, 2008; Gerner et al., 2009). Clusters 5 and 17 were more difficult to fully classify and likely represented a mixture of cells, with some but not all cells expressing *CD3E*. Cells in clusters 5 and 17 largely lacked expression of *CD5*, *CD6, TRDC, CD4,* and *CD8B* but did largely express *CD2, CD8A*, *KLRB1*, and *KLRK1* and were therefore characterized as a mixture of $CD8\alpha^+$ αβ T- and NK cells.

Cells in cluster 34 could not be characterized well enough to broadly classify as myeloid, B, T, or NK lineage leukocytes based on the porcine cell markers described and remained unclassified. Cluster 35 expressed *HBM* and *AHSP*, indicating erythrocytes. Clusters 34 and 35 were still included in further scRNA-seq analyses; however, results pertaining to these clusters were not discussed.

Gene Signatures of BulkRNA-Seq Populations Had Limitations in Resolving Single-Cell Identities

Gene set enrichment analyses (GSEA) using SEG lists defined at different levels of enrichment for each sorted bulkRNA-seq population (**Supplementary File 3**, see Materials and Methods) was performed to identify which scRNA-seq clusters were likely represented (**Figures 5A,B**, **Supplementary Figure 10A**, and **Supplementary File 8**). Some gene sets had high relative enrichment in anticipated corresponding scRNA-seq clusters, such as Myeloid gene sets to monocyte/DC clusters, CD21nB/CD21pB gene sets to B-cell clusters, and SWC6gdT gene sets to $CD2^-$ γδ T-cell clusters. Interestingly, highest relative enrichment (2.51) for the top 1% of CD21nB SEGs was noted for ASCs in cluster 29, followed by erythrocytes in cluster 35 (1.68). Within sorted NK and T-cell populations, some gene sets showed high relative enrichment for their anticipated corresponding clusters in the scRNA-seq dataset. We also noted off-target relative enrichment for gene sets in clusters not anticipated to be included in specific sorted cell populations. Cluster 28 had lower relative enrichment for CD4T and CD4CD8T SEG lists at top 5–25% SEG levels (−0.02 to 0.73) than did several non-$CD4^+$ αβ T-cell clusters. Similar phenomena were observed for CD8T top 5–25% SEG lists, whereby clusters 1, 24, and 31 had higher relative enrichment for CD8T SEG lists (0.69 to 1.56) than did clusters 14 or 18 (−0.04 to 0.95 relative enrichment) that were anticipated to be included in the CD8T population. Clusters 24 and/or 31 showed off-target relative enrichment for all T/NK gene sets to various degrees, though these cells would not be expected to make up a sizeable portion of any of those sorted cell populations.

Further comparison of porcine bulk and scRNA-seq data by CIBERSORTx deconvolution analysis largely supported our single-cell cluster designations by predicting which clusters proportionally represented the bulk RNA-seq data (**Supplementary Figure 10B** and **Supplementary File 7**). Several clusters with poor AUCell enrichment for anticipated bulkRNA-seq gene sets in **Figures 5A,B**, such as cluster 28, were predicted to constitute considerable proportions of their anticipated cell populations by CIBERSORTx deconvolution analysis. Additionally, clusters that demonstrated off-target enrichment by AUCell analysis, such as clusters 1, 9, 22, 24, and 31, were not predicted to be largely present in those off-target populations using CIBERSORTx. However, CIBERSORTx failed to predict many single-cell clusters to have notable abundances in any bulkRNA-seq populations, such as clusters 8, 19, 26, 32, and 34 having <3.33% predicted abundance for any one bulkRNA-seq sample.

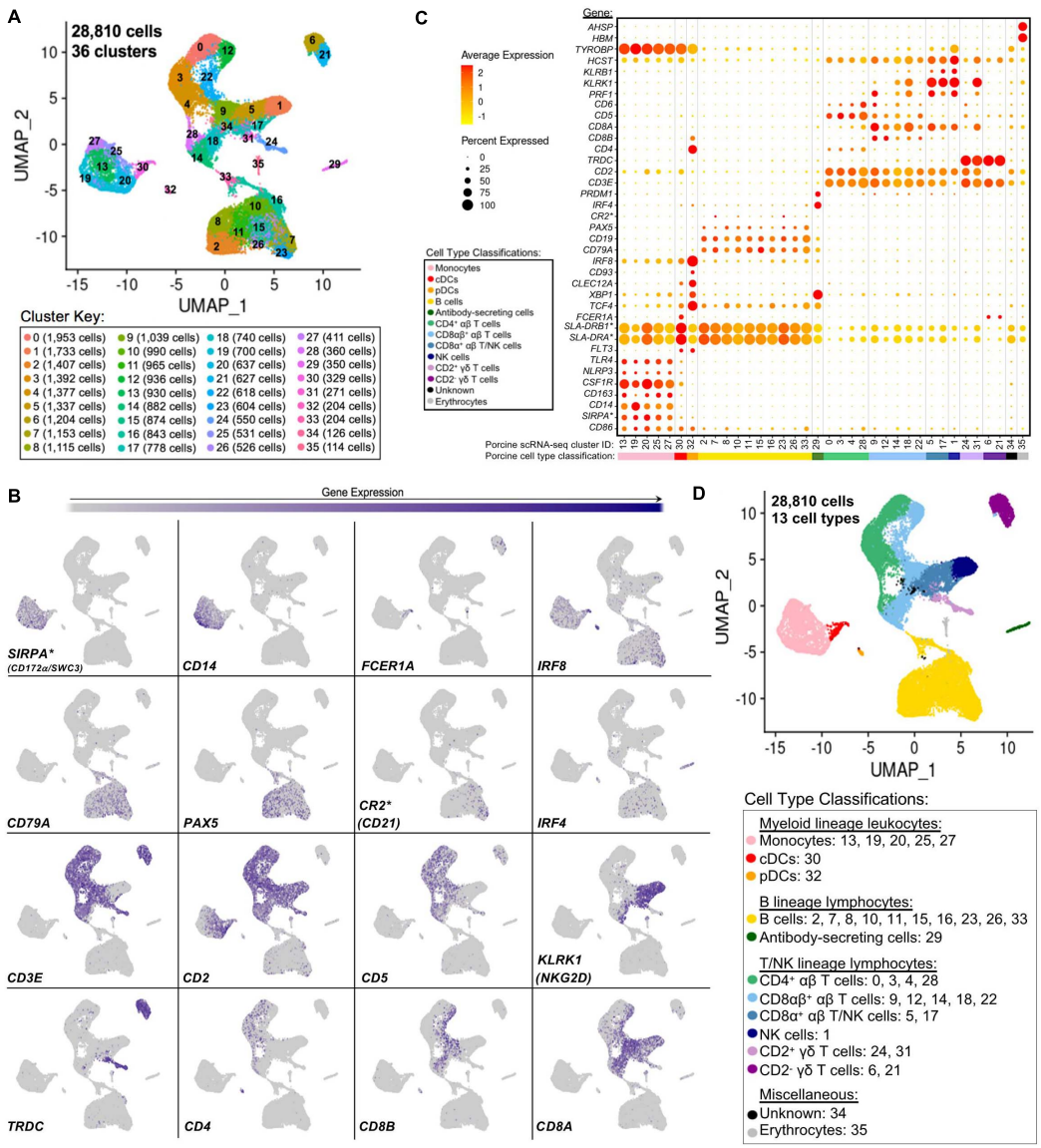


FIGURE 4 | Classification of porcine PBMC scRNA-seq clusters based on known cell type-specific gene expression. **(A)** Two-dimensional UMAP visualization of 28,810 single cells from porcine PBMCs classified into 36 designated clusters. Each point represents a single cell. Color of the point corresponds to transcriptional cluster a cell belongs to. Cells more transcriptionally similar to each other belong to the same cluster. **(B)** Visualization of selected cell type-specific gene expression overlaid onto two-dimensional UMAP coordinates of single cells. Each point represents a single cell. Color of the point corresponds to relative expression of a specified gene (bottom left of each UMAP plot) within a cell. Gray corresponds to little/no gene expression, while navy corresponds to increased gene expression. **(C)** Dot Plot visualization of selected cell type-specific gene expression for each single-cell cluster shown in A. Clusters are listed on the *x*-axis, while selected genes are listed on the *y*-axis. The size of a dot corresponds to the percent of cells in a cluster that expressed the gene. The color of a dot corresponds to the average relative expression level for the gene in the cells expressing the gene within a cluster. Color bar below the *x*-axis corresponds to porcine cell type each cluster was classified as. **(D)** Two-dimensional UMAP visualization of single cells from porcine PBMCs classified into major porcine cell types. Each point represents a single cell. Color of the cell corresponds to porcine cell type the respective cluster was designated as based on gene expression patterns for the cluster it belonged to in **(C)**. Seven PBMC samples used for scRNA-seq analysis were derived from each of three separate experiments (experiment B, $n = 2$; experiment C, $n = 3$; experiment D, $n = 2$). Between 3,042 and 6,518 cells were derived from each PBMC sample. *Refer to 'Gene name replacement' methods.

Additional GSEA comparing gene sets derived from public bulkRNA-seq data of sorted human PBMC populations with porcine single-cell gene expression profiles informed cluster identity as it relates to human immune cells (**Figure 5C**, **Supplementary Figures 10C,D**, and **Supplementary File 9**). High relative enrichment for human monocyte gene sets in porcine monocyte populations, human CD123PmDC gene sets in porcine cDCs, and human pDC gene sets in porcine pDCs was observed, in general consensus with gene expression profiles of anticipated corresponding porcine single-cell clusters. NaiveB

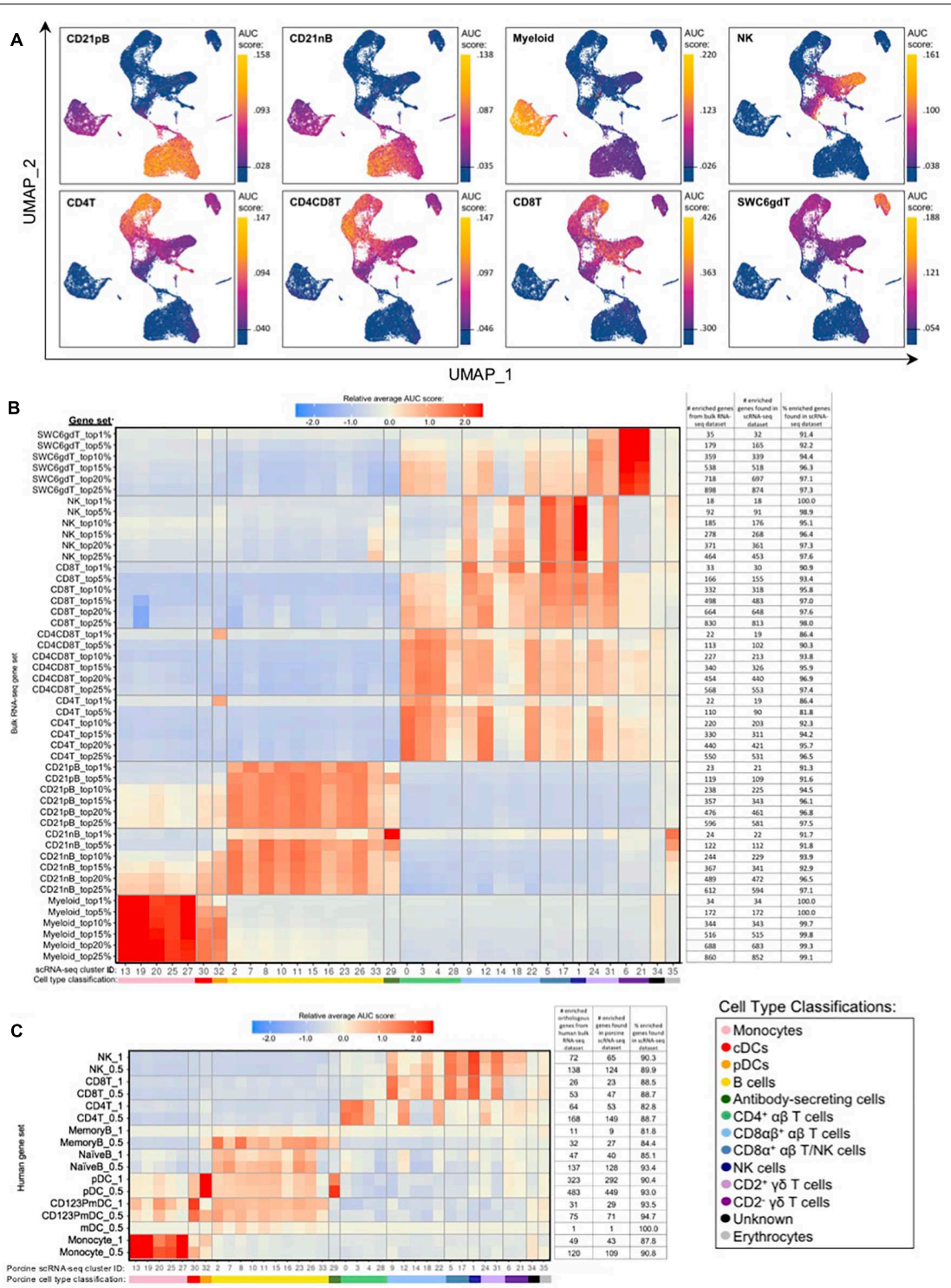


FIGURE 5 | Enrichment of gene signatures from bulkRNA-seq in porcine single-cell clusters. **(A)** Gene set enrichment scores calculated by AUCell analysis of enriched gene sets from the top 25% of SEGs in pig bulkRNA-seq sorted populations overlaid onto cells of the porcine scRNA-seq dataset visualized in two-dimensional UMAP plot. Each point represents a single cell. The color of the point corresponds to the AUC score calculated for each respective cell. Higher AUC scores correspond to a greater percentage of cells from a gene set being detected in the top 5% of expressed genes in a cell. A threshold for AUC score detection within each gene set was set as shown in **Supplementary Figure 10A** and is indicated by a horizontal line on the gradient fill scale for each plot. **(B)** Relative average gene set enrichment scores of scRNA-seq clusters calculated by AUCell analysis of enriched gene sets from porcine bulkRNA-seq sorted data. Scores are relative to other cells within a single gene set comparison (across a row of the heatmap) and are not calculated relative to scores across different gene

(Continued)

FIGURE 5 | Continued
sets (across columns in the heatmap). Gene sets were created from the top 1, 5, 10, 15, 20, or 25% of SEGs from sorted populations, as determined by highest log2FC values in the porcine bulkRNA-seq data. The number of genes included from the bulkRNA-seq dataset and the number and percent of genes detected in the scRNA-seq dataset is listed on the right of the heatmap. A color bar under scRNA-seq cluster IDs indicates the cell type classification, as according to **Figure 4D**. **(C)** Relative average gene set enrichment scores of scRNA-seq clusters calculated by AUCell analysis of enriched gene sets from human bulkRNA-seq sorted data. Scores are relative to other cells within a single gene set comparison (across a row of the heatmap) and are not calculated relative to scores across different gene sets (across columns in the heatmap). Gene sets were created from genes with high expression scores >0.5 or >1 for each respective sorted population of cells, with a greater high expression score indicating greater enrichment. The number of genes included from the bulkRNA-seq dataset and the number and percent of genes detected in the scRNA-seq dataset is listed on the right of the heatmap. A color bar under scRNA-seq cluster IDs indicates the cell type classification, as according to **Figure 4D**.

cell gene signatures had positive relative enrichment in all porcine B-cell clusters except cluster 33 at both the 0.5 and 1.0 resolution level, while the MemoryB cell signature had highest relative enrichment scores for B and ASC clusters at the 0.5 level, with little relative enrichment at the 1.0 level (likely due to a limited number of genes in the gene set). Human T/NK gene sets had off-target enrichment very similar to patterns observed in GSEA with porcine gene sets. Overall, GSEA between human bulkRNA-seq gene signatures and gene expression profiles of porcine scRNA-seq data supported many of the same findings when comparing between porcine bulkRNA-seq gene sets and gene expression profiles of porcine scRNA-seq data. Results indicated limitations of gene profiles obtained from sorted bulkRNA-seq populations in accurately describing/accounting for transcriptional heterogeneity resolved by scRNA-seq.

Integration of Porcine and Human scRNA-seq Datasets to Further Annotate Porcine Cells

We examined porcine single-cell identities by comparing the porcine scRNA-seq data to a highly annotated scRNA-seq dataset of human PBMCs, providing a higher level of resolution than available with bulkRNA-seq. Transfer of more highly specified human cell type labels onto porcine cells could reveal the most likely human counterparts for these porcine populations. Mapping scores were further calculated to determine how well porcine cells were truly represented by the human dataset (**Figure 6A**, **Supplementary Figures 11A,B**, and **Supplementary File 10**).

Many porcine clusters had >95% of cells mapping to a specific human cell type, with average mapping scores >0.9, including monocyte, pDC, cDC, and ASC clusters, suggesting high congruency between pig and human for these cell types (**Figure 6B**). All porcine B-cell clusters, omitting cluster 33, mapped primarily to human B-cell clusters, but average mapping scores were slightly lower (0.80–0.87), indicating less ideal representation in the human data. In addition, every porcine B-cell cluster had overlap with all three human B-cell types (**Figure 6A**). Of the porcine $CD4^+$ αβ T-cells, most cluster 0 cells were predicted as human CD4 naïve cells, clusters 3 and 4 cells as human CD4 T central memory (TCM) cells, and cluster 28 cells as human CD4 proliferating cells. From porcine $CD8\alpha\beta^+$ αβ T-cells, clusters 14 and 18 were largely assigned as human replicating cell types, while 90% of cluster 9 cells were predicted as human CD8 T effector memory (TEM) cells. Highest cluster 12 predictions were mainly to human CD4/CD8 naïve T-cells, and cluster 22 cells predicted to match a range of human cell populations, with the largest percentage predicted as human CD8 TEMs. Porcine $CD8\alpha^+$ αβ T/NK and NK clusters had predictions split primarily across human CD8 TEM and NK designations. Porcine $CD2^+$ γδ T-cell clusters 24 and 31 had 74 and 98%, respectively, of cells predicted as human CD8 TEM, NK, or γδ T-cells. Porcine $CD2^-$ γδ T-cell clusters 6 and 21 had the majority of cells predicted as human CD4 TCM, innate lymphoid cell (ILC), or γδ T-cells, though the average mapping scores were lower for those assigned as CD4 TCM (0.73–0.74) or gdT (0.74–0.78) than those assigned as ILCs (0.82–0.83) (**Supplementary File 10**). Overall, cross-species comparison to a well-annotated human scRNA-seq dataset helped elucidate porcine cell type identities at a higher resolution than porcine or human bulkRNA-seq datasets (**Figure 5**), though some discordance was clearly still present.

Several porcine clusters had low mapping scores to a human cell type, indicating the porcine cells may not be well represented by the human reference dataset (**Figure 6B** and **Supplementary File 10**). Therefore, *de novo* visualization was performed on the combined human and porcine data, to identify cells in the pig dataset not well represented in the human data (**Figures 6C,D**). Porcine clusters could be identified that had low similarity to human cells, and vice versa (**Figure 6C**). Specifically, porcine clusters 6, 16, 21, and 33 weakly overlapped human cells in the two-dimensional *de novo* visualization (compare **6C** and **6D**) and had lower average mapping scores to any human cell type (**Figure 6B**). Further inspection revealed clusters 6 and 21 to be $CD2^-$ γδ T-cells (identified in **Figures 4B,D**) and their limited representation in human dataset is discussed further below. In contrast, clusters 16 and 33 were B-cells, and to further understand their limited representation compared to other porcine B-cells, clusters 16 and 33 were compared by pairwise comparisons to all remaining B-cell clusters (2, 7, 8, 10, 11, 15, 23, 26; **Supplementary File 6**). Pairwise comparisons revealed significantly increased expression of 33 genes in cluster 16 and 282 genes in cluster 33 relative to every other B-cell cluster (**Table 4**). Compared to other porcine B-cell clusters, cluster 16 had significantly greater expression of several genes associated with B-cell activation (such as *BHL, ITGB7, JCHAIN, ZBTB38*) (Castro and Flajnik, 2014; Kreslavsky et al., 2017; Delecluse et al., 2019; Wong and Bhattacharya, 2020), while many genes with significantly increased expression in cluster 33 were associated with cellular replication and/or division, such as *HIST1H2AB,*

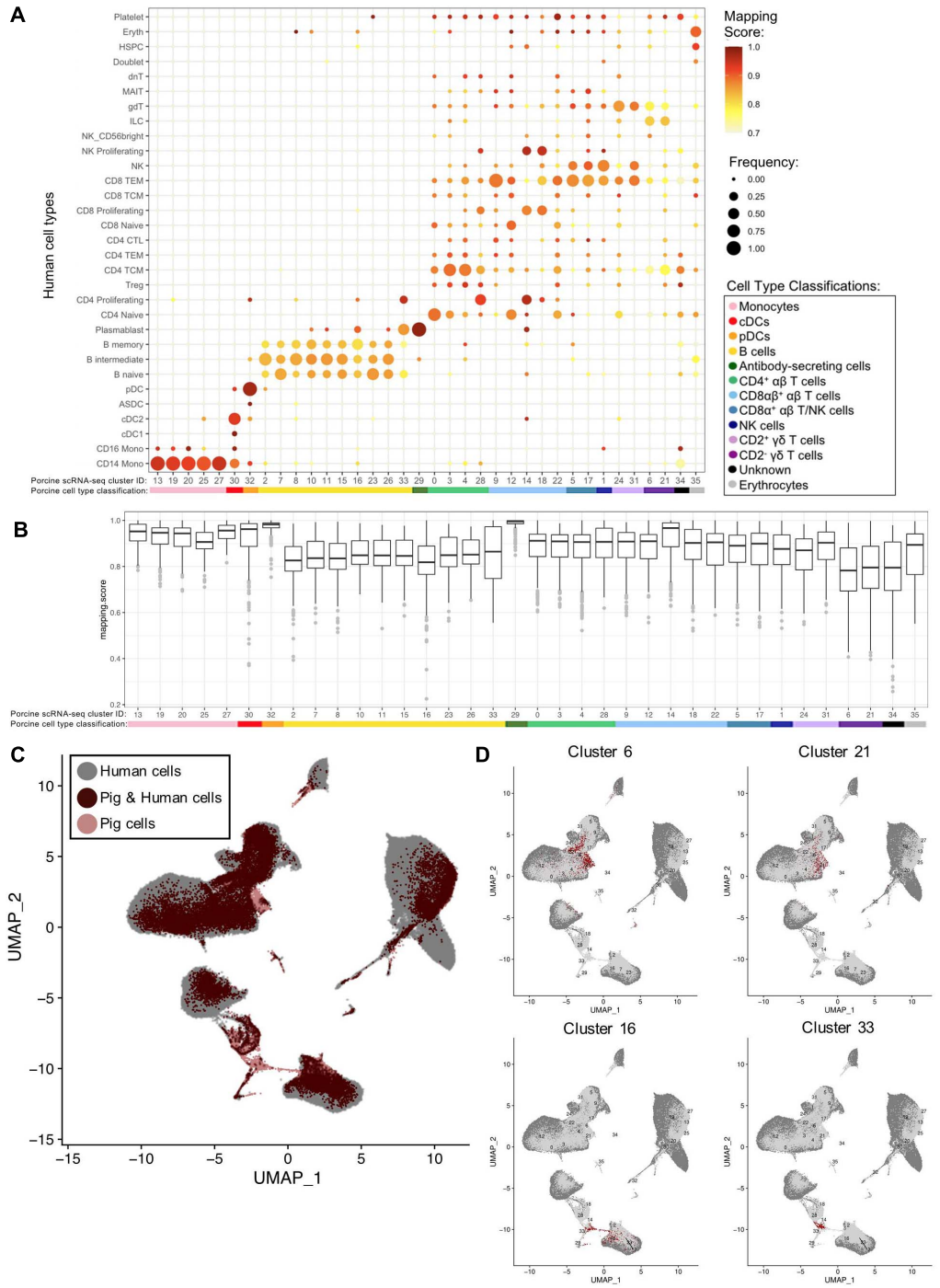


FIGURE 6 | Integration of porcine and human scRNA-seq datasets to further annotate porcine cells. **(A)** Mapping scores calculated to determine how well porcine cells were represented by the human dataset. The human cell type specific frequency (size of the circle) and mapping score for that human cell type (color) are shown for each porcine scRNA-seq cluster. Porcine cell type classifications (color) are shown below the porcine scRNA-seq cluster IDs. **(B)** Mapping scores calculated to determine how well porcine cells were represented by the human dataset. The mapping scores for each porcine scRNA-seq cluster is represented by a box and whiskers plot. Porcine cell type classifications (color) are shown below the porcine scRNA-seq cluster IDs. **(C)** To identify cells in the porcine dataset that were not well represented in the human dataset, a *de novo* visualization of the merged porcine and human data was performed. The porcine (pink) and human (gray) were plotted together using UMAP. An overlap of both porcine and human cells is shown as (dark red). Clusters of porcine cells that are not well represented in the human data can be observed by pink regions in the plot. **(D)** Two primary regions of porcine cells that were not well represented in the human data were identified in **(C)**. In order to clarify which porcine scRNA-seq clusters were represented in these regions, the porcine cluster IDs were projected onto the UMAP and cells from four clusters overlapping the identified regions were colored as dark red.

TABLE 4 | Genes with significantly increased expression in cluster 16 or 33 relative to every other B-cell cluster (2, 7, 8, 10, 11, 15, 23, 26) by every pairwise differential gene expression analysis. Underlined genes had significantly increased expression in both cluster 16 and 33.

Cluster with greater gene expression relative to all other B-cell clusters	Genes
Cluster 16	*ABRACL*, ACTG1, ANXA2, BHLHE41, BIRC5, CA8, CD52, CDK2AP2, CFD, CLIC1*, DUT, DYNLL1, ENSSSCG00000037141, ENSSSCG00000039490, GAPDH, H2AFZ, HINT1, HIST1H2AB, HMGB1*, HMGB2, HMGN2, ITGB7, JCHAIN, LSM5*, PCLAF, PCNA, S100A11, SPCS1, SRSF10, STMN1, TUBB, TYMS, ZBTB38*
Cluster 33	*ACTG1, AHCY, ANP32E*, ANXA2, AP3S1*, ARL6IP1, ASF1B, ASPM, ATAD2, ATAD5, ATP5MC3, AURKA*, AURKB, BANF1*, BBS7, BIRC5, BRCA1, BUB1, BUB1B*, BUB3, CALM3, CBX3*, CBX5, CCDC167, CCDC34, CCNB1, CCNB2, CCNE2, CDC2*, CDC20, CDC25B, CDC45, CDCA3, CDCA5, CDCA7, CDCA8, CDK2, CDK4, CDKN2C, CDKN3, CDT1, CENPA, CENPE, CENPF, CENPM, CENPN, CENPS, CENPT, CENPU, CENPW, CEP55, CEP57*, CIP2A, CKAP2, CKAP2L, CKAP5, CKS1B, CKS2, CLIC1*, CLSPN, CMC2*, COX17, COX5A, CSPP1, CTCF, CXXC1, CYB5B, DBF4*, DDX39A, DEK*, DEPDF1B, DHFR*, DIAPH3, DLGAP5, DNMT1, DTYMK, DUT, DYNLL1, E2F2, E2F8, ECT2, ENSSSCG00000034527, ENSSSCG00000037071, ENSSSCG00000037185, ERH, ESCO2, E2H2, FABP3, FAM72A*, FBXO5, GAPDH, GFER*, GGCT*, GINS1*, GINS2, GMNN, GON7, GPN3, GTSE1*, H1-2*, H2AFV, HSAFY, H2AFZ, HDGF, HELLS, HINT1, HIST1H1D, HIST1H1E, HIST1H2AB, HIS1H2AG, HIST2H2AC, HMGA1, HMGB1*, HMGB2, HMGB3*, HMGN2, HMGN1*, HMGN5*, HMMR, HNRNPA2B1*, HNRNPAB*, HNRNPS, HNRNPH1, HSP90AA1, HSPA4L, IDI1*, JPT1*, KIF11, KIF15, KIF18A, KIF20A, KIF20B, KIF22, KIF23*, KIF2C, KIF4A, KIFC1, KNL1*, KNSTRN, KPNA2, KPNB1, LGALS1, LMNB1, LSM2, LSM3, LSM4*, LSM5*, LSM6, LSM8, LYAR, MAD2LA, MAGOHB*, MAZ, MCM3, MCM4, MCM5, MCM6, MCM7, MDH1, MELK, MIS18A, MKI67*, MND1, MNS1, MPHOSPH9, MSH6, HTFR1, MTHFD1, MXD3, MYBL2, NANS, NASP, NCAPD2, NCAPD3, NCAPG, NCAPG2, NCAPH, NCAPH2, NDC80, NEK2, NRM, NSD2, NT5C, NUCKS1, NUDC, NUF2, NUSAP1*, NUTF2*, NXT2, ORC1, ORC6, PBK, PCLAF, PCNA, POC1A, POLR2K, POMP, POP7, PPIA*, PRIM1*, PRR11, PTBP1, PTMA, PTTG1*, RACGAP1, RAD21, RAD51, RAD51AP1, RAN, RANBP1, RBMX, RFC3, RFC4, RHNO1, RNASEH2B, RPA2, RPA3, RRM2*, S100A11, S100A6, SEPHS1*, SFPQ, SGO1, SGO2, SHCBP1, SIVA1, SKA1*, SKA2, SLBP, SMC1A, SMC2, SMC3, SMC4, SNRPA1, SNRPD1, SNRPD3, SNRPE, SNRPF, SNRPG*, SPC24, SPC25, SPTSSA, SQLE, SRSF10, SRSF1, SRSF7, STMN1, SUZ12, SYNE2, TACC3*, TCF19, TEX30, TK1, TMEM258, TMPO, TMSB15B*, TOP2A, TOPBP1, TPX2, TRA2B, TRIM28, TRIM59, TTK, TUBA1B, TUBA1C*, TUBB, TUBB4B, TYMS, UBALD2, UBE2C, UBE2S, UBE2T, UHRF1, USP1, UXT, VIM VRK1, WEE1*, YBX1*, YEATS4, YWHAQ, ZNF367*

**Refer to gene name replacement in Materials and Methods section.*

HMGB2, STMN1, MKI67, PCLAF, UBE2C (Dabydeen et al., 2019; Giotti et al., 2019).

Different Activation States of Porcine CD4$^+$ αβ T-Cells Based on CD8α Expression

We further compared scRNA-seq gene expression profiles amongst only CD4$^+$ αβ T-cell clusters to gain functional inferences and correspondence to CD8α$^-$ vs. CD8α$^+$ phenotypes that were used to sort CD4$^+$ αβ T-cells for bulkRNA-seq. CD4$^+$ αβ T-cell clusters (0, 3, 4, 28) were comprised of 5,082 total cells (**Figure 7A**). Hierarchical clustering and pairwise DGE (**Supplementary File 7**), as well as random forest (RF) analyses, a deep-learning classification method, (see Methods; **Supplementary File 11**), cumulatively revealed clusters 3 and 4 to be the most transcriptionally similar to each other. Clusters 3 and 4 had the smallest hierarchical distance, fewest DEGs (67), and largest RF error rate (19.5) between them, while cluster 28 was the most distantly related to the other 3 clusters (**Figure 7B**).

CD8A gene expression was detected in a subset of cells in the CD4$^+$ αβ T-cell clusters (3.5, 13.1, 20.9, 39.7% of cells in clusters 0, 3, 4, 28, respectively; **Figure 7C**). *CD8A* expression was significantly greater in clusters 4 and 28 compared to cluster 0 by pairwise DGE analyses (**Supplementary File 7**) but not in cluster 3 compared to 0, due to not meeting a minimum threshold of cells (20%) expressing the gene in either cluster implemented for DGE analysis. However, cluster 3 had significantly greater expression of *CD8A* compared to cluster 0 when removing the minimum cell expression threshold (average log2FC = 0.37, adjusted p-value = 5.52×10^{-21}). GSEA of DEGs identified by pairwise DGE analysis of CD4T and CD4CD8T populations recovered from bulkRNA-seq (**Supplementary File 2**) revealed genes significantly enriched in CD4T compared to CD4CD8T populations were relatively enriched in cluster 0, while genes significantly enriched in CD4CD8T compared to CD4T populations showed greater relative enrichment in clusters 4 and 28 and to a lesser extent in cluster 3 (**Figure 7D** and **Supplementary File 12**).

The top genes contributing to overall transcriptional heterogeneity amongst four clusters of CD4$^+$ αβ T-cells, as determined by RF analysis (**Figures 7E,F** and **Supplementary File 13**), highly overlapped with genes identified in overall DGE analysis (**Figure 7G** and **Supplementary File 13**). Of eight genes with mutually highest permutation and impurity scores from overall RF analysis (**Figures 7E,F**), one gene had significantly greater expression in cluster 0 compared to all other clusters (*RPS3A*), while the other seven genes had significantly greater expression in clusters 3, 4, and 28 compared to cluster 0 (*FCGR3A*, TMSB10, COX1, S100A6, GPX1, CRIP1*, S100A11*), as determined by pairwise DGE analyses (**Supplementary File 7**).

Genes associated with a naïve phenotype, including *CCR7, SELL, LEF1,* and *TCF7* (Szabo et al., 2019; Kim et al., 2020) had significantly increased expression in cluster 0 (**Figure 7G** and **Supplementary Files 9, 13**), in line with the result obtained by comparing to human scRNA-seq data that indicated a good alignment of cluster 0 with human naïve CD4 T-cells (**Figure 6A**). From **Figure 6A**, clusters 3 and 4 aligned with human CD4 Tcm (central memory) cells, and cluster 28 aligned with human CD4 proliferating cells. Correspondingly, genes associated with

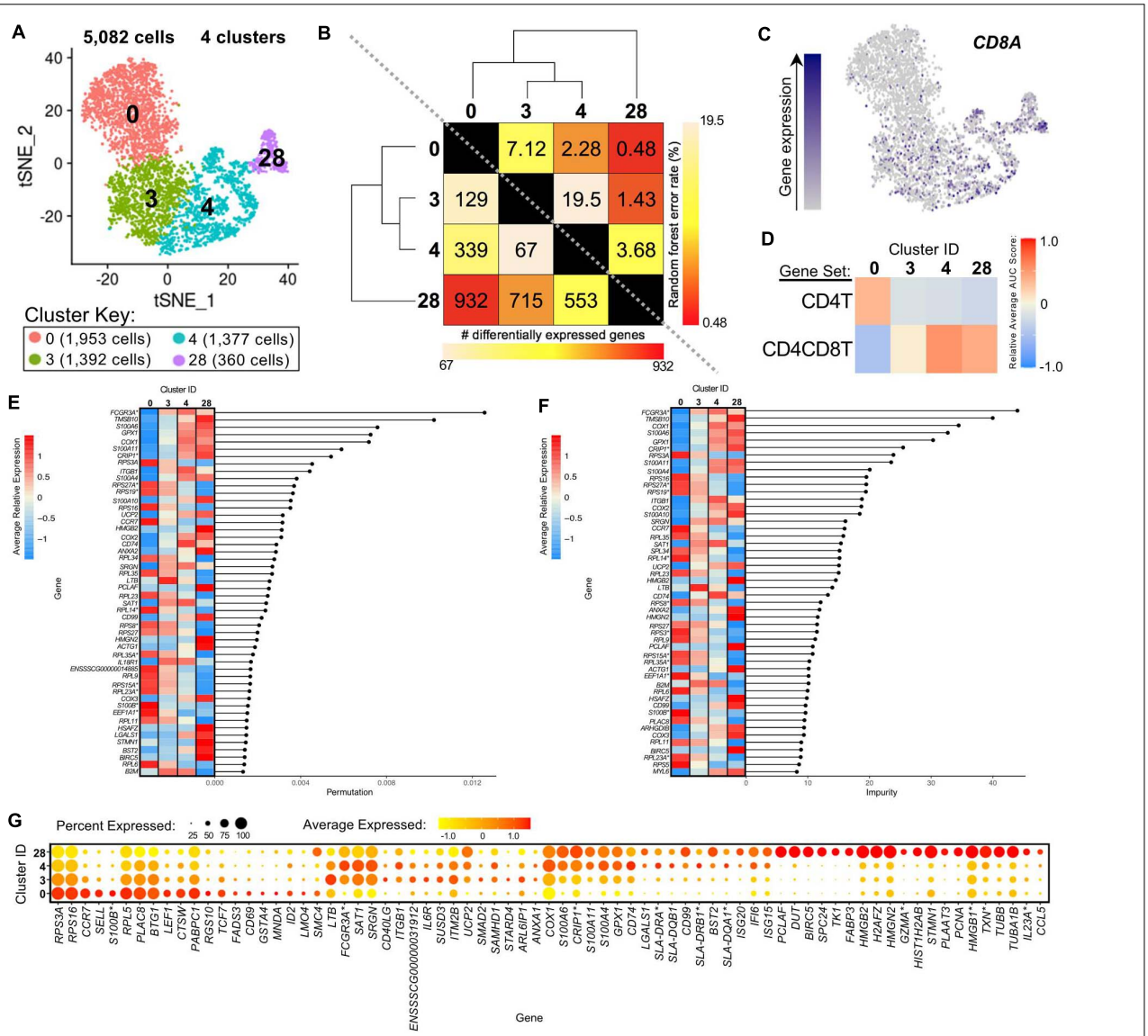


FIGURE 7 | Transcriptional heterogeneity of porcine CD4$^+$ αβ T-cells at single-cell resolution. **(A)** Two-dimensional t-SNE plot of 5,082 cells belonging to clusters designated as CD4$^+$ αβ T-cells (clusters 0, 3, 4, and 28) in **Figure 4D**. Each point represents a single cell. Color of the cell corresponds to transcriptional cluster a cell belongs to. Cells more transcriptionally similar to each other belong to the same cluster. **(B)** Transcriptomic relationship amongst CD4$^+$ αβ T-cell clusters as calculated by three methods: hierarchical clustering (as seen by hierarchical trees on both axes), pairwise random forest analyses (as seen on top right diagonal); and pairwise DGE analyses (as seen on bottom left diagonal). Longer branches on the hierarchical tree corresponds to greater hierarchical distance. Lower numbers of DEGs by DGE analysis and higher out-of-bag (OOB) error rates from random forest analyses indicate greater pairwise transcriptional similarity. **(C)** Visualization of *CD8A* expression overlaid onto t-SNE coordinates of single CD4$^+$ αβ T-cells. Each point represents a single cell. Color of the point corresponds to relative expression of CD8A within a cell. Gray corresponds to little/no gene expression, while navy corresponds to increased gene expression. **(D)** Relative average gene set enrichment scores of CD4$^+$ αβ T-cell clusters calculated by AUCell analysis of DEG sets from pairwise DGE analysis of the CD4T and CD4CD8T populations from porcine bulkRNA-seq. Scores are relative to other cells within a single gene set comparison (across a row of the heatmap) and are not calculated relative to scores across gene set (across columns in the heatmap). **(E,F)** Genes with the largest effects in discriminating CD4$^+$ αβ T-cells by cluster identities were determined, as indicated by high permutation **(E)** and/or impurity scores **(F)** calculated from a trained random forest model. Average relative expression for each of these genes within clusters is also depicted by a heatmap. **(G)** Dot plot of up to the top 20 DEGs having logFC > 0 from overall DGE analysis of only CD4 + ab T-cell clusters. Clusters are listed on the *y*-axis, while selected DEGs are listed on the *x*-axis. The size of a dot corresponds to the percent of cells in a cluster that expressed the gene. The color of a dot corresponds to the average relative expression level for the gene in the cells expressing the gene within a cluster. *Refer to 'Gene name replacement' methods.

activation, such as *ITGB1, CD40LG, IL6R,* and MHC II-associated genes (*CD74, SLA-DRA, SLA-DQB1, SLA-DRB1*, SLA-DQA1**) (Grewal and Flavell, 1996; Gerner et al., 2009; Zemmour et al., 2018; Zhu et al., 2020) had significantly greater expression in clusters 3, 4, and/or 28, and cluster 28 expressed many genes specific for cellular replication and division (*PCLAF, BIRC5, TK1, PCNA*) (Dabydeen et al., 2019; Giotti et al., 2019; **Figure 7G** and **Supplementary Files 9, 13**). Overall, we leveraged

single-cell gene expression profiles to confirm likely identity of cluster 0 as naïve $CD4^+CD8\alpha^-$ αβ T-cells and clusters 3, 4, and 28 as potentially previously activated $CD4^+CD8\alpha^+$ αβ T-cells.

Heterogeneity Between/Amongst $CD2^+$ and $CD2^-$ γδ T-Cells

Clusters predicted to be porcine γδ T-cells were examined to reveal transcriptional distinctions within this cell type. Four clusters containing 2,652 cells were previously identified as $CD2^-$ γδ T-cells (clusters 6, 21) or $CD2^+$ γδ T-cells (clusters 24, 31) (**Figure 8A**). We could further segregate these clusters by *CD2* and *CD8A* expression into $CD2^-CD8\alpha^-$ (clusters 6, 21), $CD2^+CD8\alpha^-$ (cluster 24), and $CD2^+CD8\alpha^+$ (cluster 31) designations used to functionally define porcine γδ T-cells previously (Stepanova and Sinkora, 2013; Sedlak et al., 2014; **Figure 8B**).

$CD2^-$ γδ T-cell clusters 6 and 21 were most closely related to one another by hierarchical clustering, had the fewest pairwise DEGs (30), and had the highest pairwise RF analysis error rate (23.5), indicating clusters 6 and 21 to be the most transcriptionally similar γδ T-cell clusters of the four clusters (**Figure 8C** and **Supplementary Files 7, 14**). $CD2^+$ γδ T-cell clusters 24 and 31 were most similar to each other by hierarchical clustering, had the second fewest pairwise DEGs (236), and had the second highest pairwise RF error rate (5.12), indicating clusters 24 and 31 to be most similar to each other. When performing pairwise comparison between any $CD2^-$ and $CD2^+$ γδ T-cell clusters, the number of DEGs increased and RF error rates decreased, indicating greater transcriptional differences between cells of the $CD2^-$ and $CD2^+$ γδ T-cell lineages than amongst them (**Figure 8C** and **Supplementary Files 7, 14**).

The top genes contributing to overall transcriptional heterogeneity amongst γδ T-cell clusters, as determined by RF analysis (**Figures 8D,E** and **Supplementary File 14**), overlapped with genes identified with significant and highest logFC expression in overall DGE analysis (**Figure 8F** and **Supplementary File 14**). Six of the top seven genes with mutual highest impurity (the best features that correctly split the data) and permutation scores from RF analysis (**Figures 8D,E**) were also DEGs between both $CD2^-$ compared to both $CD2^+$ γδ T-cell clusters by pairwise DGE analysis (**Supplementary File 7**), again indicating large transcriptional differences between $CD2^-$ and $CD2^+$ γδ T-cells. In total, 31 genes had significantly greater expression in both $CD2^-$ γδ T-cell clusters compared to both $CD2^+$ γδ T-cell clusters, and 49 genes had significantly greater expression in both $CD2^+$ γδ T-cell clusters compared to both $CD2^-$ γδ T-cell clusters (**Table 5**), as determined using the pairwise DGE analyses (**Supplementary File 7**).

Intra-lineage heterogeneity of $CD2^-$ γδ T-cells (between clusters 6 and 21) and $CD2^+$ γδ T-cells (between clusters 24 and 31) demonstrated additional complexity beyond the inter-lineage heterogeneity between $CD2^-$ and $CD2^+$ γδ T-cells. Pairwise comparison between clusters 24 and 31 (**Supplementary Data 8**) revealed 80 genes with significantly greater expression in cluster 24 ($CD2^+CD8\alpha^-$ γδ T-cells) and 156 genes with significantly greater expression in cluster 31 ($CD2^+CD8\alpha^+$ γδ T-cells). Genes with the greatest logFC expression (logFC > 1.5) in cluster 31 compared to cluster 24 were related to cellular activation and/or effector functions (*CCL5, GNLY, FCGR3A*, KLRK1, GZMA*, NKG7, FCER1G, GZMB*) (Rincon-Orozco et al., 2005; Pizzolato et al., 2019; Szabo et al., 2019). Of the 30 DEGs between clusters 6 and 21 (**Supplementary Data 8**), three genes had significantly greater expression in cluster 6, while 27 genes had significantly greater expression in cluster 21. Several genes with greater expression in cluster 21 encoded for activation- or stress-induced molecules, including *GPX1, LGALS1, ITGB1, LTB,* several genes encoding for S100 proteins (*S100A4, S100A6, S100A10, S100A11*), and genes related to MHCII presentation (*CD74, SLA-DRA**) (Blaser et al., 1998; Ware, 2005; Gerner et al., 2009; Steiner et al., 2011; Kesarwani et al., 2013; Siegers, 2018). Genes encoding transcriptional regulators playing important roles in cell fate determination, including *ID3* and *GATA3,* had greater expression in cluster 6, while *ID2* expression was significantly greater in cluster 21 (Blom et al., 1999; Zhang et al., 2014; Rodríguez-Gómez et al., 2019).[7]

DISCUSSION

We present the first comprehensive annotation of the global transcriptome of all major circulating porcine blood mononuclear cells. We applied bulkRNA-seq to determine transcriptomes of eight sorted PBMC populations and scRNA-seq to annotate transcriptomic diversity of PBMCs into transcriptionally distinct clusters. Deep RNA sequencing detected significant heterogeneity between sorted populations except for T-cell populations, while further heterogeneity was unmasked by scRNA-seq. Collectively, the data sets revealed specific immune functional expression patterns and highlighted substantial diversity in some subsets, such as T-cells. The combined approach helps to unite porcine transcriptomics and cellular immunology, as transcriptional differences and functional relationships of porcine immune cells have remained unclear due to lack of sufficient reagents to label distinct porcine immune cell populations. While cross-species comparisons have been done with many RNA-seq datasets of partially purified cell populations (Kapetanovic et al., 2013; Herrera-Uribe et al., 2020), our new porcine data demonstrates global similarity to human bulkRNA-seq and scRNA-seq transcriptomes that can be used to further unravel porcine cell function and extend comparative immune investigation.

Gene expression patterns from the bulkRNA-seq datasets revealed distinct transcript profiles enriched in biological pathways characteristic of each respective cell population, based on previous findings in pig and other species (Alter et al., 2004; Palmer et al., 2006; Wang et al., 2008; Foissac et al., 2019; Monaco et al., 2019; Summers et al., 2020). However, bulkRNA-seq data from the porcine sorted populations had limited ability to identify genes with specific transcriptional patterns for some sorted lymphocyte populations. The transcriptomes of eight different cell types we provide include three types of transcriptomes that have not reported before in pig, including NK, CD21pB and CD21nB. Lists of SEGs, pairwise DGE between all populations

[7] http://biocc.hrbmu.edu.cn/CellMarker/#

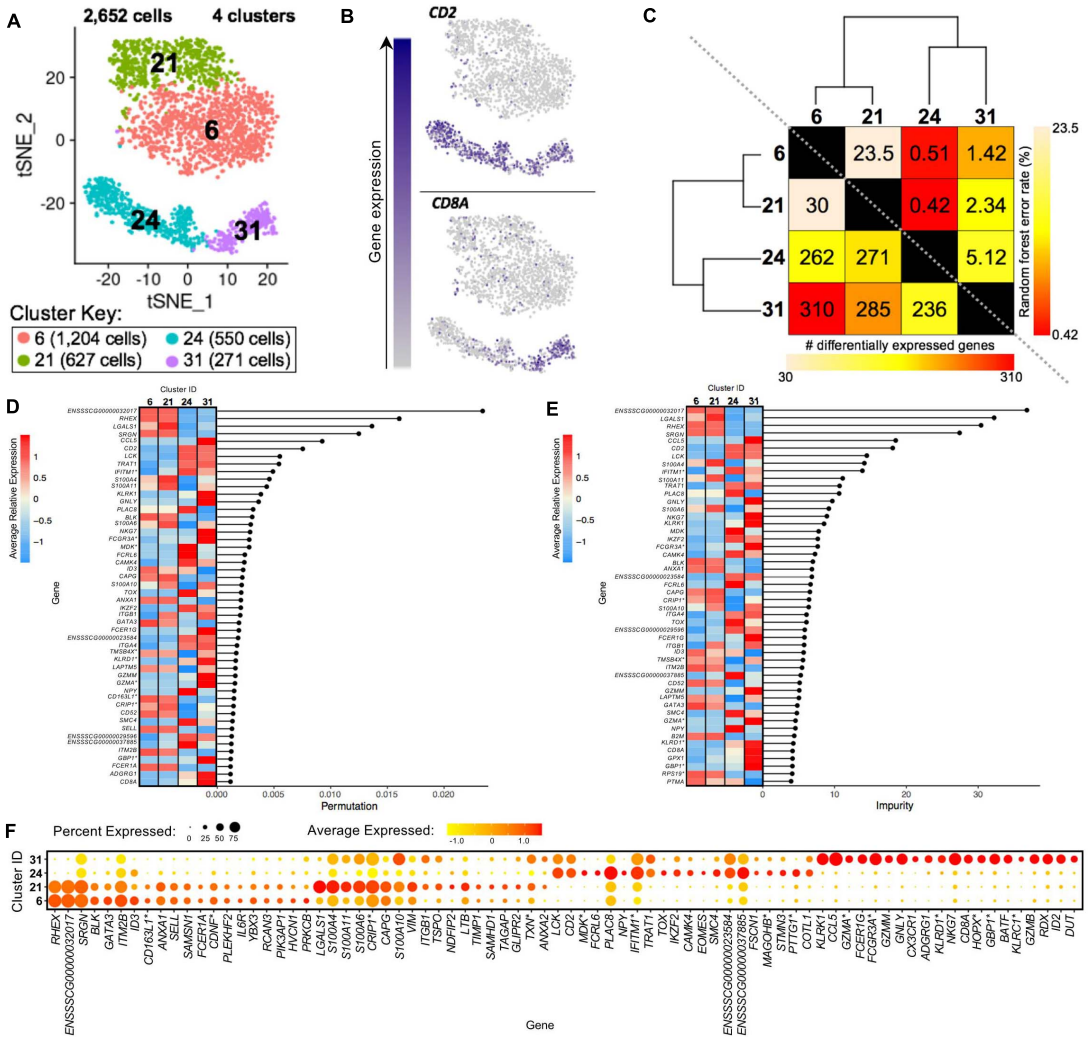


FIGURE 8 | Transcriptional heterogeneity of porcine γδ T-cells at single-cell resolution. **(A)** Two-dimensional t-SNE plot of 2,652 cells belonging to clusters designated as $CD2^{-}$ γδ T-cells (clusters 6, 21) or $CD2^{+}$ γδ T-cells (clusters 24, 31) in **Figure 4D**. Each point represents a single cell. Color of the cell corresponds to transcriptional cluster a cell belongs to. Cells more transcriptionally similar to each other belong to the same cluster. **(B)** Visualization of selected gene expression overlaid onto t-SNE coordinates of single γδ T-cells. Each point represents a single cell. Color of the point corresponds to relative expression of a specified gene (top left of each t-SNE plot) within a cell. Gray corresponds to little/no gene expression, while navy corresponds to increased gene expression. **(C)** Transcriptomic relationship amongst γδ T-cell clusters as calculated by three methods: hierarchical clustering (as seen by hierarchical trees on both axes), pairwise random forest analyses (as seen on top right diagonal); and pairwise DGE analyses (as seen on bottom left diagonal). Longer branches on the hierarchical tree corresponds to greater hierarchical distance. Lower numbers of DEGs by DGE analysis and higher out-of-bag (OOB) error rates from random forest analyses indicate greater pairwise transcriptional similarity. **(D,E)** Genes with the largest effects in discriminating γδ T-cells by cluster identities were determined, as indicated by high permutation **(D)** and/or impurity scores **(E)** calculated from a trained random forest model. Average relative expression for each of these genes within clusters is also depicted by a heatmap. **(F)** Dot plot of up to the top 20 DEGs having logFC > 0 from overall DGE analysis of only γδ T-cell clusters. Clusters are listed on the *y*-axis, while selected DEGs are listed on the *x*-axis. The size of a dot corresponds to the percent of cells in a cluster that expressed the gene. The color of a dot corresponds to the average relative expression level for the gene in the cells expressing the gene within a cluster. *Refer to 'Gene name replacement' methods.

and cell type-specific genes data sets presented here, could be used for further analysis in other pig or even in cross-species comparisons. Notably, we were able to identify a large number of HEGs in the Myeloid population. Some HEGs in Myeloid cells were reported as a Myeloid cell markers in pig (e.g., *CD14* and *CD36*) (Fairbairn et al., 2013) and other HEGs may be considered as new potential cell markers. Also, in comparison to sorted CD4T and CD8T populations reported in a previous porcine RNA-seq study (Foissac et al., 2019), we observed concordant transcriptional patterns in essentially equivalent populations. However, we extended transcriptional annotation to two additional T-cell populations (CD4CD8T, SWC6gdT), thus identifying transcriptional differences across more T-cell populations. We demonstrated the utility of an established NanoString CodeSet (Van Goor et al., 2020; Dong et al., 2021) to validate RNA-seq results and further profile porcine sorted PBMC populations. At the bulk RNAseq level, we concluded substantial transcriptional heterogeneity was present across sorted T-cell and B-cell populations, as fewer enriched or cell type-specific genes were detected. As described below, the lack of

TABLE 5 | Genes differentially expressed between both CD2$^-$ γδ T-cell clusters (clusters 6 and 21) and both CD2$^+$ γδ T-cell clusters (clusters 24 and 31).

Population with greater gene expression	Genes
CD2$^-$ γδ T-cells (clusters 6, 21)	*AP3S1*, ANXA1, BLK, CAPG, CDNF*, CD163L1*, EMP3, ENSSSCG00000032017, ENSSSCG00000033734, FCER1A, GATA3, IL6R, ITM2B, LGALS1, LTB, MAN2B1*, MYL12A*, PARK7, PIK3AP1, PLEKHF2, PPP1CC, RCAN3, RHEX, RPS19*, SAMSN1, SELL, SLC25A24, SRGN, TIMP1, VIM, YBX3*
CD2$^+$ γδ T-cells (clusters 24, 31)	*ABI3*, ACTG1, ARPC1B, ARPC5L, BIN1, CAMK4, CCDC12*, CD2, COTL1, CTSD, DYNLRB1, ENSSSCG00000023584, ENSSSCG00000027196, ENSSSCG00000029596, ENSSSCG00000038825, FAM49B, FSCN1, FYB1, GBP7*, GIMAP4*, H2AFV, IFITM1*, IFI6, IKZF2, IKZF3, IL2RB, ISG15, ITGA4, ITGB2, ITM2C, KRAS, LCK, MAGOHB*, NT5C3A*, PIK3R1, PRKCH*, PSIP1, PTPRC, RESF1, S100A1, SLC9A3R1, SMC4, SNRK, STK17B, STMN3, TRAT1, UBAC2, WCR1, WIPF1**

**Refer to gene name replacement in Materials and Methods section.*

identification of cell type-specific genes was likely caused by the lack of further sub-setting during sorting to separate functionally distinct cells. However, we were able to find several specific transcriptional patterns in B- and T-cells using bulkRNA-seq, and some of the identified genes encode for transmembrane proteins. Beyond further description of well-annotated genes, we also demonstrated that up to 18% of our predicted cell-type specific and enriched genes are currently poorly annotated, i.e., genes with no recognized human ortholog. These data thus increase the functional annotation of these genes, as co-expression patterns linking such genes with known genes can be an important component for Gene Ontology classifications and disease-association gene prediction (van Dam et al., 2018), and is an important proposed outcome of the FAANG project (Giuffra et al., 2019).

Comparison of our sorted population expression patterns to a similar human RNA-seq dataset revealed both similarities and differences between species. While we compared the transcriptomes of the sorted cells with human populations that were isolated using similar cell markers, we cannot exclude that we are biasing this comparison due to different immunoreagent markers used across species. However, we did find similar transcriptional patterns across immune cell populations that are intrinsic to a lineage, such as the porcine Myeloid population correlating with the human myDC123 population, in agreement with other studies (Auray et al., 2016).

Previous global gene expression studies using either porcine whole blood or specific immune cell types have failed to thoroughly describe all major PBMC populations (Freeman et al., 2012; Dawson et al., 2013; Mach et al., 2013; Auray et al., 2016; Foissac et al., 2019). Providing the transcriptomes of bulk sorted cell populations will be readily useful to the majority of porcine immunology research labs that use sorting techniques to analyze porcine immune cell function and RNA expression patterns, as new lists of co-expressed genes in these cell populations are now available. However, our combined analysis of such bulkRNAseq data with the scRNAseq data demonstrated that the former approach has significant heterogeneity, limiting the ability to resolve specific cell types for deeper transcriptional interrogation. A combined analysis provided evidence confirming our hypothesis that scRNA-seq would lead to identification of more specific and novel transcriptional signatures to improve annotation and understanding of circulating porcine immune cells.

Single-cell RNA-sequencing provides many noted benefits in transcriptomic analysis, however there are limitations to the approach. Of benefit, scRNA-seq captured transcriptomes of cells excluded from our bulkRNA-seq analysis, as scRNA-seq approach did not rely on protein marker expression and selection of sorting criteria based on specific marker phenotypes. As mentioned above, scRNA-seq also established that greater levels of cellular heterogeneity exist, since sequencing was resolved to the level of individual cells rather than a sorted population. We recognize the scRNAseq-predicted clusters may contain transitory cell states that may be very challenging to further study for the relationship between cellular function and transcriptional patterns (Bassler et al., 2019). Further, we assumed single-cell gene expression profiles would be indicative of protein expression for cell type-specific markers; however, gene expression for many such markers, including *SIRPA** and *CR2** that encode proteins used for bulk RNA-seq cell sorting, was sparse. Sparsity of data is a known limitation of the scRNA-seq approach utilized herein, while methods such as imputation have been proposed to improve sensitivity (Andrews et al., 2021). We chose not to use imputation due to our current inability to estimate effects on cell patterns through comparison to an external reference (Andrews et al., 2021). Thus, these limitations made it difficult to decipher between low- and non-expression for some genes of interest, including canonical markers used for identifying cell types in the immunology literature. Instead, reliance on gene expression profiles of multiple markers was used. For example, *SIRPA** expression was observed at low levels in monocyte clusters but was virtually absent in DC clusters, though both porcine monocytes and DCs express CD172α protein. Because DCs express CD172α at lower levels than monocytes (Piriou-Guzylack and Salmon, 2008; Auray et al., 2016), *SIRPA** expression in DCs may have been below our limit of detection using scRNA-seq, as it was insufficiently expressed in DCs but not in monocytes. We utilized a droplet-based partitioning method for scRNA-seq that can detect a large number of cells but a lower number of transcripts per cell. By this method, we could retain a large number of cells (>25,000 cells from seven samples) at the expense of limited sequencing depth per cell (minimum of 500 unique genes and 1,000 unique transcripts per cell). Utilizing higher sequencing depth per cell or different partitioning platforms for scRNA-seq that have more efficient transcript capture per cell will be beneficial for deeper analysis of specific cells/genes of interest. It is likely some gene expression profiles are not predictive of protein expression, due to post-transcriptional regulation mechanisms. Using newly available co-expression lists to formulate more refined cell sorting regimens and scRNAseq analysis of such sorted populations will also increase the ability

to define transcriptomes of such cell types (Nestorowa et al., 2016). It was notable that the lists of genes predicted to be significantly enriched in the 36 scRNAseq clusters had overall a very similar fraction of poorly annotated genes (average of 18%; cite in **Supplementary File 6**) to those predicted for bulkRNAseq, indicating that even the genes with expression patterns predicted to be more discriminatory contribute a similar level of genome annotation improvement.

We used multiple methods to compare these high-dimensional expression datasets to further interpret genes predicted to be different between sorted cell populations, between clusters, or between human and pig. GSEA and/or deconvolution analyses of bulkRNA-seq to scRNA-seq datasets was only partially effective in correlating sorted populations with assumed corresponding clusters in the scRNA-seq dataset (regardless of inter-species or intra-species comparison). At a higher level of resolution, both methods were able to assign most corresponding cell-type designations between scRNA-seq and bulkRNA-seq data. However, several different scRNA-seq clusters were not predicted to make up a large portion of any bulkRNA-seq sample. While methodology could account for these differences, it is more likely that CIBERSORTx was unable to discriminate between certain clusters due to their high similarity. For example, cells that could have been predicted to be assigned to cluster 8, which makes up a large proportion of the scRNA-seq data, may have been assigned to other similar B-cell clusters. The ability to discriminate between similar clusters may have been impacted by down sampling each cluster to include the same number of cells for the analysis. Overall, deconvolution was useful in assigning cell type level data but in some instances, it could not fully deconvolute bulk RNAseq to the cluster specific level.

Integration of porcine PBMC scRNA-seq with a human PBMC scRNA-seq dataset did allow further resolution of porcine cluster annotations and yielded high confidence of homology between many porcine and human single cell populations. While we cannot completely discount the potential for recognized cell types in our scRNA-seq dataset not being present in sorted populations used for bulkRNA-seq (or vice-versa), it seems more likely this is similar evidence to that described above indicating that the same level of resolution simply was not captured by bulkRNA-seq and could not well represent all cell types found in the scRNA-seq data. Integration with another scRNA-seq dataset, even when accounting for cross-species comparison, was in many ways more informative for further annotating porcine single cells, highlighting the enhanced ability of scRNA-seq to define cellular landscapes. Moreover, cross-species integration extended our knowledge of comparative immunology between humans and pigs, as we could identify most similar human counterparts by reference-based prediction. Conversely, we also identified clusters of CD2- γδ T-cells (clusters 6 and 21) and B-cells enriched for activation or cycling-specific genes (clusters 16 and 33) that were more prevalent in porcine data by *de novo* visualization of single cells using the combined human and porcine scRNA-seq data. $CD2^-$ γδ T-cells are frequent in porcine circulation but are reported absent in humans and mice (Stepanova and Sinkora, 2013), and our analyses supported the presence in pigs but not humans. On the other hand, B-cells with transcriptional profiles characteristic of activated or cycling cells, similar to porcine clusters 16 and 33, likely still occur in humans, albeit with low prevalence in circulation. B-cell ontogeny and activation are less fully understood in pigs than in humans, and it's possible peripheral B-cells in clusters 16 and 33 arise from a developmental, activation, or circulation process specific to pigs. In pigs, the majority of leukocytes exit lymph nodes through the vasculature and directly re-enter the blood rather than efferent lymph, as observed in humans (Sasaki et al., 1994; Saalmüller and Gerner, 2016). Thus, it's possible the different patterns of egress for activated cells leaving sites of immune induction might contribute to a higher frequency of activated B-cells entering circulation in pigs compared to humans.

While we did not perform deeper biological query of all cell types identified in our scRNA-seq dataset, we did attempt to deduce biological significance for the different $CD4^+$ αβ T-cell populations that have unique aspects in pigs. Deeper query of $CD4^+$ αβ T-cells was performed, as there is functional interest in determining activation states of porcine $CD4^+$ αβ T-cells based on CD8α expression, which may be gained upon activation and retained in a memory state (Summerfield et al., 1996; Zuckermann, 1999; Saalmüller et al., 2002; Gerner et al., 2009). We found it difficult to identify $CD4^+$ αβ T-cell clusters as $CD8\alpha^+$ or $CD8\alpha^-$ due to sparsity in *CD8A* expression but could leverage comparison of CD4T and CD4CD8T populations from bulkRNA-seq to formulate gene sets enriched in each CD4 expressing T-cell population. GSEA helped identify one cluster of $CD4^+CD8\alpha^-$ αβ T-cells that corresponded mostly to human naïve CD4 T-cells, while three clusters of $CD4^+CD8\alpha^+$ αβ T-cells corresponded to human memory or proliferating CD4 T-cells. Collectively, these data reinforce previous porcine literature, elucidate parallels to human cells, and provide greater insight into the spectrum of activation states present in $CD4^+CD8\alpha^+$ αβ T-cells. Future analysis of activated T-cells or trajectory analysis may provide even further insight on the transition of activation states in porcine peripheral T-cells.

Pigs are a 'γδ high' species, named as such because they have a higher proportions of γδ T-cells in circulation, largely attributed to the presence of $CD2^-$ γδ T-cells that are absent in humans and mice (Stepanova and Sinkora, 2013). Three major γδ T-cell populations are characterized in pigs: $CD2^-CD8\alpha^-$ γδ T-cells that express SWC6 and $CD2^+CD8\alpha^{-/+}$ γδ T-cells that do not express SWC6, where $CD2^-CD8\alpha^-$ γδ T-cells become $CD2^+CD8\alpha^+$ upon activation (Stepanova and Sinkora, 2013; Sedlak et al., 2014). As our sorting strategy for bulkRNA-seq utilized an anti-SWC6 antibody rather than a pan-γδ T-cell-specific antibody; thus, γδ T-cells for bulk RNA-seq included $CD2^-CD8\alpha^-$ γδ T-cells in the SWC6gdT population or $CD2^+CD8\alpha^+$ γδ T-cells found in combination with $CD4^-CD8\alpha^+$ αβ T-cells in the CD8T population. $CD2^+CD8\alpha^-$ γδ T-cells were expected to be excluded in cell sorting. In future sorting strategies, it may be beneficial to utilize a pan-γδ T-cell reactive antibody and/or identify $CD4^-CD8^+$ αβ T-cells with anti-CD8β antibody, which should not label with $CD2^+CD8\alpha^+$ γδ T-cells (Gerner et al., 2009). though this may still exclude potential $CD4^-CD8\alpha^+CD8\beta^-$ αβ T-cells, such as we observed in clusters 5 and 17. Despite limitations in sorting, the bulkRNA-seq profiles were still informative when comparing to

scRNA-seq data. The highest relative enrichment of SWC6gdT gene signatures was detected in $CD2^-$ γδ T-cell clusters, while $CD2^+$ γδ T-cell clusters showed relative enrichment to a lesser level, indicating some conserved gene expression between $CD2^-$and $CD2^+$ γδ T-cells. Comparison between $CD2^+$ γδ T-cell clusters further supported previous biological understanding, where $CD2^+CD8\alpha^+$ γδ T-cells had greater expression of genes related to cellular activation and cytotoxicity relative to $CD2^+CD8\alpha^-$ γδ T-cells (Yang and Parkhouse, 1997; Stepanova and Sinkora, 2013; Sedlak et al., 2014). On the other hand, $CD2^-$ γδ T-cells are less well described than $CD2^+$ γδ T-cells, largely due to lack of comparable populations in humans or mice that may be used for biological inference. Integration with human scRNA-seq data supported previous observations of the absence of $CD2^-$ γδ T-cells in humans, as close counterparts for $CD2^-$ γδ T-cell clusters could not be found by *de novo* visualization, and reference-based integration indicated closest human counterparts to be a mixture of primarily γδ T-cells, ILCs, and CD4 T_{CM}s, and mapping scores were highest for human ILCs rather than γδ T-cells, indicating human ILCs to be the closest, albeit still poor, human match. Nonetheless, we were able to highlight transcriptional distinctions that better annotate $CD2^-$ γδ T-cells, including DEGs between $CD2^-$ and $CD2^+$ γδ T-cells that defined the two γδ T-cell lineages and between two clusters of $CD2^-$ γδ T-cells that have not yet been described.

CONCLUSION

This study provides a first-generation atlas annotating circulating porcine immune cell transcriptomes at both the cell surface marker-sorted population and single-cell levels. These findings illuminate the landscape of immune cell molecular signatures useful for porcine immunology and a deeper annotation of the genome, a goal of the FAANG project. These results also provide useful resources to identify new porcine cell biomarkers for discrimination and isolation of specific cell types, urgently needed in the field.

DATA AVAILABILITY STATEMENT

Raw sequencing data from bulkRNA-seq and scRNA-seq are available through the European Nucleotide Archive (project: PRJEB43826), https://www.ebi.ac.uk/ena/browser/view/PRJEB43826. Processed and finalized data for single-cell datasets are available at https://data.nal.usda.gov/dataset/data-reference-transcriptomics-porcine-peripheral-immune-cells-created-through-bulk-and-single-cell-rna-sequencing. Scripts used for data analyses are available at https://github.com/USDA-FSEPRU/PorcinePBMCs_bulkRNAseq_scRNAseq.

ETHICS STATEMENT

The animal study was reviewed and approved by NADC Animal Care and Use Committee.

AUTHOR CONTRIBUTIONS

CL and CT conceptualized and supervised research. JH-U, JW, KB, and CL collected samples and isolated cells. KB performed cell staining and FACS. JH-U performed bulk RNA isolations. TS performed bulkRNA-seq library preparation and sequencing. JL supervised the NanoString assay data collection. JH-U, LD, and HL performed bulkRNA-seq analyses. HL performed NanoString analyses. JW, LD, SS, and HL performed scRNA- seq analyses. JH-U, JW, CL, and CT interpreted the data and drafted the manuscript. All authors contributed to the writing of the materials and methods, edited the manuscript, and approved the final version.

FUNDING

This work was supported by (1) the National Institute of Food and Agriculture (NIFA) Project 2018-67015-2701, (2) the NRSP-8 Swine Genome Coordination project, (3) appropriated funds from USDA-ARS CRIS project 5030-31320-004-00D and 3040-31000-100-00D, and (4) an appointment to the Agricultural Research Service (ARS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the U.S. Department of Agriculture (USDA). ORISE is managed by ORAU under DOE contract number DE-SC0014664, (5) USDA ARS CRIS Project 8042-32000-102. All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of USDA, ARS, DOE, or ORAU/ORISE. Mention of trade names or products is for information purposes only and does not imply endorsement by the USDA. USDA is an equal opportunity employer and provider.

ACKNOWLEDGMENTS

We acknowledge the DNA facility of the Iowa State University for provision of technical support and sequencing platforms utilized in this study. We are grateful to the NADC animal care staff for their efforts. We thank Dr. Catherine Ernst at Michigan State University for providing the Yorkshire pigs used in the bulkRNA-seq study. We also thank Samuel Humphrey for cell sorting technical assistance. We also thank Kristen Walker for performing the NanoString data collection and initial analyses. We also thank Zahra Bond for technical assistance with cell isolation and FACS, and Dr. Julian Trachsel for early data visualization assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.689406/full#supplementary-material

REFERENCES

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, 1083–1086. doi: 10.1038/nmeth.4463

Alter, G., Malenfant, J. M., and Altfeld, M. (2004). CD107a as a functional marker for the identification of natural killer cell activity. *J. Immunol. Methods* 294, 15–22. doi: 10.1016/j.jim.2004.08.008

Alvarez, J. I., Kébir, H., Cheslow, L., Charabati, Chabarati, M., Larochelle, C., et al. (2015). JAML mediates monocyte and CD8 T cell migration across the brain endothelium. *Ann. Clin. Transl. Neurol.* 2, 1032–1037. doi: 10.1002/acn3.255

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* 16:57. doi: 10.1186/s13059-015-0622-4

Andrews, T. S., Kiselev, V. Y., McCarthy, D., and Hemberg, M. (2021). Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat. Protoc.* 16, 1–9. doi: 10.1038/s41596-020-00409-w

Antsiferova, M., Huber, M., Meyer, M., Piwko-Czuchra, A., Ramadan, T., MacLeod, A. S., et al. (2011). Activin enhances skin tumourigenesis and malignant progression by inducing a pro-tumourigenic immune cell response. *Nat. Commun.* 2:576. doi: 10.1038/ncomms1585

Arceo, M., Ernst, C. W., Lunney, J. K., Choi, I., Raney, N. E., Huang, T., et al. (2013). Characterizing differential individual response to Porcine Reproductive and Respiratory Syndrome Virus infection through statistical and functional analysis of gene expression. *Front. Livestock Genomics* 3:321. doi: 10.3389/fgene.2012.00321

Auray, G., Keller, I., Python, S., Gerber, M., Bruggmann, R., Ruggli, N., et al. (2016). Characterization and Transcriptomic Analysis of Porcine Blood Conventional and Plasmacytoid Dendritic Cells Reveals Striking Species-Specific Differences. *J. Immunol.* 197, 4791–4806. doi: 10.4049/jimmunol.1600672

Auray, G., Talker, S. C., Keller, I., Python, S., Gerber, M., Liniger, M., et al. (2020). High-Resolution Profiling of Innate Immune Responses by Porcine Dendritic Cell Subsets. *Front. Immunol.* 11:1429. doi: 10.3389/fimmu.2020.01429

Bassler, K., Schulte-Schrepping, J., Warnat-Herresthal, S., Aschenbrenner, A. C., and Schultze, J. L. (2019). The Myeloid Cell Compartment-Cell by Cell. *Annu. Rev. Immunol.* 37, 269–293. doi: 10.1146/annurev-immunol-042718-041728

Bates, D., Machlet, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Statist. Softw.* 67, 1–48.

Belizário, J. E., Neyra, J. M., Setúbal Destro, and Rodrigues, M. F. (2018). When and how NK cell-induced programmed cell death benefits immunological protection against intracellular pathogen infection. *Innate Immun.* 24, 452–465. doi: 10.1177/1753425918800200

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 57, 289–300. doi: 10.2307/2346101

Berry, M. P., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A., Oni, T., et al. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* 466, 973–977. doi: 10.1038/nature09247

Blaser, C., Kaufmann, M., Müller, C., Zimmermann, C., Wells, V., Mallucci, L., et al. (1998). Beta-galactoside-binding protein secreted by activated T cells inhibits antigen-induced proliferation of T cells. *Eur. J. Immunol.* 28, 2311–2319. doi: 10.1002/(SICI)1521-4141(199808)28:08<2311::AID-IMMU2311>3.0.CO;2-G

Blom, B., Heemskerk, M. H., Verschuren, M. C., van Dongen, J. J., Stegmann, A. P., Bakker, A. Q., et al. (1999). Disruption of alpha beta but not of gamma delta T cell development by overexpression of the helix-loop-helix protein Id3 in committed T cell progenitors. *EMBO J.* 18, 2793–2802. doi: 10.1093/emboj/18.10.2793

Bordet, E., Frétaud, M., Crisci, E., Bouguyon, E., Rault, S., Pezant, J., et al. (2019). Macrophage-B Cell Interactions in the Inverted Porcine Lymph Node and Their Response to Porcine Reproductive and Respiratory Syndrome Virus. *Front. Immunol.* 10:953. doi: 10.3389/fimmu.2019.00953

Byrne, K. A., Tuggle, C. K., and Loving, C. L. (2020). Differential induction of innate memory in porcine monocytes by β-glucan or bacillus Calmette-Guerin. *Innate Immun.* 2020:1753425920951607. doi: 10.1177/1753425920951607

Castro, C. D., and Flajnik, M. F. (2014). Putting J chain back on the map: how might its expression define plasma cell development? *J. Immunol.* 193, 3248–3255. doi: 10.4049/jimmunol.1400531

Chaussabel, D., and Baldwin, N. (2014). Democratizing systems immunology with modular transcriptional repertoire analyses. *Nat. Rev. Immunol.* 14, 271–280.

Chaussabel, D., Pascual, V., and Banchereau, J. (2010). Assessing the human immune system through blood transcriptomics. *BMC Biol.* 8:84. doi: 10.1186/1741-7007-8-84

Choi, J., Baldwin, T. M., Wong, M., Bolden, J. E., Fairfax, K. A., Lucas, E. C., et al. (2019). Haemopedia RNA-seq: a database of gene expression during haematopoiesis in mice and humans. *Nucleic Acids Res.* 47, D780–D785. doi: 10.1093/nar/gky1020

Costanzo, M. C., Kim, D., Creegan, M., Lal, K. G., Ake, J. A., Currier, J. R., et al. (2018). Transcriptomic signatures of NK cells suggest impaired responsiveness in HIV-1 infection and increased activity post-vaccination. *Nat. Commun.* 9:1212. doi: 10.1038/s41467-018-03618-w

Dabydeen, S. A., Desai, A., and Sahoo, D. (2019). Unbiased Boolean analysis of public gene expression data for cell cycle gene identification. *Mol. Biol. Cell* 30, 1770–1779. doi: 10.1091/mbc.E19-01-0013

Davis, W. C., Zuckermann, F. A., Hamilton, M. J., Barbosa, J. I., Saalmüller, A., Binns, R. M., et al. (1998). Analysis of monoclonal antibodies that recognize gamma delta T/null cells. *Vet. Immunol. Immunopathol.* 60, 305–316. doi: 10.1016/s0165-2427(97)00107-4

Dawson, H. D., Loveland, J. E., Pascal, G., Gilbert, J. G., Uenishi, H., Mann, K. M., et al. (2013). Structural and functional annotation of the porcine immunome. *BMC Genomics* 14:332. doi: 10.1186/1471-2164-14-332

Delecluse, S., Tsai, M., Shumilov, A., Bencun, M., Arrow, S., Beshirova, A., et al. (2019). Epstein-Barr Virus induces expression of the LPAM-1 integrin in B cells *in vitro* and *in vivo*. *J. Virol.* 93, e1618–e1618. doi: 10.1128/JVI.01618-18

Denyer, M. S., Wileman, T. E., Stirling, C. M., Zuber, B., and Takamatsu, H. H. (2006). Perforin expression can define CD8 positive lymphocyte subsets in pigs allowing phenotypic and functional analysis of natural killer, cytotoxic T, natural killer T and MHC un-restricted cytotoxic T-cells. *Vet. Immunol. Immunopathol.* 110, 279–292. doi: 10.1016/j.vetimm.2005.10.005

Dong, Q., Lunney, J. K., Lim, K. S., Nguyen, Y., Hess, A. S., Beiki, H., et al. (2021). Gene expression in tonsils in swine following infection with porcine reproductive and respiratory syndrome virus. *BMC Vet. Res.* 17:88. doi: 10.1186/s12917-021-02785-1

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191. doi: 10.1038/nprot.2009.97

Fairbairn, L., Kapetanovic, R., Beraldi, D., Sester, D. P., Tuggle, C. K., Archibald, A. L., et al. (2013). Comparative analysis of monocyte subsets in the pig. *J. Immunol.* 190, 6389–6396. doi: 10.4049/jimmunol.1300365

Faldyna, M., Samankova, P., Leva, L., Cerny, J., Oujezdska, J., Rehakova, Z., et al. (2007). Cross-reactive anti-human monoclonal antibodies as a tool for B-cell identification in dogs and pigs. *Vet. Immunol. Immunopathol.* 119, 56–62. doi: 10.1016/j.vetimm.2007.06.022

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., et al. (2019). Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol.* 17:108. doi: 10.1186/s12915-019-0726-5

Freeman, T. C., Ivens, A., Baillie, J. K., Beraldi, D., Barnett, M. W., Dorward, D., et al. (2012). A gene expression atlas of the domestic pig. *BMC Biol.* 10:90. doi: 10.1186/1741-7007-10-90

Gerner, W., Käser, T., and Saalmüller, A. (2009). Porcine T lymphocytes and NK cells–an update. *Dev. Comp. Immunol.* 33, 310–320. doi: 10.1016/j.dci.2008.06.003

Gerner, W., Talker, S. C., Koinig, H. C., Sedlak, C., Mair, K. H., and Saalmuller, A. (2015). Phenotypic and functional differentiation of porcine alphabeta T cells: current knowledge and available tools. *Mol. Immunol.* 66, 3–13. doi: 10.1016/j.molimm.2014.10.025

Giotti, B., Chen, S. H., Barnett, M. W., Regan, T., Ly, T., Wiemann, S., et al. (2019). Assembly of a parts list of the human mitotic cell cycle machinery. *J. Mol. Cell Biol.* 11, 703–718. doi: 10.1093/jmcb/mjy063

Giuffra, E., Tuggle, C. K., and Consortium, F. (2019). Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annu. Rev. Anim. Biosci.* 7, 65–88. doi: 10.1146/annurev-animal-020518-114913

Grewal, I. S., and Flavell, R. A. (1996). The role of CD40 ligand in costimulation and T-cell activation. *Immunol. Rev.* 153, 85–106. doi: 10.1111/j.1600-065x.1996.tb00921.x

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W., Zheng, S., Butler, A., et al. (2020). Integrated analysis of multimodal single-cell data. *bioRxiv*. [preprint].

Herrera-Uribe, J., Liu, H., Byrne, K. A., Bond, Z. F., Loving, C. L., and Tuggle, C. K. (2020). Changes in H3K27ac at Gene Regulatory Regions in Porcine Alveolar Macrophages Following LPS or PolyIC Exposure. *Front. Genet.* 11:817. doi: 10.3389/fgene.2020.00817

Hicks, S. C., and Irizarry, R. A. (2015). quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome Biol.* 16, 117–117. doi: 10.1186/s13059-015-0679-0

Hicks, S. C., Okrah, K., Paulson, J. N., Quackenbush, J., Irizarry, R. A., and Bravo, H. C. (2018). Smooth quantile normalization. *Biostatistics* 19, 185–198. doi: 10.1093/biostatistics/kxx028

Huang, T. H., Uthe, J. J., Bearson, S. M., Demirkale, C. Y., Nettleton, D., Knetter, S., et al. (2011). Distinct peripheral blood RNA responses to *Salmonella* in pigs differing in *Salmonella* shedding levels: intersection of IFNG, TLR and miRNA pathways. *PLoS One* 6:e28768. doi: 10.1371/journal.pone.0028768

Ihn, H. J., Kim, D. H., Oh, S. S., Moon, C., Chung, J. W., Song, H., et al. (2011). Identification of Acvr2a as a Th17 cell-specific gene induced during Th17 differentiation. *Biosci. Biotechnol. Biochem.* 75, 2138–2141. doi: 10.1271/bbb.110436

Kapetanovic, R., Fairbairn, L., Downing, A., Beraldi, D., Sester, D. P., Freeman, T. C., et al. (2013). The impact of breed and tissue compartment on the response of pig macrophages to lipopolysaccharide. *BMC Genomics* 14:581. doi: 10.1186/1471-2164-14-581

Kesarwani, P., Murali, A. K., Al-Khami, A. A., and Mehrotra, S. (2013). Redox regulation of T-cell function: from molecular mechanisms to significance in human health and disease. *Antioxid Redox. Signal* 18, 1497–1534. doi: 10.1089/ars.2011.4073

Kim, C., Jin, J., Weyand, C. M., and Goronzy, J. J. (2020). The Transcription Factor TCF1 in T Cell Differentiation and Aging. *Int. J. Mol. Sci.* 21:21186497. doi: 10.3390/ijms21186497

Knetter, S. M., Bearson, S. M., Huang, T. H., Kurkiewicz, D., Schroyen, M., Nettleton, D., et al. (2015). *Salmonella enterica* serovar Typhimurium-infected pigs with different shedding levels exhibit distinct clinical, peripheral cytokine and transcriptomic immune response phenotypes. *Innate Immun.* 21, 227–241.

Kobayashi, E., Hanazono, Y., and Kunita, S. (2018). Swine used in the medical university: overview of 20 years of experience. *Exp. Anim.* 67, 7–13. doi: 10.1538/expanim.17-0086

Koltes, J. E., Cole, J. B., Clemmens, R., Dilger, R. N., Kramer, L. M., Lunney, J. K., et al. (2019). A Vision for Development and Utilization of High-Throughput Phenotyping and Big Data Analytics in Livestock. *Front. Genet.* 10:1197. doi: 10.3389/fgene.2019.01197

Kreslavsky, T., Vilagos, B., Tagoh, H., Poliakova, D. K., Schwickert, T., Wohner, M., et al. (2017). Essential role of the transcription factor Bhlhe41 in regulating the development, self-renewal and BCR repertoire of B-1a cells. *Nat. Immunol.* 18, 442–455. doi: 10.1038/ni.3694

Krueger, J., Jules, F., Rieder, S. A., and Rudd, C. E. (2017). CD28 family of receptors inter-connect in the regulation of T-cells. *Recept. Clin. Investig.* 4:e1581.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:r29. doi: 10.1186/gb-2014-15-2-r29

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, 882–883. doi: 10.1093/bioinformatics/bts034

Li, S., Labaj, P. P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., et al. (2014). Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotech.* 32, 888–895. doi: 10.1038/nbt.3000

Liu, G., Wang, B., Chen, Q., Li, Y., Li, B., Yang, N., et al. (2020). Interleukin (IL)-21 Promotes the Differentiation of IgA-Producing Plasma Cells in Porcine Peyer's Patches via the JAK-STAT Signaling Pathway. *Front. Immunol.* 11:1303. doi: 10.3389/fimmu.2020.01303

Liu, S., and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* 5:F1000FacultyRev–182. doi: 10.12688/f1000research.7223.1

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lun, A. T. L., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., Marioni, J. C., et al. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* 20:63. doi: 10.1186/s13059-019-1662-y

Mach, N., Gao, Y., Lemonnier, G., Lecardonnel, J., Oswald, I. P., Estellé, J., et al. (2013). The peripheral blood transcriptome reflects variations in immunity traits in swine: towards the identification of biomarkers. *BMC Genomics* 14:894. doi: 10.1186/1471-2164-14-894

Maroilley, T., Lemonnier, G., Lecardonnel, J., Esquerré, D., Ramayo-Caldas, Y., Mercat, M. J., et al. (2017). Deciphering the genetic regulation of peripheral blood transcriptome in pigs through expression genome-wide association study and allele-specific expression analysis. *BMC Genomics* 18:967. doi: 10.1186/s12864-017-4354-6

Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y. Y., Carré, C., et al. (2019). RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.* 26, 1627.e–1640.e. doi: 10.1016/j.celrep.2019.01.041

Munyaka, P. M., Kommadath, A., Fouhse, J., Wilkinson, J., Diether, N., Stothard, P., et al. (2019). Characterization of whole blood transcriptome and early-life fecal microbiota in high and low responder pigs before, and after vaccination for Mycoplasma hyopneumoniae. *Vaccine* 37, 1743–1755. doi: 10.1016/j.vaccine.2019.02.016

Nestorowa, S., Hamey, F. K., Pijuan Sala, B., Diamanti, E., Shepherd, M., Laurenti, E., et al. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* 128, e20–e31. doi: 10.1182/blood-2016-05-716480

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, 773–782. doi: 10.1038/s41587-019-0114-2

Palmer, C., Diehn, M., Alizadeh, A. A., and Brown, P. O. (2006). Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics* 7:115. doi: 10.1186/1471-2164-7-115

Paulson, J. N., Chen, C. Y., Lopes-Ramos, C. M., Kuijjer, M. L., Platig, J., Sonawane, A. R., et al. (2017). Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. *BMC Bioinformatics* 18:437. doi: 10.1186/s12859-017-1847-x

Piriou-Guzylack, L., and Salmon, H. (2008). Membrane markers of the immune cells in swine: an update. *Vet. Res.* 39:54. doi: 10.1051/vetres:2008030

Pizzolato, G., Kaminski, H., Tosolini, M., Franchini, D. M., Pont, F., Martins, F., et al. (2019). Single-cell RNA sequencing unveils the shared and the distinct cytotoxic hallmarks of human TCRVδ1 and TCRVδ2 γδ T lymphocytes. *Proc. Natl. Acad. Sci. U S A.* 116, 11906–11915. doi: 10.1073/pnas.1818488116

Qian Liu, M. M. (2016). Evaluation of methods in removing batch effects on RNA-seq data. *Infect. Dis. Transl. Med.* 2, 3–9. doi: 10.11979/idtm.20161001002

Rincon-Orozco, B., Kunzmann, V., Wrobel, P., Kabelitz, D., Steinle, A., and Herrmann, T. (2005). Activation of V gamma 9V delta 2 T cells by NKG2D. *J. Immunol.* 175, 2144–2151. doi: 10.4049/jimmunol.175.4.2144

Rodríguez-Gómez, I. M., Talker, S. C., Käser, T., Stadler, M., Reiter, L., Ladinig, A., et al. (2019). Expression of T-Bet, Eomesodermin, and GATA-3 Correlates With Distinct Phenotypes and Functional Properties in Porcine γδ T Cells. *Front. Immunol.* 10:396. doi: 10.3389/fimmu.2019.00396

Rusmini, M., Griseri, P., Lantieri, F., Matera, I., Hudspeth, K. L., Roberto, A., et al. (2013). Induction of RET dependent and independent pro-inflammatory programs in human peripheral blood mononuclear cells from Hirschsprung patients. *PLoS One* 8:e59066. doi: 10.1371/journal.pone.0059066

Rusmini, M., Griseri, P., Matera, I., Pontarini, E., Ravazzolo, R., Mavilio, D., et al. (2014). Expression variability and function of the RET gene in adult peripheral blood mononuclear cells. *J. Cell Physiol.* 229, 2027–2037. doi: 10.1002/jcp.24660

Saalmüller, A., and Gerner, W. (2016). The immune system of swine. *Encyclop. Immunobiol.* 1, 538–548. doi: 10.1016/B978-0-12-374279-7.12019-3

Saalmüller, A., Werner, T., and Fachinger, V. (2002). T-helper cells from naive to committed. *Vet. Immunol. Immunopathol.* 87, 137–145. doi: 10.1016/s0165-2427(02)00045-4

Sasaki, K., Pabst, R., and Rothkotter, H. J. (1994). The unique ultrastructure of high-endothelial venules in inguinal lymph nodes of the pig. *Cell Tissue Res.* 276, 85–90. doi: 10.1007/BF00354787

Schroyen, M., and Tuggle, C. K. (2015). Current transcriptomics in pig immunity research. *Mamm. Genome* 26, 1–20. doi: 10.1007/s00335-014-9549-4

Sedlak, C., Patzl, M., Saalmüller, A., and Gerner, W. (2014). CD2 and CD8α define porcine γδ T cells with distinct cytokine production profiles. *Dev. Comp. Immunol.* 45, 97–106. doi: 10.1016/j.dci.2014.02.008

Shi, W., Liao, Y., Willis, S. N., Taubenheim, N., Inouye, M., Tarlinton, D. M., et al. (2015). Transcriptional profiling of mouse B cell terminal differentiation defines a signature for antibody-secreting plasma cells. *Nat. Immunol.* 16, 663–673. doi: 10.1038/ni.3154

Siegers, G. M. (2018). Integral Roles for Integrins in γδ T Cell Function. *Front. Immunol.* 9:521. doi: 10.3389/fimmu.2018.00521

Song, L., and Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *Gigascience* 4:48. doi: 10.1186/s13742-015-0089-y

Steiner, J., Marquardt, N., Pauls, I., Schiltz, K., Rahmoune, H., Bahn, S., et al. (2011). Human CD8(+) T cells and NK cells express and secrete S100B upon stimulation. *Brain Behav. Immun.* 25, 1233–1241. doi: 10.1016/j.bbi.2011.03.015

Stepanova, K., and Sinkora, M. (2013). Porcine γδ T lymphocytes can be categorized into two functionally and developmentally distinct subsets according to expression of CD2 and level of TCR. *J. Immunol.* 190, 2111–2120. doi: 10.4049/jimmunol.1202890

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., et al. (2019). Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888.e–1902.e. doi: 10.1016/j.cell.2019.05.031

Summerfield, A., Meurens, F., and Ricklin, M. E. (2015). The immunology of the porcine skin and its value as a model for human skin. *Mol. Immunol.* 66, 14–21. doi: 10.1016/j.molimm.2014.10.023

Summerfield, A., Rziha, H. J., and Saalmüller, A. (1996). Functional characterization of porcine CD4+CD8+ extrathymic T lymphocytes. *Cell Immunol.* 168, 291–296. doi: 10.1006/cimm.1996.0078

Summers, K. M., Bush, S. J., and Hume, D. A. (2020). Network analysis of transcriptomic diversity amongst resident tissue macrophages and dendritic cells in the mouse mononuclear phagocyte system. *PLoS Biol.* 18:e3000859. doi: 10.1371/journal.pbio.3000859

Sutermaster, B. A., and Darling, E. M. (2019). Considerations for high-yield, high-throughput cell enrichment: fluorescence versus magnetic sorting. *Sci. Rep.* 9:227. doi: 10.1038/s41598-018-36698-1

Swindle, M. M., Makin, A., Herron, A. J., Clubb, F. J. Jr., and Frazier, K. S. (2012). Swine as models in biomedical research and toxicology testing. *Vet. Pathol.* 49, 344–356. doi: 10.1177/0300985811402846

Szabo, P. A., Levitin, H. M., Miron, M., Snyder, M. E., Senda, T., Yuan, J., et al. (2019). Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat. Commun.* 10:4706. doi: 10.1038/s41467-019-12464-3

Takamatsu, H. H., Denyer, M. S., Stirling, C., Cox, S., Aggarwal, N., Dash, P., et al. (2006). Porcine gammadelta T cells: possible roles on the innate and adaptive immune responses following virus infection. *Vet. Immunol. Immunopathol.* 112, 49–61. doi: 10.1016/j.vetimm.2006.03.011

Toka, F. N., Nfon, C. K., Dawson, H., and Golde, W. T. (2009). Accessory-cell-mediated activation of porcine NK cells by toll-like receptor 7 (TLR7) and TLR8 agonists. *Clin. Vaccine Immunol.* 16, 866–878. doi: 10.1128/CVI.00035-09

Tsang, J. S., Schwartzberg, P. L., Kotliarov, Y., Biancotto, A., Xie, Z., Germain, R. N., et al. (2014). Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell* 157, 499–513.

Tu, Y. H., Cooper, A. J., Teng, B., Chang, R. B., Artiga, D. J., Turner, H. N., et al. (2018). An evolutionarily conserved gene family encodes proton-selective ion channels. *Science* 359, 1047–1050. doi: 10.1126/science.aao3264

Upadhyay, G. (2019). Emerging Role of Lymphocyte Antigen-6 Family of Genes in Cancer and Immune Cells. *Front. Immunol.* 10:819.

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.* 19, 575–592. doi: 10.1093/bib/bbw139

Van Goor, A., Pasternak, A., Walker, K., Hong, L., Malgarin, C., MacPhee, D. J., et al. (2020). Differential responses in placenta and fetal thymus at 12?days post infection elucidate mechanisms of viral level and fetal compromise following PRRSV2 infection. *BMC Genomics* 21:763. doi: 10.1186/s12864-020-07154-0

Vieira Braga, F. A., Teichmann, S. A., and Chen, X. (2016). Genetics and immunity in the era of single-cell genomics. *Hum. Mol. Genet.* 25, R141–R148. doi: 10.1093/hmg/ddw192

Villani, A. C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 356:aah4573. doi: 10.1126/science.aah4573

Wang, H., Horbinski, C., Wu, H., Liu, Y., Sheng, S., Liu, J., et al. (2016). NanoStringDiff: a novel statistical method for differential expression analysis based on NanoString nCounter data. *Nucleic Acids Res.* 44:e151. doi: 10.1093/nar/gkw677

Wang, M., Windgassen, D., and Papoutsakis, E. T. (2008). Comparative analysis of transcriptional profiling of CD3+, CD4+ and CD8+ T cells identifies novel immune response players in T-cell activation. *BMC Genomics* 9:225. doi: 10.1186/1471-2164-9-225

Ware, C. F. (2005). Network communications: lymphotoxins, LIGHT, and TNF. *Annu. Rev. Immunol.* 23, 787–819. doi: 10.1146/annurev.immunol.23.021704.115719

Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., et al. (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience* 9:giaa051. doi: 10.1093/gigascience/giaa051

Wilson, N. K., and Göttgens, B. (2018). Single-Cell Sequencing in Normal and Malignant Hematopoiesis. *Hemasphere* 2:e34. doi: 10.1097/HS9.0000000000000034

Wolock, S. L., Lopez, R., and Klein, A. M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* 8, 281.e–291.e. doi: 10.1016/j.cels.2018.11.005

Wong, R., and Bhattacharya, D. (2020). ZBTB38 is dispensable for antibody responses. *PLoS One* 15:e0235183. doi: 10.1371/journal.pone.0235183

Yang, H., and Parkhouse, R. M. (1996). Phenotypic classification of porcine lymphocyte subpopulations in blood and lymphoid tissues. *Immunology* 89, 76–83. doi: 10.1046/j.1365-2567.1996.d01-705.x

Yang, H., and Parkhouse, R. M. (1997). Differential expression of CD8 epitopes amongst porcine CD8-positive functional lymphocyte subsets. *Immunology* 92, 45–52. doi: 10.1046/j.1365-2567.1997.00308.x

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Alvarez-Jarreta, J., Amode, M. R., et al. (2020). Ensembl 2020. *Nucleic Acids Res.* 48, D682–D688. doi: 10.1093/nar/gkz966

Young, M. D., and Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* 9:giaa151. doi: 10.1093/gigascience/giaa151

Zemmour, D., Zilionis, R., Kiner, E., Klein, A. M., Mathis, D., and Benoist, C. (2018). Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nat. Immunol.* 19, 291–301. doi: 10.1038/s41590-018-0051-0

Zhang, B., Lin, Y. Y., Dai, M., and Zhuang, Y. (2014). Id3 and Id2 act as a dual safety mechanism in regulating the development and population size of innate-like γδ T cells. *J. Immunol.* 192, 1055–1063. doi: 10.4049/jimmunol.1302694

Zheng, G. X., 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

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10:1523. doi: 10.1038/s41467-019-09234-6

Zhu, L., Yang, P., Zhao, Y., Zhuang, Z., Wang, Z., Song, R., et al. (2020). Single-Cell Sequencing of Peripheral Mononuclear Cells Reveals Distinct Immune Response Landscapes of COVID-19 and Influenza Patients. *Immunity* 53, 685.e–696.e. doi: 10.1016/j.immuni.2020.07.009

Zhu, Y., Yao, S., Iliopoulou, B. P., Han, X., Augustine, M. M., Xu, H., et al. (2013). B7-H5 costimulates human T cells via CD28H. *Nat. Commun.* 4:2043. doi: 10.1038/ncomms3043

Zuckermann, F. A. (1999). Extrathymic CD4/CD8 double positive T cells. *Vet. Immunol. Immunopathol.* 72, 55–66.

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.

Copyright © 2021 Herrera-Uribe, Wiarda, Sivasankaran, Daharsh, Liu, Byrne, Smith, Lunney, Loving and Tuggle. 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: 28 June 2021
doi: 10.3389/fgene.2021.655707

RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and *Cis*-Regulation Analysis by Allele-Specific Expression in Livestock Species

Frédéric Jehl[1†], Fabien Degalez[1†], Maria Bernard[2,3†], Frédéric Lecerf[1], Laetitia Lagoutte[1], Colette Désert[1], Manon Coulée[1], Olivier Bouchez[4], Sophie Leroux[5], Behnam Abasht[6], Michèle Tixier-Boichard[3], Bertrand Bed'hom[3], Thierry Burlot[7], David Gourichon[8], Philippe Bardou[2], Hervé Acloque[3], Sylvain Foissac[5], Sarah Djebali[5], Elisabetta Giuffra[3], Tatiana Zerjal[3], Frédérique Pitel[5], Christophe Klopp[2‡] and Sandrine Lagarrigue[1*‡]*

[1] INRAE, INSTITUT AGRO, PEGASE UMR 1348, Saint-Gilles, France, [2] INRAE, SIGENAE, Genotoul Bioinfo MIAT, Castanet-Tolosan, France, [3] INRAE, AgroParisTech, Université Paris-Saclay, GABI UMR 1313, Jouy-en-Josas, France, [4] INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France, [5] INRAE, INPT, ENVT, Université de Toulouse, GenPhySE UMR 1388, Castanet-Tolosan, France, [6] Department of Animal and Food Sciences, University of Delaware, Newark, DE, United States, [7] NOVOGEN, Maugueérand, Le Foeil, France, [8] INRAE, PEAT UE, Nouzilly, France

In addition to their common usages to study gene expression, RNA-seq data accumulated over the last 10 years are a yet-unexploited resource of SNPs in numerous individuals from different populations. SNP detection by RNA-seq is particularly interesting for livestock species since whole genome sequencing is expensive and exome sequencing tools are unavailable. These SNPs detected in expressed regions can be used to characterize variants affecting protein functions, and to study *cis*-regulated genes by analyzing allele-specific expression (ASE) in the tissue of interest. However, gene expression can be highly variable, and filters for SNP detection using the popular GATK toolkit are not yet standardized, making SNP detection and genotype calling by RNA-seq a challenging endeavor. We compared SNP calling results using GATK suggested filters, on two chicken populations for which both RNA-seq and DNA-seq data were available for the same samples of the same tissue. We showed, in expressed regions, a RNA-seq precision of 91% (SNPs detected by RNA-seq and shared by DNA-seq) and we characterized the remaining 9% of SNPs. We then studied the genotype (GT) obtained by RNA-seq and the impact of two factors (GT call-rate and read number per GT) on the concordance of GT with DNA-seq; we proposed thresholds for them leading to a 95% concordance. Applying these thresholds to 767 multi-tissue RNA-seq of 382 birds of 11 chicken populations, we found 9.5 M SNPs in total, of which ∼550,000 SNPs per tissue and population with a reliable GT (call rate ≥ 50%) and among them, ∼340,000 with a MAF ≥ 10%. We showed that such RNA-seq data from one tissue can be used to (*i*) detect SNPs with a strong predicted impact on

OPEN ACCESS

Edited by:
James Reecy,
Iowa State University, United States

Reviewed by:
Stephen J. Bush,
University of Oxford, United Kingdom
Melissa Susan Monson,
Iowa State University, United States

***Correspondence:**
Christophe Klopp
christophe.klopp@inrae.fr
Sandrine Lagarrigue
sandrine.lagarrigue@agrocampus-ouest.fr

[†]These authors share first authorship

[‡]These authors share last authorship

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

Received: *19 January 2021*
Accepted: *01 June 2021*
Published: *28 June 2021*

Citation:
Jehl F, Degalez F, Bernard M, Lecerf F, Lagoutte L, Désert C, Coulée M, Bouchez O, Leroux S, Abasht B, Tixier-Boichard M, Bed'hom B, Burlot T, Gourichon D, Bardou P, Acloque H, Foissac S, Djebali S, Giuffra E, Zerjal T, Pitel F, Klopp C and Lagarrigue S (2021) RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. Front. Genet. 12:655707. doi: 10.3389/fgene.2021.655707

proteins, despite their scarcity in each population (16,307 SIFT deleterious missenses and 590 stop-gained), (*ii*) study, on a large scale, *cis*-regulations of gene expression, with ∼81% of protein-coding and 68% of long non-coding genes (TPM ≥ 1) that can be analyzed for ASE, and with ∼29% of them that were *cis*-regulated, and (*iii*) analyze population genetic using such SNPs located in expressed regions. This work shows that RNA-seq data can be used with good confidence to detect SNPs and associated GT within various populations and used them for different analyses as GTEx studies.

Keywords: RNA-seq, SNP calling, genotype calling, SNP annotation, allele-specific expression, livestock, chicken

INTRODUCTION

RNA-seq is currently the method of choice to study transcriptome expression in replacement of gene chips (Mortazavi et al., 2008). This technology is commonly used to study gene expression patterns in a variety of organisms including plant, animal or human groups to better understand the genetic mechanisms intervening in the determinism of phenotypes (Gondret et al., 2017), diseases (Savary et al., 2020) or response to environmental changes (Jehl et al., 2019) among others. The RNA-seq has other more specific applications taking advantage of its sequencing step. For example RNA-seq allows transcript and gene modeling as shown by long non-coding atlas reported in different species (Derrien et al., 2012; Jehl et al., 2020). It also allows to combine SNP information, at the RNA level with gene expression to study the variation which affects gene-expression levels: it is a powerful technology to identify such expression quantitative trait locus (eQTL) either through GWAS mapping (if the individual number is sufficient) or through allele-specific expression (ASE) analysis as shown by growing number of studies on a variety of species since the beginning of the RNA-seq technology in the 2010s (Montgomery et al., 2010; Pickrell et al., 2010; Battle et al., 2013; Lagarrigue et al., 2013b; Chamberlain et al., 2015; Deelen et al., 2015; The GTEx Consortium, 2020), among them the famous studies from the human GTEx consortium (The GTEx Consortium, 2020). Finally RNA-seq allows RNA editing analysis, a phenomenon resulting in nucleotide changes observed at RNA level, occurring after its transcription from DNA level (Kleinman et al., 2012). In these two last applications, RNA-seq is in general combined with DNA-seq used for genotyping individuals. However, RNA-seq can also detect genomic variations in expressed regions like DNA-seq, as described by Piskol et al. (2013). It is particularly interesting in non-model species (wild or domesticated, for example livestock species) in which no exome capturing tools have been developed as an alternative to DNA-seq data, which remains costly to generate and store. In this context, RNA-seq presents several advantages compared to the DNA-seq. First, the number of RNA-seq data sets publicly available is much higher than the number of DNA-seq data sets, for many species (chicken, pig, cow, and other non-model species) since these data have accumulated over the past several years and continue to accumulate in different populations and within populations. Moreover, within populations, different conditions are studied, increasing the number of studied animals, allowing to better detect, in a given population, variants with low frequencies. Second, RNA-seq data allows studying coding region variations that have potential functional impacts. Some of these SNPs can induce a loss of the protein function. These loss-of-function variants are extensively studied because of their possible contribution to phenotypes (Genome Aggregation Database Consortium et al., 2020). In addition, they represent a powerful source of information to understand gene functions (Genome Aggregation Database Consortium et al., 2020). However, these loss of function SNPs are rather rare because purged by negative selection in natural populations but can be detected with a certain number of samples. In well-known model-species or human, these coding region variants are accessible using whole exome sequencing (WES), as shown by the recent work of the Genome Aggregation Database (gnomAD) (Lek et al., 2016). This consortium analyzed 125,748 human exomes (and much fewer whole genomes: 15,708) from public sources and identified 443,769 high-confidence predicted loss-of-function variants, defined in the work of gnomAD as being either gain of stop (non-sense variants), frameshift or splice site variants. For non-model species such as livestock species, for which the WES method is usually not available, RNA-seq can thus fulfill the same objective, with a similar advantage that is, producing a smaller data volume, thus facilitating data storage and decreasing costs (Battle et al., 2013). Third, RNA-seq data provides expression levels of loci harboring SNPs, allowing to study allele-specific expression as we previously mentioned, and hence, to study *cis*-regulation on a large scale, in multiple tissues and multiple populations. Fourth, the transcribed regions are well spread over the genome and much more numerous than previously thought. Thousands of novel long non-coding genes exist across the genome, as highlighted by the ENCODE project (Derrien et al., 2012). RNA-seq data can therefore provide sets of numerous and well distributed SNPs throughout the genome. Finally, these data could be used to study population genetic diversity from a different point of view compared to the SNP chips, by offering various sets of SNPs with more or less severe functional impacts and not neutral SNPs.

Despite the aforementioned advantages RNA-seq is not yet often used for SNP detection in coding regions. Indeed, SNP detection and genotype calling by RNA-seq present three main challenges. First, the transcriptome is composed of mature transcripts (i.e., spliced), making mapping of RNA-seq reads that overlap exon-exon junctions, more difficult, compared to DNA-seq read alignment (Pan et al., 2008). However, RNA-seq

mapping methods seem to be well mastered in recent years, even though it is important to remain cautious for SNPs detected close to exon-exon junctions (Peng et al., 2012; Lagarrigue et al., 2013b). Second, RNA editing, by definition, could represent a strong limitation for SNP detection by RNA-seq, mainly because it introduces variations at the RNA level, which are absent at the DNA level. Nevertheless, as we will discuss later, RNA editing has such features that it only slightly impedes reliable RNA-seq based variation detection in standard conditions. Third, genes exhibit highly variable expression levels, leading to the read depths ranging from a few reads to millions of reads, contrarily to the DNA-seq which offers a rather homogeneous read depth across the genome (see **Figures 1A,B**). Indeed, coding and non-coding transcripts can be expressed at vastly different levels, ranging from few copies to millions of copies per cell, in different cell types and developmental or physiological stages. Moreover, the transcriptome is also composed of a small portion of immature under processing transcripts (composed of exons and introns), less supported by reads but enriched in introns that are more variable in sequence compared to exons (Sims et al., 2014). In summary, these variations in read depth from one gene to another, and within a gene (between introns and exons) constitute a major challenge for SNP detection (see **Figures 1A,B**, left), and more importantly, for individual genotype calling (see **Figures 1A,B**, right). Indeed, reliable SNP detection at the population level benefits from the information accumulation born by the reads across individuals, in contrast to genotype calling. This last point might explain why only few studies have used RNA-seq data for variant detection and genotype calling since the first publications. Consequently, neither the number of SNPs that could be detected using RNA-seq, nor the percentage of individuals with a given genotype (a prerequisite for computing allelic frequencies), are known. To our best knowledge, since Piskol et al. (2013), less than a dozen studies were focused on large-scale SNP detection tools from RNA-seq data (Quinn et al., 2013; Tang et al., 2014; Wang et al., 2014; Wolfien et al., 2016; Oikkonen and Lise, 2017; Cornwell et al., 2018; Adetunji et al., 2019). The reference tools for read mapping and variant detection have been evolving very rapidly, and these studies have tested different tools, and among them, only Adetunji et al., 2019 (Adetunji et al., 2019) used the most recent tools proposed by ENCODE for RNA-seq data, i.e., STAR (Dobin et al., 2013) for read mapping and GATK (Van der Auwera et al., 2013) for variant detection. Three of the above-mentioned studies were interested in determining the concordance of SNP and genotype detection between RNA-seq and DNA-seq, the latter being the gold standard for SNP detection. However, these studies used only few samples (from 1 to 4) and had not at their disposal both RNA-seq and DNA-seq data on the same tissues of the same individuals.

In this context, this work aims at detecting SNPs from RNA-seq data in chicken. The first goal was to set up a procedure allowing SNP detection and genotype (GT) calling from RNA-seq data using reference tools (STAR for read mapping and GATK for SNP detection). We tested the SNP reliability according to three filters suggested by the GATK team and compared the detected SNPs with those obtained using DNA-seq data. This comparison was performed in two independent chicken populations for which RNA-seq and DNA-seq data were available on the same biological samples (i.e., the same tissue of the same

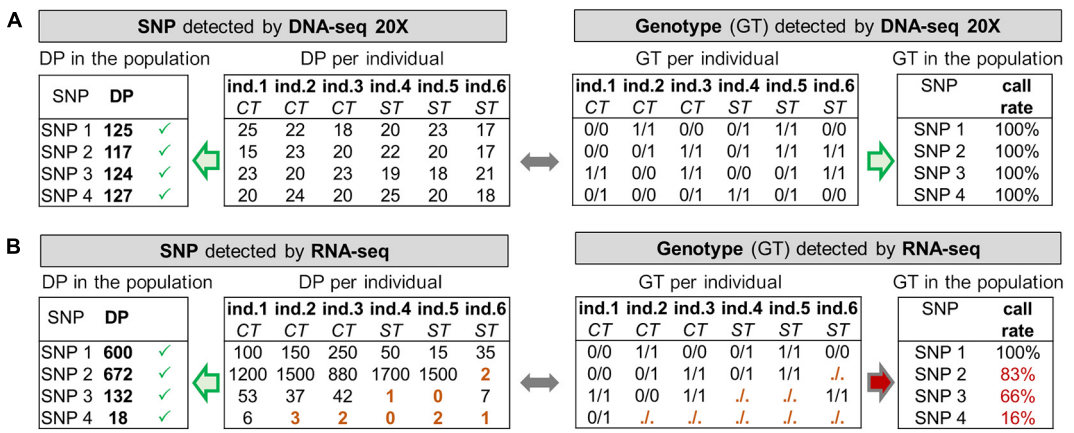


FIGURE 1 | Toy example with simulated data illustrating the need for read depth (DP) filters in RNA-seq and differences with DNA-seq. **(A)** DNA-seq data offers a globally homogeneous genome coverage (20X in our case), all SNPs are therefore detected by GATK at the individual level with a DP of 20 reads on average (*"DP per individual"*), and at the population level with a DP of 6 × 20 = 120 reads on average (*"DP in the population* resulting from the addition of *"DP per individual")*. All genotypes (GT) can therefore be computed at the individual level (*"GT per individual"*), resulting in a genotype call rate of 100% for every SNP (*"GT in the population"*). **(B)** RNA-seq data offers a heterogeneous coverage of the genome depending on the expression of the genes harboring the SNPs. *At the population level,* 4 SNPs having a sufficiently high DP are detected by GATK. At the individual level, SNP 1 shows good read coverage across all samples whereas SNP 3 is on a gene that has a lower expression, in particular in the stress (ST) condition compared to the control (CT). SNP 4 is on an overall very lowly expressed gene. *In terms of genotype (GT) per individual*, some cannot be provided by GATK (noted "./.") because of a too low DP (i.e., 5 reads, see brown GT and DP) and are not considered for the GT call rate. For SNP 3, most of the individuals from the ST condition have no GT and for SNP 4, only one GT is called whereas in both case the SNP is detected at the population-level. "GT in the population" provides for each SNP their call-rate for the genotypes (CR): SNP 1 has 100% of the samples with a GT whereas SNP 4 has 16% and cannot be used to compute meaningful genotype frequencies.

individuals). In this paper, the workflow was used at the tissue level to provide results for RNA-seq experimental settings with only one analyzed tissue which represent a quite common case. This, however, corresponds to the least favorable case compared to multi-tissue experimental projects, since it does not allow cumulating the sequences from tissues per individual. We then analyzed the effects on the number of detected SNPs by this workflow performed at the tissue level when using additional tissues of a same population.

Because a large proportion of SNPs detected by RNA-seq was reliable, we further applied this procedure to 11 different chicken populations: a population derived from the wild Red Jungle Fowl population, an Egyptian Fayoumi population, six commercial and experimental laying hen populations and three commercial and experimental broiler populations. Our three goals were to (*i*) provide an estimation of the number of SNPs and GT that can be detected using RNA-seq data per tissue and population, (*ii*) present an overview of the predicted consequences of the SNPs located in coding regions, in particular, the number of high-confidence predicted loss-of-function variants, as defined in the work of gnomAD, and finally (*iii*) give an overview of the potential of RNA-seq for allele-specific expression (ASE) analysis by estimating the number of genes that could be analyzed for ASE with the number of SNPs detected per gene. We then identified the *cis*-regulated genes in the liver of 2 of the 11 populations using the phASER tool (Castel et al., 2016) and the proportion of cis-regulated hepatic genes shared by the two populations. Finally, we illustrated the possibility of using RNA-seq data to explore genetic diversity between populations using different hepatic RNA-seq SNP sets with variable percentage of severe predicted protein consequence.

MATERIALS AND METHODS

RNA-Seq and DNA-Seq Data

Raw data of both DNA-seq and/or RNA-seq are available on the ENA and SRA archives under accession numbers: PRJEB28745 (RpRm DNA-seq and RNA-seq, Novo1 and Novo2, RNA-seq); PRJEB43829 (FLLL, DNA-seq); PRJNA330615 and PRJNA248570 (FLLL, RNA-seq); PRJEB26695 (red jungle fowl, RNA-seq); PRJEB34341 (Naked neck, RNA-seq); PRJEB34310 (Fayoumi, RNA-seq); PRJEB27455 (FrAg, RNA-seq); PRJEB43662 (Cobb, RNA-seq); PRJNA612882 (HerX, RNA-seq) (Fu et al., 2015). RNA sequencing was conducted on all samples using an Illumina HiSeq (Illumina, California, United States) system, with 2 × 150 bp or 100 bp. Libraries were prepared following Illumina's instructions by purifying poly-A RNAs (TruSeq RNA Sample Prep Kit). Illumina adapters containing indexing tags were added for subsequent identification of samples.

For the comparison of SNPs detected by RNA-seq versus DNA-seq, we used two populations for which both data types were obtained from same liver samples collected on the same birds. The population A was composed of 15 birds from an experimental layer population (RpRm, PRJEB28745) composed of birds diverging for feed efficiency (Rp and Rm) after a 40-year diverging selection (Bordas et al., 1992). The population B was composed of 8 birds from an experimental broiler population (FLLL, PRJNA330615) composed of birds diverging for body fat content (FL and LL) (Roux et al., 2015).

For the rest of the work, we used RNA-seq data from 11 populations (see Additional File 1 for the detail of the number of birds, the tissues and the number of samples): a red jungle fowl population (called RJFh with 36 birds and 3 tissues); 3 broiler populations, the FLLL presented previously but here extended with 32 birds and 2 tissues) and two commercial ones, the Cobb 500 (Cobb Vantress, named Cobb with 48 birds and 2 tissues) and a 3-way cross produced by Heritage Breeders, LLC (named HerX, 23 birds and 1 tissue), 6 layer populations with 2 commercial brown-egg subpopulations from the Novogen company, Novo1 with 32 birds and 1 tissue and Novo2 with 40 birds and 2 tissues, 2 experimental brown-egg populations with the RpRm presented previously but here extended (with 88 birds and 5 tissues) and an experimental dwarf chicken layer line homozygous for the Naked Neck mutation (named LSnu with 16 birds and 2 tissues) and 2 other layer populations with a leghorn breed (FrAg) with 4 birds and 2 tissues) and the Fayoumi (FAyo), an Egyptian breed with 16 birds and 2 tissues; finally an experimental population (Rmx6) issued from crosses between 2 experimental lines (Frésard et al., 2014) with 19 embryos harvested from the same batch at embryonic day 4.5 (stage 26).

RNA-Seq Read Mapping and Variant Detection

For all samples, RNA-seq variants were detected using the snakemake (Koster and Rahmann, 2012) pipeline, available at this reference: (GitLab, 2019). For each population, samples were analyzed by tissue. FASTQ files were trimmed for Illumina adapter using TrimGalore version 0.4.5 (Krueger, 2021). STAR v.2.5.2b (Dobin et al., 2013) was used with default parameters for the read mapping on the Gallus_gallus-5.0 reference genome, after the multi-sample 2-pass mapping procedure, with a GTF file enriched in long non-coding genes [available on http://www.fragencode.org (LNChickenAtlas); Section: Galgal5—Ensembl v94; Genome annotation: LNCextendedEns94.gtf.gz; (Jehl et al., 2020)]. Uniquely mapped reads (selected on a mapping quality score equal to 255) were then post-processed following the GATK best practices for RNA-seq data [duplicates were marked, reads overlapping intron were split and mapping quality score were reassigned, indel were realigned and base were recalibrated thanks to the known variants from Ensembl v94's dbSNP (Ensembl, 2018)]. Variant detection was done for each sample using the "HaplotypeCaller" function of GATK (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013) 3.7.0 with option "—stand_call_conf 20.0," "—min_base_quality_score 10" and "—min_mapping_quality_score 20" (which are the defaults values), generating one gVCF file per sample. The "GenotypeGVCFs" function was then used with option "—stand_call_conf 20.0," to jointly genotype all these samples into one VCF per tissue. The VCF file obtained at the end of the pipeline was then used as the input to two other steps,

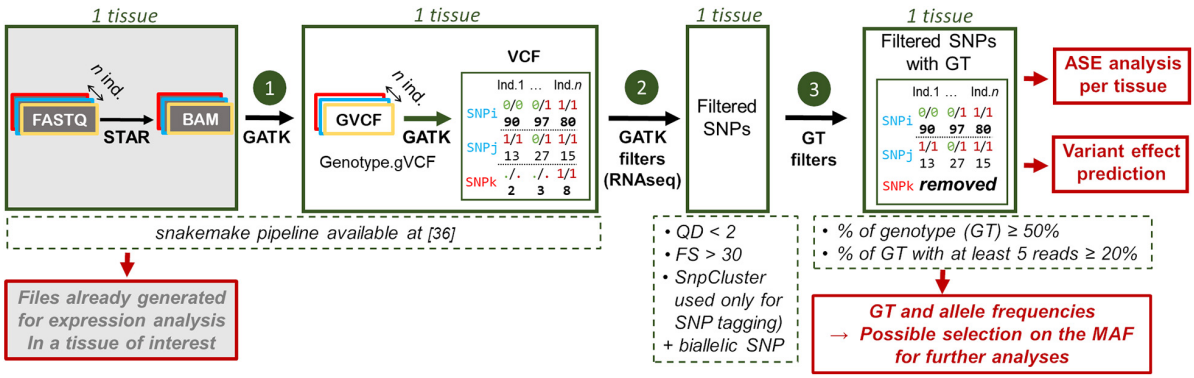


FIGURE 2 | Workflow used to detect SNPs from RNA-seq data. The input files are indicated in gray. GATK filters: QD quality depth, FS: Fisher strand and SnpCluster: 3 or more SNPs in a sliding window of 35 bp. This last criterion was used only for tagging and not for filtering SNPs. For each SNP, are given the genotype (e.g., 0/0) for 3 individuals and under each genotype the associated read number (e.g., 90).

as summarized in **Figure 2**. First, biallelic SNPs were then extracted using the "SelectVariant" function with option "—selectType SNP—restrictAllelesTo BIALLELIC." Variants were also filtered using "VariantFiltration" with two of the three suggested filters, "QD < 2" and "FS > 30," as we discussed in the Results and Discussion section. Finally, we selected the SNPs with genotypes associated with each individual and that met the criteria established in results and Discussion section, i.e. (5.reads.DP) genotype CR ≥ 20% and CR ≥ 50%. Genotype and allele frequencies were then computed, making possible to work on SNPs selected on the minor allele frequency (MAF). These VCF files containing the SNP with their associated genotypes can be used for allele specific expression (ASE) analysis in each tissue of interest.

It is important to note that all previous treatments were conducted in this paper at the tissue level to provide SNP detection results for RNA-seq experimental settings with only one analyzed tissue, which is quite common and corresponds to the least favorable case. This implies that we had one bird's genotype per tissue. For the multi-tissue analysis step of this paper, gVCF files generated per tissue were combined and genotypes were computed from all the tissues information using "CombineGVCFs" and "GenotypeGVCFs" generating per bird as many genotypes as tissues analyzed. Genotype concordance between tissues for a same bird was very high (∼99% of SNPs) and increased with coverage (see result section). Therefore, for the rare cases of discordance, we kept the genotype of the tissue with the highest coverage when they were different. However, outside from this study, for projects in which RNA-seq of different tissues per animal are available when the SNP detection analysis is started, we advise users of our pipeline to define in the first step a sample as a specific individual. This strategy allows to gain power in SNP detection by gathering all BAM tissue files per animal.

DNA-Seq Read Mapping and Variant Detection

DNA-seq read mapping and variant detection were performed using standard tools. The BWA-MEM algorithm (Li, 2013) from BWA-0.7.17 was used with default parameters for the read mapping on the Gallus_gallus-5.0 reference genome (GCA_000002315.3). Variant detection was done for each sample using the "HaplotypeCaller" function of GATK (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013) 3.7.0 with option "-variant_index_type LINEAR," "-variant_index_parameter 128000," "-mmq 30" and "-mbq 10 2," generating one gVCF file per sample. The "CombineGVCFs" and "GenotypeGVCFs" (with "stand_call_conf 20.0" option) functions were then used to combine these gVCF into one VCF per population (one VCF for the 15 RpRm and one VCF for the 8 FLLL). Biallelic SNPs were then extracted using the "SelectVariant" function with option "—selectType SNP—restrictAllelesTo BIALLELIC." Variant were filtered using "VariantFiltration" with all the recommended filters for DNA-seq: "FS > 60.0," "QD < 2.0," "SOR > 3.0," "MQ < 40.0," "MQRankSum < −12.5" and "ReadPosRankSum < −8.0."

Gene and Exon Expression Quantification

Gene expression was quantified with RSEM (Li and Dewey, 2011) v.1.3.0, at the gene-level, using the GTF file LNCextendedEns94.gtf.gz available on http://www.fragencode.org (LNChickenAtlas; section Galgal5) and corresponding to the genes from the Ensembl annotation used as reference, extended with lncRNAs loci available in other public databases (NCBI, NON-CODE, etc.) (Muret et al., 2017). To compute expression at the exon level, we used FeatureCount v1.6.2 (Liao et al., 2014) with options -t "exon" and -g "exon_id." We defined for each exon a metric called RpKb (Read per Kilobase) as the mean number of reads mapped at the exon divided by its length in kilobases. To define an expression threshold, we compared the expression of exons to the expression of a set of randomly selected loci in the genome as done previously in Jehl et al. (2020). The background noise corresponds to the expression of a set of artificial loci randomly distributed across chicken chromosomes 1–33 using the "shuffle" function from the BEDTools suite

v2.29 (Quinlan and Hall, 2010). These artificial loci had the same length distribution as the LNC genes known to be the less expressed compared to PCG and were positioned at a distance of at least 5kb of the closest known transcribed regions. The expression of these randomly selected regions was well below the expression of the exons. We set as an expression threshold for the exons a $\log_{10}(\text{RpKb} + 1)$ value of 0.5, corresponding to the first quartile of expression in both RpRm and FLLL (see Additional File 2).

Variant Functional Predictions

Variant Effect Predictor (VEP) v92 (McLaren et al., 2016) with the GTF file enriched in long non-coding genes ("—gtf") was used for effect prediction of 9,496,283 SNPs. "—everything" and "—total_length" options were applied to respectively, obtain SIFT score predictions and length of cDNA, CDS and proten positions (Ng, 2003; Sim et al., 2012).

Detection of Homopolymers and Exon-Exon Junctions

Regions with 5 or more repeated nucleotides (homopolymers) and regions spanning 5 bp of each extremity of a junction were detected using home-made scripts.

Hierarchical Clustering Analysis

The hierarchical clustering was performed on a set of 67,341 SNPs obtained using liver RNA-seq data from the 10 populations presented in **Table 1** (liver unavailable for Rmx6). This set corresponds to the SNPs common to the 10 populations and passes the GT criteria (see "Results and discussion") for each population. The analysis was produced by using the function "snpgdsHCluster" of the R (R Core Team, 2019) package SNPRelate v1.8.0 (Zheng et al., 2012).

Allele-Specific Expression (ASE) Analysis

Prior to the quantification of allele specific expression, sequences need to be aligned against masked version of the genome to avoid favoring reference alleles. At the population level, polymorphic (allele frequency < 100%) and bi-allelic filtered (GATK—FS and QD criteria) SNP were extracted using the GATK "SelectVariants" tool. These last variants were then used to mask the reference genome using "maskfasta" tool from the BEDTools suite v2.29. Tissue sample sequence were aligned to this masked version of the genome using the multi-sample 2-pass mapping procedure of STAR 2.6. Non-duplicated ("MarkDuplicates" function from GATK 4.1.2, with "READ_NAME_REGEX" set to null) properly paired (if paired sequences) uniquely mapped reads (samtools 1.9 with –f 2 and –q 255 options) were selected. "SplitNCigar" tool from GATK were finally used to split alignment overlapping exon/intron junction and rescaled mapping quality. The phASER tool (Castel et al., 2016) and its downstream tool phASER Gene AE were used to detect ASE among the liver samples of the RpRm and FLLL populations. Briefly, phASER phases, in each sample, SNPs from a user-provided VCF, using the reads from the previously processed BAM file of the sample. This produces a list of haplotypes upon which phASER counts the number of reads associated to each "super-allele." Then, in each sample, phASER Gene AE selects one haplotype per gene, using the genes' boundaries from a user-provided BED file, allowing the study of the gene's ASE using the selected haplotype.

Using base quality of 10, and mapping quality of 20, we provided a VCF containing the SNP that met the criteria established here. After selection of one haplotype per gene using phASER Gene AE, we considered only the genes represented by a haplotype with at least 10 reads associated to at least 1 super-allele. To assess ASE in each sample, we screened for read number imbalance between the super-alleles using a binomial test

TABLE 1 | SNP counts per population retained at each step of the selection.

Population				Total SNP			Selected GT					Selected GT and MAF ≥ 10%				
Pop.	#ind.	#smpl.	#tiss.	Liver[a]	Multi-tiss.[b]	b/a	Liver[c]	Multi-tiss.[d]	d/c	c/a	d/b	Liver[e]	Multi-tiss.[f]	f/e	e/a	f/b
RJFh	36	72	3	1,050,035	2,604,288	2.48	265,750	578,726	2.18	0.25	0.22	152,029	319,268	2.10	0.14	0.12
Cobb	48	96	2	3,771,992	5,464,266	1.45	949,127	1,678,364	1.77	0.25	0.31	558,020	952,445	1.71	0.15	0.17
FLLL	32	64	2	1,729,800	2,033,207	1.18	535,228	1,109,324	2.07	0.31	0.55	368,280	714,523	1.94	0.21	0.35
HerX	23	23	1	1,332,709	1,332,709	1.00	481,314	481,314	1.00	0.36	0.36	307,859	307,859	1.00	0.23	0.23
Novo1	32	32	1	1,459,352	1,459,352	1.00	447,594	447,594	1.00	0.31	0.31	264,804	264,804	1.00	0.18	0.18
Novo2	44	104	2	1,289,199	2,146,975	1.67	390,195	738,109	1.89	0.30	0.34	243,892	449,446	1.84	0.19	0.21
RpRm	112	286	5	1,841,778	4,032,988	2.19	555,928	1,279,458	2.30	0.30	0.32	307,049	631,868	2.06	0.17	0.16
Rmx6	19	19	1	–	2,123,217	–	–	715,822	–	–	0.34	–	483,379	–	–	0.23
FrAg	4	7	2	1,247,253	1,732,440	1.39	784,397	1,055,772	1.35	0.63	0.61	520,277	583,742	1.12	0.42	0.34
Lsnu	16	32	2	1,487,176	2,284,902	1.54	590,399	836,800	1.42	0.40	0.37	384,720	534,938	1.39	0.26	0.23
Fayo	16	32	2	1,320,244	2,033,207	1.54	496,412	698,932	1.41	0.38	0.34	288,464	396,446	1.37	0.22	0.19
Mean				1,652,954	2,477,050	1.54	549,634	874,565	1.64	0.35	0.37	339,539	512,611	1.55	0.22	0.22
Union	382	767		5,490,587	9,496,283		1,685,406	3,276,615				1,255,554	2,243,766			
Intersection				221,374	241,960		67,341	73,223				2,442	1,442			

In columns—Pop., population; #ind., bird number; #smpl., sample number; #tiss., tissue number; Multi-tiss., Multi-tissues. Superscripts are used to show which ratio are presented. Total SNP: SNPs detected at the population level (i.e., with at least one ALT allele); Selected GT: SNPs with at least 50% of genotypes (CR ≥ 50%) and 20% of GT with reads ≥ 5 reads [(5.reads.DP)genotypeCR ≥ 20%, see "Results and discussion"]; Selected GT with minor allele frequency (MAF) ≥ 10%.
In lines—Union: SNPs detected in at least one population; Intersection: SNPs detected in each of the 10 populations (i.e., each population has at least one ALT allele).

(*binom.test* R function) with the null hypothesis that, for a given gene, each super-allele had the same number of associated reads. *P*-values were corrected using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) with a false discovery rate of 0.05. We considered a gene to be ASE if it presented a significant read number imbalance in at least 2 samples.

RESULTS AND DISCUSSION

SNP Detection by RNA-Seq: Genome Location

We compared the repartition of the SNPs detected by DNA-seq and RNA-seq among different genomic regions (**Figure 3A**). The chicken genome is composed at equal parts of intergenic (50%) and genic (50%) sequences, with 43% of introns and 7% of exons. As expected, DNA-seq SNPs were mostly distributed across the non-coding part of the genome (46% in intergenic regions, 52% in introns) and at a lower proportion (2%) in exonic regions. This distribution is expected since coding regions are generally under stronger selection pressure than non-coding regions (Zhao et al., 2003). With RNA-seq (all the samples being systematically treated with DNAse), we expected to find most of the SNPs in exonic regions, which represent the majority of expressed regions. However, the majority of the detected SNPs were located in intronic (61%) and intergenic (29%) regions. Higher SNP counts in intronic regions can be explained by the presence of unspliced transcripts (premature transcripts), very lowly expressed compared to spliced transcripts, but sufficiently to be supported by reads, and by the lower selection pressure on these regions compared to the exons. SNPs located in "intergenic regions" are likely to be located in new genes or in not yet annotated part of genes (particularly 3′UTR and 5′UTR). Within exons, the proportion of SNPs in 3′UTR, 5′UTR and CDS were similar between RNA-seq and DNA-seq (32, 7, 61%), but significantly different from the proportion of these regions in the genome (20, 5, 75%) showing a lower selection pressure in 3′UTR regions than in CDS regions.

SNP Detection by RNA-Seq: Concordance With Those Detected by DNA-Seq

We detected SNPs using either RNA-seq or DNA-seq data obtained from the liver of the same 15 laying hens (see population A in **Figure 3B**, left). We found 7,786,492 biallelic SNPs using the DNA-seq data filtered with the standard criteria of GATK (see section "Materials and Methods") and considered them as reliable. Using the RNA-seq data filtered with some of the filters

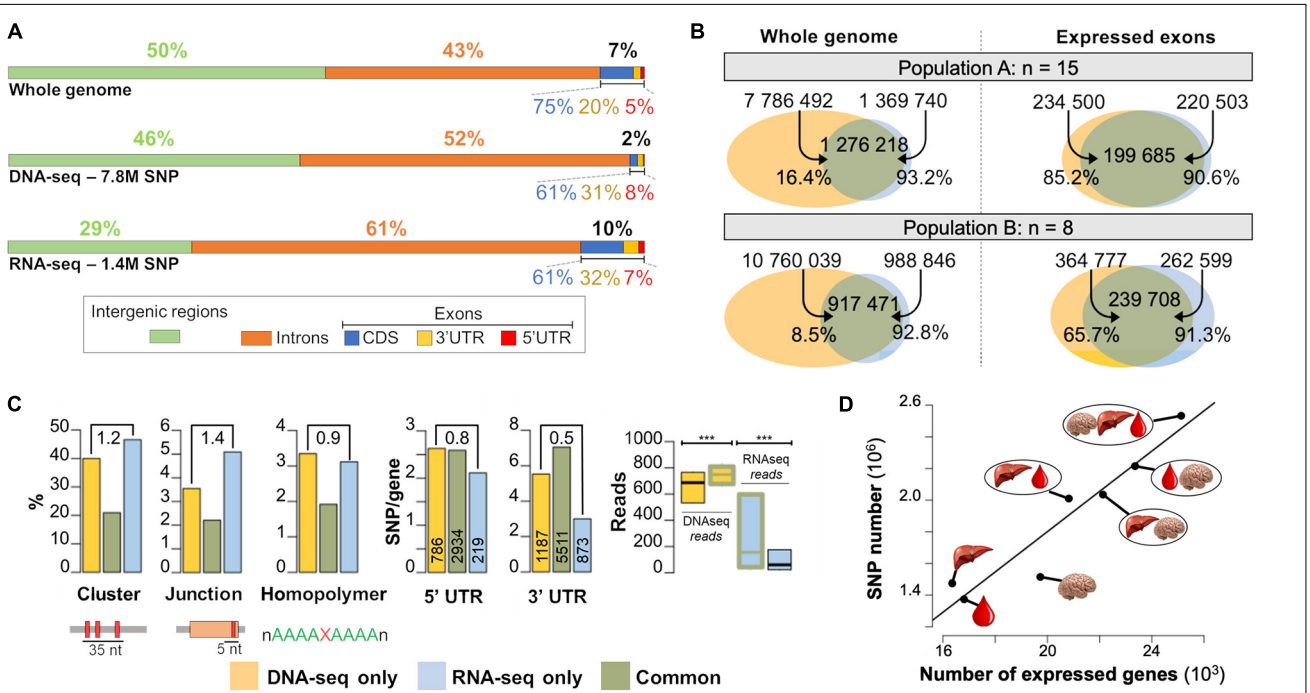


FIGURE 3 | Differences and common features of SNPs detected by RNA-seq and DNA-seq. **(A)** Percentage of the genome comprising each type of feature (top) and the proportion of SNPs detected by DNA-seq (middle) and RNA-seq (bottom) across these genomic features. **(B)** Number of SNPs detected by DNA-seq only (yellow set), RNA-seq only (blue set) and by both methods (gray set) at the whole genome level (left) and expressed exon level (right) in two independent populations A (*n* = 15, layers) and B (*n* = 8, broilers). **(C)** Left: Percentages of SNPs in SNP clusters (i.e., 3 or more SNPs in a sliding window of 35 bp, as per GATK definition), in junctions, homopolymers, in population A by DNA-seq only, RNA-seq only and both methods (common). Middle: Number of SNPs detected in 5′ and 3′UTR by gene (*Y*-axis) and the gene number (vertical numbers) in population A. The ratio "RNA-seq specific SNPs/DNA-seq specific SNPs" is indicated at the top of each plot. Right: read supporting SNP distribution at the population level in DNA-seq or RNA-seq data. **(D)** Evolution of the number of detected SNPs as a function of the number of expressed genes using one tissue alone or groups of tissues. Tissues used were liver (figured as a liver), blood (figured as a blood drop) and hypothalamus (figured as a brain).

suggested by GATK (see section "Materials and Methods" and comments below), we found 1,369,740 SNPs. As expected, the number of SNPs detected with RNA-seq is much lower than that in DNA-seq, because only variants present in transcribed regions were detected. Note that the impact of all these filters on the SNP number was provided in the Additional File 3 for DNA-seq and RNA-seq and was quite low, more than 98% of SNP were kept after filtering whatever the population.

To provide a meaningful comparison of both methods, we used the SNPs detected in expressed exons, assessed using RNA-seq with the metric described in section "Materials and Methods." We detected in population A 147,474 expressed exons among the 162,145 exons of the 16,814 expressed genes (on average 8.8 exons per gene). As shown in **Figure 3B** right, in these exons, 85.2% of the 234,500 SNPs detected by DNA-seq were also detected by RNA-seq. In population B, which was composed of only 8 broiler chickens, we found that 65.7% of the SNPs detected with DNA-seq in the expressed exons were also detected by RNA-seq. Assuming SNPs detected by DNA-seq represent the "truth," these percentages represent the sensitivity, or recall, of RNA-seq for SNP detection. This difference in RNA-seq sensitivity between populations A and B is likely due to the number of samples per population (15 versus 8), that affects the extent to which reads at each position are accumulated across the samples (see **Figure 1**).

Concerning the precision of RNA-seq, among the 220,503 SNPs detected by RNA-seq in population A, and the 262,599 SNPs from population B, 90.6 and 91.3%, respectively, were detected by DNA-seq 20X showing a reasonable precision of RNA-seq for the SNP detection. These results are consistent with the findings of Guo et al. (2017), who compared the percentage of SNPs detected using RNA-seq versus exome sequencing and found around 85% concordance. Regarding the 9.4% (20,818 SNPs) RNA-seq specific SNPs, we analyzed different factors that could underlie their detection to highlight those that should be treated with caution (**Figure 3C**) and verify these factors in DNA-seq variants set or in the set of variants called by both methods. We consider the SNPs detected by DNA-seq as true since DNA-seq are now routinely used for SNP detection with the well-proven GATK filters. First, we observed that a large proportion of RNA-seq specific SNPs (46.6%) and DNA-seq specific SNP (40.0%) belonged to a "SNP cluster" (i.e., 3 or more SNPs in a sliding window of 35 bp, as per GATK definition) (**Figure 3C**). This filter is one of the three filters proposed by GATK for RNA-seq SNP detection, but not for DNA-seq detection and the GATK team notes that these filters are not definitive and should be validated by users. Therefore, in the light of these observations, we decided not to remove the "SNP clusters" from our RNA-seq dataset as for DNA-seq dataset, but only to flag them as belonging to a so-called SNP cluster. Indeed, this filter removed 39,783 true SNPs (i.e., True positives detected by both DNA-seq and RNA-seq methods) and consequently the benefit of the precision increase (from 90.6 to 93.5) by removing "SNP clusters" was too small relatively to the recall decrease (from 0.85 to 0.68). The 20,818 RNA-seq specific SNPs can be explained by other factors of lowest impact: (*i*) 5.09% were located at 5 bp or less of an exon-exon junction, *versus* 3.55% for those detected only by DNA-seq; the corresponding ratio, that is significantly greater than 1 (1.4, $p \leq 10^{-17}$, χ^2 test), was expected since RNA-seq deals with spliced transcripts (**Figure 3C**) and therefore RNA-seq read mapping by the aligner is more complicated and more error-prone than DNA-seq read mapping. Since most of them are also observed in DNA-seq, we consider that the SNPs in the vicinity (i.e., 5 bp) of the junctions can be kept, but should be validated by another technique. Note that these SNPs represent only 0.48% of the total SNPs detected by RNA-seq. (*ii*) 3.1% were located in low complexity regions, defined as repetition of at least 5 identical nucleotides, versus 3.4% for the ones detected only by DNA-seq (**Figure 3C**). (*iii*) 2.7 and 5.5 SNPs per gene for RNA-specific SNPs were observed in 5′UTR and 3′UTR regions, respectively, with a fewer 3′UTR SNPs compared to those detected by DNA-seq only (0.5, $p \leq 10^{-16}$, χ^2 test) (**Figure 3C**). This may be due to the fact that mature transcripts undergo exonucleases action, degrading their 3′ extremities and causing their absence in RNA-seq libraries (Gallego Romero et al., 2014). (*iv*) Last, another factor that could be responsible for these RNA-seq specific SNPs is RNA editing, however, according to the literature, it is unlikely that most of the remaining SNPs are due to this mechanism. In mammals, in which RNA editing is well studied, Adenosine-to-Inosine (A-to-I) editing due to ADAR1 and ADAR2 enzymes is the most common editing form and mostly occur in inverted pairs of Alu interspersed repeats (Porath et al., 2014). In chicken *Alu*-like family of interspersed repeats also exist and they are called CR1 (Olofsson and Bernardi, 1983). These editing events tend to occur in clusters, a phenomenon called hyper-editing that introduces $\geq$ 20 mismatches in the sequencing reads (Carmi et al., 2011), that are therefore discarded by the aligner either because of a multi-mapping or no mapping. The prevalence of editing is still discussed: RNA editing is rarely detected when standard mapping filters are used, as shown in mice (Lagarrigue et al., 2013a) and chickens (Frésard et al., 2015; Roux et al., 2016; Shafiei et al., 2019), with less than 200 events, and in humans (Kleinman et al., 2012; Tan et al., 2017) with less than 1000 events per tissue. By contrast, RNA editing is frequently detected when working in repeated regions and rescuing unaligned reads (Picardi et al., 2017). Finally, we observed that SNPs detected only by one method were supported by significantly less reads (either of RNA- or DNA-seq) than the SNPs detected by both methods (**Figure 3C**).

SNP Detection by RNA-Seq: Impact of the Number of Tissues That Are Analyzed

Using blood and hypothalamus samples collected on the same 15 animals (population A), we studied the effect of detecting the SNPs in more than one tissue. RNA-seq from each tissue was not generated at the same time and have been analyzed separately at different occasions. Results are displayed in **Figure 3D**. We detected 1,369,740 SNPs in the liver (as previously stated), 1,481,627 in the blood and 1,511,909 in the hypothalamus, while 16,814 genes were expressed in liver, 16,346 in blood, and 19,733 in hypothalamus. As expected, using combinations of two or three tissues, the number of detected SNPs increased in relation with the number of expressed genes (spearman correlation = 0.96, $p = 3 \times 10^{-3}$) by cumulating the information on all tissues

in which a same gene is more or less expressed. Note that here, we have used our pipeline in a sub-optimal manner, by analysing RNA-seq data per tissue instead of combining the tissues together to increase detection power and reliability. For projects in which RNA-seq from different tissues per animal are all available before SNP detection analysis, we advise users to pool for each animal the RNA-seq files. For SNPs detected in more than one tissue, the concordance between genotypes detected in different tissues was very high, 98.9% without read filtering. Considering genotypes supported by at least 5 reads (respectively 10 reads) the concordance raised to 99.5% (respectively 99.9%).

Genotype (GT) Calling by RNA-Seq Importance of Genotype Call Rate (CR) and Read Depth at the Individual Scale for Selecting SNPs With Enough Reliable Genotypes for *in fine* Calculating Genotype and Allele Frequencies

While reliable SNPs can be detected in the population thanks to some individuals that bear them, it does not necessarily mean that there are enough reads for each individual to produce a genotype (GT). This was exemplified in **Figure 1B** by the brown cells (SNPs 3 and 4), for individuals 4 and 5 ("stress" group) for SNP 3 or most of the individuals of the population for SNP 4. These cases are quite frequent in practice because of gene expression variability between individuals in a given tissue, especially when different conditions are analyzed or also when a SNP is located in an intron of an immature transcript (weakly abundant compared to the mature transcript). Therefore, genotype call rate (CR), defined as the percentage of individuals with a genotype in the population, can be highly variable (e.g., from 16 to 100% in **Figure 1B**, right) from one SNP to another, depending on the number of reads observed in each individual (DP per individual). With 20X DNA-seq data, most of the SNP have a genotype CR close to 100%, as depicted in **Figure 1A**.

These observations indicate that a genotype can be observed with a certain call-rate but its reliability will depend on the DP supporting it. The GT reliability was estimated by the genotype concordance between RNA-seq and DNA-seq, assuming that GT detected by DNA-seq represents the truth. This concordance corresponds to the precision of RNA-seq for GT calling. We tested the RNA-seq precision according to different criteria. First, we conjointly studied in **Figure 4A** the effects of the criteria "genotype CR" and "DP supporting the genotype" on the RNA-seq precision (genotype concordance between RNA-seq and DNA-seq). We found a concordance (of roughly 90%) when no threshold was applied on the DP (purple line); it increased to around 95% for a CR $\geq$ 20% with a DP $\geq$ 5 reads and over 97% for a CR $\geq$ 20% with a DP $\geq$ 10 reads. We then evaluated the impact of the CR alone (without a DP threshold, *x*-axis) versus the CR with a DP $\geq$ 5 reads (*y*-axis), on the genotype concordance between RNA-seq and DNA-seq (solid green isoclines) and on the number of SNPs selected according to the different criteria (dashed blue isoclines) (**Figure 4B**).

Interestingly, only the CR with DP $\geq$ 5 reads have an effect on the genotype concordance and the percentage of selected SNPs, while no such effect is observed for the no DP filtering CR (*x*-axis) comprised between 0 and 50%, as shown by the horizontal isoclines. Hence, we propose for our subsequent analysis on different RNA-seq datasets to select SNPs within the red surface of **Figure 4B** with a (5.reads.DP) genotype CR $\geq$ 20% ensuring a concordance (precision) of almost 95% and a CR $\geq$ 50% ensuring a sufficient number of GT per SNP to calculate the allelic frequencies. We can note that most of the SNPs on this surface have a genotype concordance of more than 97%. We can also note in most of the populations analyzed in the next section that more than 98% of SNPs with (5.reads.DP) genotype CR $\geq$ 20% have a CR $\geq$ 50% (Additional File 4).

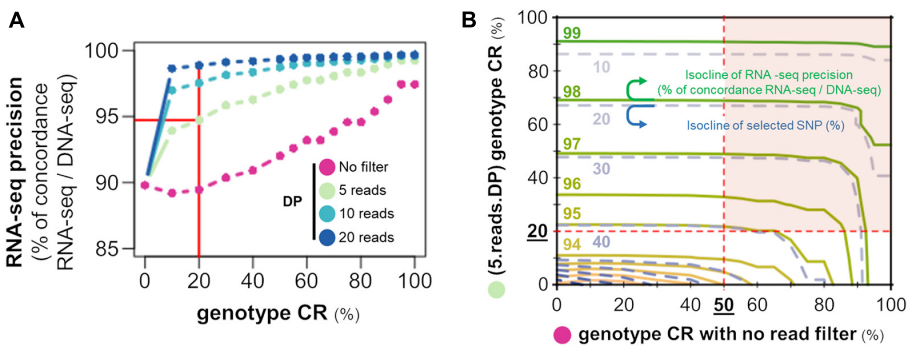


FIGURE 4 | The passage from SNP to GT necessitates a read depth threshold. **(A)** Evolution of the percentage of genotype concordance between RNA-seq and DNA-seq (*y*-axis) for the 15 RpRm birds as a function of genotype call rate in the population (CR: *x*-axis) supported by at least 5 (light green), 10 (light blue), or 20 (dark blue) reads or without read filter (purple curve). The red lines correspond to the criteria used in the further analysis (CR $\geq$ 20% with a DP $\geq$ 5 reads) and the corresponding RNA-seq precision. **(B)** Isoclines of the percentage of genotype concordance between RNA-seq and DNA-seq (solid green lines) and of the percentage of SNPs selected out to the original set (dashed blue lines) according to the CR with no read filter (*x*-axis) and the CR with at least 5 reads [(5.reads.DP) genotype CR (%), *y*-axis]. Red surface: SNPs selected after filtering on (5.reads.DP) genotype CR $\geq$ 20% and a CR $\geq$ 50%.

Number of SNPs and Genotypes Detected by RNA-Seq in 11 Populations

As shown in **Table 1** which gives an overview of the SNP diversity in 11 chicken populations, we detected between 1.1 and 3.8 M SNPs per population using liver RNA-seq datasets. Using all the tissues available (1–5 tissues depending on the population), we detected more SNPs, consistently with our previous result (see **Figure 3D**): between 1.7 and 5.5 M SNPs with a fold increase of × 1.18 to × 2.48 depending on the number and nature of analyzed tissues. Across populations and using all tissues, we found a total of 9.5 M SNPs having at least one alternative allele in at least one population (SNP union), and 241,960 SNPs that had at least one alternative allele in each of the 11 populations (SNP intersection). The union of our SNPs contains 23% (2,175,528) yet-unreported SNPs in the reference Ensembl v94 dbSNP database [(Ensembl, 2018): 21 M SNPs]. The intersection of our SNPs contains 5.1% (12,203 SNPs) of the SNPs present in the 600K genotyping array (Kranis et al., 2013).

We then filtered SNPs on genotype call rate and read depth (**Table 1**, "Selected GT") and found between around 0.4 and 1.7 M SNPs using all tissues, 37% of the SNPs observed previously. These results on 11 populations show that a large number of SNPs (two thirds) were detected at the population level thanks to the accumulation of reads across all individuals of the population, but that within each individual, read counts are not sufficient to reliably determine a genotype. Nevertheless, the number of SNPs with a genotype per population remains in the order of magnitude of several hundred thousand to a few millions with a union of 3.3 M and an intersection of 73,223 SNPs. In the liver, for which data was available in all but one population (Rmx6), the union and intersection are of the same order of magnitude: 1.7 M and 67,341 SNPs, respectively. After selecting for a MAF (minor allele frequency) $\geq$ 10% in order to discard rare SNPs or those resulting from sequencing errors, the number of SNPs was halved in all populations with a grand total of 2.2 and 1.3 M for the multi-tissue and liver union, respectively. As expected, the intersection drastically decreased to approximatively 2,000 SNPs, since this set corresponds to the SNPs with a MAF $\geq$ 10% in each of the 11 populations. The list of the 9.5 M of SNPs including 3.3 M with a GT and 2.2 M with MAF $\geq$ 10% is available on http://www.fragencode.org/lnchickenatlas.html.

Rare Deleterious Variants Detection in the Populations

We predicted the impacts of the 9,496,283 SNPs detected in at least one population using the VEP tool (McLaren et al., 2016) which predicts the potential consequences of the SNPs in each of the transcripts carrying them: we found 33,304,412 consequences. As expected, the vast majority of the SNPs affected non-coding regions (**Figure 5A**) and among the 472,319 SNPs affecting a coding-region, a majority were synonymous (63%) or non-deleterious missense (28%) as shown in **Figure 5B**.

Among all these predictions, we focused on the predicted consequences with the most severe putative impacts as defined by the gnomAD consortium, which only considers the PCG (Protein Coding Genes) (Genome Aggregation Database Consortium et al., 2020): variants in the splice regions, start and stop codon loss or stop codon gain even if the severity of the latter depends on its position in the coding sequence. We also added missense variants with a SIFT score $\leq$ 0.05. As reported by gnomAD (The GTEx Consortium, 2020), these SIFT-deleterious SNPs generally have a low frequency in the populations and can be mistaken for sequencing errors. Hence, it is crucial to select SNPs with genotypes (as defined previously) and a MAF $\geq$ 10% in at least one population (i.e., the ALT allele observed for example at least 4 times in a population of 16 individuals as for FAyo and LSnu populations) to make sure that the deleterious allele is not spurious. Thanks to our data from 382 individuals from the 11 populations, we listed a total of 25,344 strong predicted impacts (**Figure 5C**), corresponding to 14,496 SNPs and 67,58 genes, among them were 590 predictions of stop gained (404 genes), 8,126 of a coding or non-coding gene splice site change (donor and acceptor), 16,307 SIFT-predicted deleterious missenses and 321 other predictions (start lost, stop lost). Out of these 25,344 deleterious-predicted impacts, we found 5,654 (22%) predictions corresponding to 2,872 (20% of 14,496 SNPs) variants in 1,884 genes for which the homozygous ALT/ALT genotype was absent, in all populations in which the ALT allele was detected and, respectively, 7,740 (31%) predictions corresponding to 4,072 (28% of 14,496 SNPs) variants in 2,515 genes with ALT/ALT frequency $\leq$ 5%. The analysis of tolerated missense SNP show that the higher the SIFT score (i.e., tolerated variant), the lower the percentage of SNP with a low frequency ($\leq$ 5%) of ALT/ALT genotype (**Figure 5D**). The same analysis performed with 217,119 synonymous variants showed lower percentages with 9% SNPs with ALT/ALT genotype absent and 13% SNP with ALT/ALT frequency $\leq$ 5%. Such results are compatible with a homozygous state which is lethal or strongly negatively selected (28 versus 13%, $p \leq 10^{-20}$, χ^2 test), suggesting an important role for the genes associated to these variants with severe-predicted impact. Such variants obtained using RNA-seq data constitute a new complementary resource to Ensembl dbSNP allowing to explore variants (deleterious or not) according to their genotypic and allelic frequencies in different populations of a farm species. For example, two deleterious missense SNPs (SIFT-score = 0) are presented in **Figure 6**. One is already reported in dbSNP (Ensembl genome browser 94, 2020) and affects XBP1 protein by changing a positive charged amino acid (Arginine, R) into an aromatic and hydrophobic amino acid (Tryptophan, W) (**Figure 6A**). This SNP is observed in two of the ten analyzed populations, FLLL and Novo2, with 5 and 10 heterozygous birds among 48 and 40 animals analyzed, respectively, whereas no ALT/ALT homozygous birds were observed (**Figure 6B**). This gene is ubiquitously expressed in chicken as in human (**Figure 6C**). It codes the "Tax-Responsive Element-Binding Protein 5" transcription factor which has important cellular and physiological roles related to the "unfolded protein response" pathway in the endoplasm reticulum [(Lee et al., 2003) and for review (Glimcher et al., 2020)] and also to hepatic insulin resistance (Zhou et al., 2011).

The second SNP, not reported in dbSNP, affects the SERGEF protein (alias DelGEF) by changing an aromatic, hydrophobic and positive charged amino acid (Histidine, H) into an

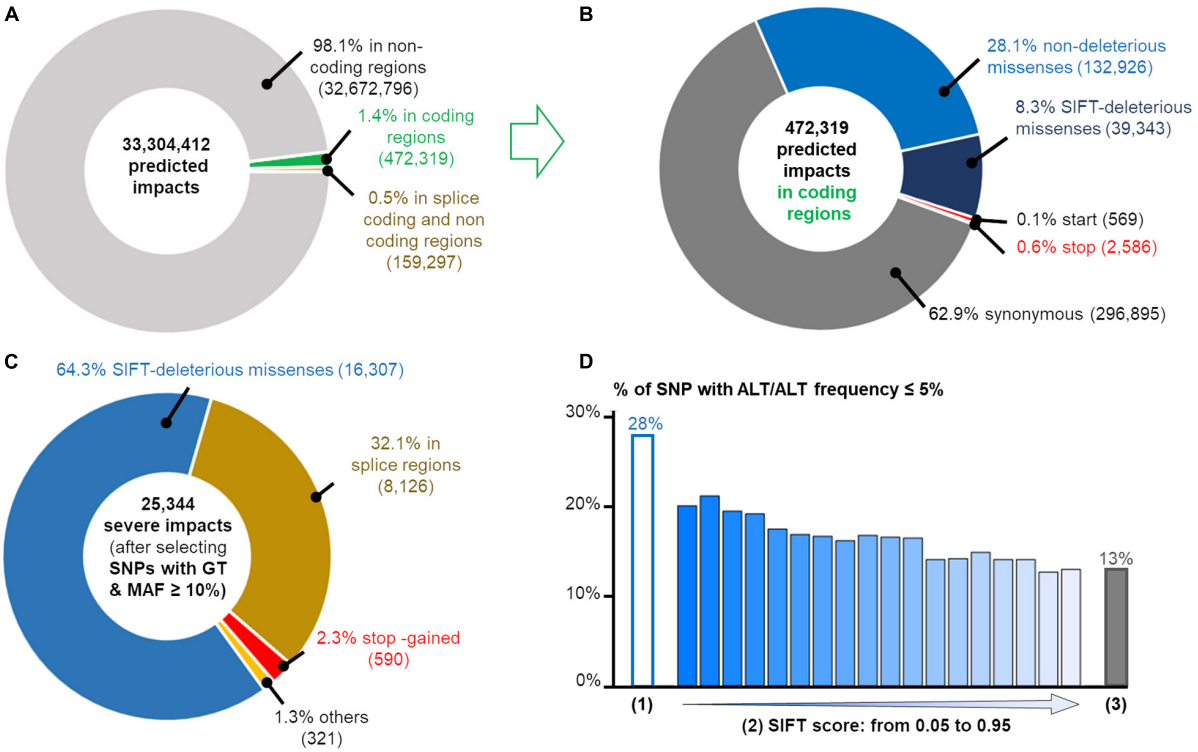


FIGURE 5 | Annotation of 9,496,283 SNPs using Variant effect Predictor (VEP) (McLaren et al., 2016). **(A)** Distribution of variant effect predictions among non-coding (light gray), splice regions related to coding and non-coding genes (orange) and coding (green) regions. **(B)** SNP annotation in coding regions: synonymous (dark gray), non-deleterious (light blue) and deleterious (blue) missenses, and consequences affecting stop (red) and start (orange) codons. Total number of consequences are indicated between parentheses. **(C)** Annotation of SNPs predicted as deleterious and filtered according to the GT criteria (as defined previously) and a MAF ≥ 10% in at least one population. **(D)** Percentage of SNP with a very low frequency (≤ 5%) of ALT/ALT genotype for three SNP sets: (1) = the 25,344 deleterious SNP described on the left; (2) = the tolerated SIFT-missense SNP according to the SIFT score and (3) = the synonymous SNP set. The splice sites correspond to the donor or acceptor splice sites of coding and long non-coding genes.

unchanged amino acid (Tyrosine, Y) (**Figure 6A**). This SNPs was observed in two populations, LSnu and Fayoumi, with 6 and 3 heterozygous birds among 16 animals, respectively, whereas no ALT/ALT homozygous birds were observed. This gene is also relatively ubiquitously expressed in chicken as in human (**Figure 6C**). The functions of this gene, which codes the "Secretion Regulating Guanine Nucleotide Exchange Factor" seem to be poorly known: 9 publications found in PubMed with the key words, SERGEF or DELGEF. As illustrated by these two examples (XBP1 and SERGEF), the analysis of various populations allowed to increase the number of rare deleterious variants detected.

Potential for Allele-Specific Expression Analysis in Various Populations

Allele-specific expression (ASE) analysis requires a heterozygous SNP in the expressed feature, to test an eventual imbalance in the expression between the two parental chromosomes. Usually, the expression is evaluated using RNA-seq and the SNPs are detected using DNA-seq, which is expensive when working on a dozen or more individuals. Since we have shown that RNA-seq allows detecting a large number of reliable SNPs in expressed regions, we studied in this section, the potential of RNA-seq data for performing ASE analysis. To this end, the **Figure 7A** provides the average numbers of genes across various populations, having at least one SNP with different filters (SNPs with an associated GT, a MAF ≥ 10% and an heterozygous status in at least 25% of the population). We also indicated the average SNP number per gene (column "S/g") to give an idea of the RNA-seq potential to test ASE along the gene. We indicated the results for two types of genes: the protein-coding genes (PCG) and the long non-coding genes (lncRNA), which are increasingly considered as important regulators of gene expression but are also known to be less expressed than PCG (Derrien et al., 2012; Muret et al., 2017; Le Béguec et al., 2018). This is the reason why we studied two expression thresholds: 0.1 and 1 TPM commonly used when working on lncRNA and PCG, respectively. Finally, results in **Figure 7A** are presented either for SNPs detected in exons (i.e., mature transcripts) (top) or for SNPs detected in exons or introns hence including immature transcripts (bottom).

The first key result is that the number of genes with at least one SNP are similar in both cases (exons only versus exons + introns), meaning that there are enough SNPs to study ASE in exonic regions only, i.e., mature transcript, despite a much lower number of SNPs per gene when SNPs are

A

gnName	chr	pos	ref	alt	SNPid	SIFTsc.	codon	aa	position
XBP1	15	7763534	G	A	rs738179721	0	Cgg/Tgg	R/W	105/255
SERGEF	5	12450547	G	A	NA	0	Cat/Tat	H/Y	217/298

B

gnName	pop.	# ind.	allele frequencies		genotype frequencies		
			ref	alt	ref/ref	ref/alt	alt/alt
XBP1	FLLL	48	0.95	0.05	0.9	0.1	0
XBP1	Novo2	40	0.875	0.125	0.75	0.25	0
SERGEF	LSnu	16	0.815	0.185	0.63	0.37	0
SERGEF	FAyo	16	0.905	0.095	0.81	0.19	0

C

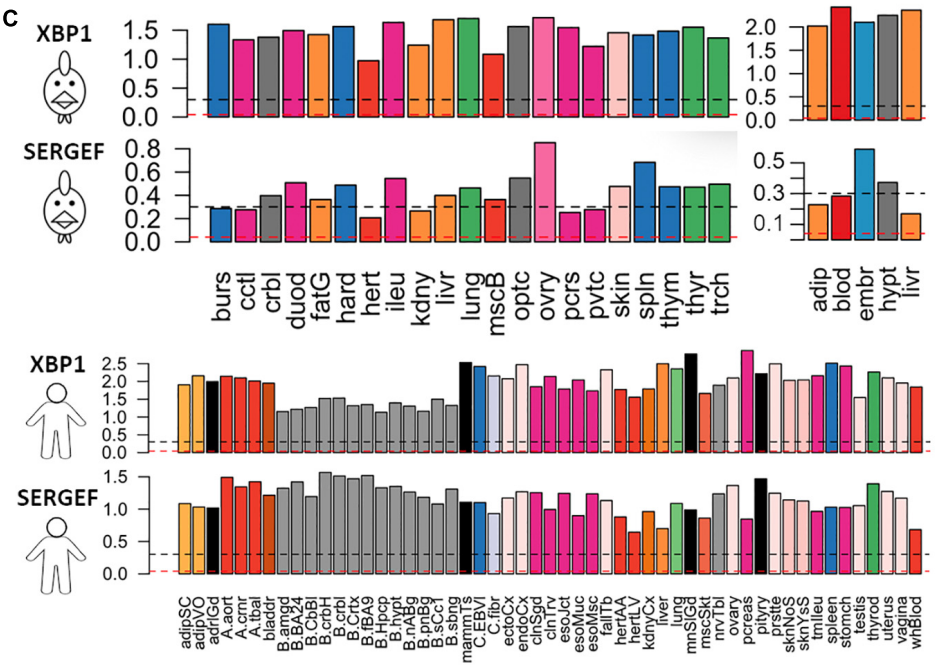


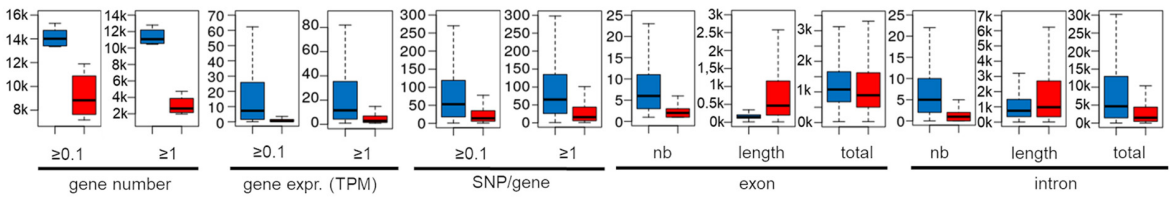
FIGURE 6 | Two examples of deleterious missense SNPs impacting two protein coding genes (XBP1 and SERGEF). **(A)** genomic position of the SNP with its identifier (SNPid) in Ensembl dbSNP and its impact on the protein with SIFTsc.: SIFT score, codon/modified codon, amino acid/modified amino acid and its position in the protein. **(B)** pop.: population with the individual size (# ind.) observed per population and the frequencies of the alleles and genotypes. **(C)** Tissue expressions [log10(TPM + 1)] in chicken using two datasets composed of 21 tissues (ERP014416) (left) and 5 tissues RpRm population) (right) and in human through the 53 tissues from the GTEX consortium (The GTEx Consortium, 2020). Abbreviations for the 21-tissue dataset: burs, bursa of Fabricius; cctl, cecal tonsils; crbl, cerebellum; duod, duodenum; fatG adipose tissue around the gizzard; hard, harderial gland; hert, heart; ileu, ileum; kdny, kidney; livr, liver; lung, lung; mscB breast muscle; optc, optical lobe; ovry, ovary; pcrs, pancreas; pvtc, proventriculus; skin, skin; spln, spleen; thym, thymus; thyr, thyroid gland; trch, trachea; and for the 5-tissue dataset: adip, abdominal adipose tissue; blod, blood; embr, 4.5 day embryos; hypt, hypothalamus; livr, liver; for more details in these 3 datasets and associated samples see Jehl et al. (2020). Black dashed line: gene expression with TPM $\geq$ 1 and red dashed line: TPM $\geq$ 0.1.

only selected in exons (**Figures 7A,B**). When working with exonic SNPs, there are on average 17–28 SNPs without filter (8–10 SNPs after all filters) per gene showing the possibility to test ASE along genes. Despite a lower exonic length in lncRNA compared to the PCG (**Figure 7B**), this number is higher for lncRNA compared to PCG (22–28 versus 15–17) probably due to lower selective pressure on lncRNA compared to PCG. The second key result, after applying 2 filters (GT and MAF $\geq$ 10%), is that 81% of PCG (9,232) and 68% of lncRNA (2,028) expressed at TPM $\geq$ 1 are analyzable for ASE. These numbers decreased a little after applying an additional filter related to the heterozygosity percentage, with 72% of PCG and 56% of lncRNA (i.e., about 10,000 genes). The variability of this "ASE analyzable genes" percentage is moderate (Additional File 5): on average 72% from 65 to 89% with an except for the "RpRm" (48%) probably due to its high consanguinity and its large size, the filter of 25% of heterozygosity impacting more the populations with a larger sample size. The same tendencies regarding the percentage of genes that can be analyzed were observed for the PCG (TPM $\geq$ 0.1) and for lncRNA (both for TPM $\geq$ 0.1 and $\geq$ 1) (Additional File 5). We can note that the selected lncRNA percentage satisfying the filters is always lower than the selected PCG percentage ($-$15% for genes with an expression $\geq$ 1TPM and $-$30% for genes with an expression $\geq$ 0.1TPM). This is mainly due to the lower expression of lncRNA compared to PCG (Jehl et al., 2020; **Figure 7B**), despite higher sequence variability for the former.

A

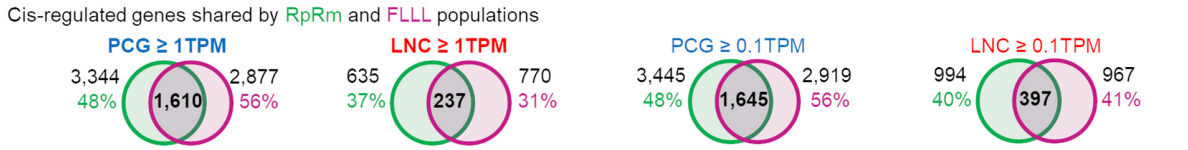
expr. threshold	biotype	expr. gene	no filter SNP	no filter gene	no filter S / g	no filter %	GT filter SNP	GT filter gene	GT filter S / g	GT filter %	MAF ≥ 10% SNP	MAF ≥ 10% gene	MAF ≥ 10% S / g	MAF ≥ 10% %	Het ≥ 25% SNP	Het ≥ 25% gene	Het ≥ 25% S / g	Het ≥ 25% %
SNP in exons																		
1 TPM	PCG	11 384	181 513	10 770	17	95	144 895	10 522	14	92	89 263	9 232	10	81	67 591	8 281	8	72
	LNC	2 982	72 040	2 608	28	88	46 401	2 365	20	79	27 727	2 028	14	68	17 756	1 701	10	56
0.1 TPM	PCG	14 183	212 598	13 116	16	92	153 716	11 560	13	82	94 251	10 063	9	71	70 063	8 792	8	62
	LNC	9 228	175 253	7 899	22	86	70 903	4 517	15	49	42 108	3 753	11	40	25 616	2 893	8	31
SNP in genes (i.e. exons + introns)																		
1 TPM	PCG	11 384	1 087 288	11 098	97,6	98	390 248	10 750	36	94	242 720	9 725	25	85	156 265	8 701	17	76
	LNC	2 982	110 267	2 670	41,7	90	59 006	2 403	24	81	35 356	2 080	17	69	22 408	1 756	12	58
0.1 TPM	PCG	14 183	1 166 864	13 541	85,6	96	403 222	11 825	34	84	250 173	10 586	23	75	159 993	9 246	17	65
	LNC	9 228	252 298	8 060	31,1	87	91 564	4 612	19	50	54 772	3 869	14	42	33 193	3 006	11	32

B



C Percentage of cis-regulated genes among the expressed genes in liver

expr. threshold	biotype	pop: RpRm expr.gene	Nb of individuals with ASE genes 2	3	4	pop: FLLL expr.gene	Nb of individuals with ASE genes 2	3	4	Mean 2
1 TPM	PCG	9748	34.3	26.2	21.0	11707	24.6	18.0	13.7	29.5
	LNC	1828	34.7	27.8	23.4	3528	21.8	15.7	11.7	28.3
0.1 TPM	PCG	11254	30.6	23.3	18.7	12178	24.0	17.6	13.3	27.3
	LNC	4265	23.3	18.1	15.1	5326	18.2	13.0	11.7	20.8



D

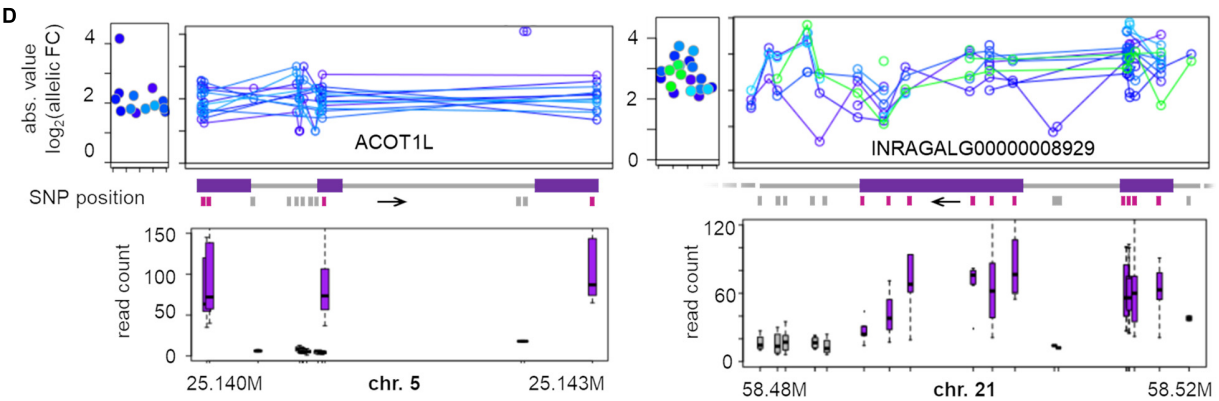


FIGURE 7 | Overview of the analyzable genes for allele-specific expression in the liver of various populations for two gene biotypes, Protein Coding Gene (PCG) and Long Non-Coding gene (lncRNA), at two gene expression thresholds (0.1 TPM and 1 TPM) and for 3 filters. **(A)** Average numbers for all populations analyzed here. These average numbers are provided for both PCG (blue) and lncRNA (red) biotype, with minimum expression of 0.1 or 1 TPM ("expr. threshold"), and considering only the SNPs in exons (top part) or in the whole gene, i.e., in both exons and introns (bottom part). **(B)** Feature of lncRNA and PCG. **(C)** Percentage of gene with a significant allele specific expression in two populations RpRm (in left) and FLLL (in right) in comparison to the expressed gene number. Venn diagrams provide the number of ASE genes (in at least 2 individuals) shared by RpRm and FLLL populations. **(D)** Overview of the ASE of ACOT1L (left) and INRAGALG00000008929 (right). For each ASE sample, absolute values of the log2 allelic fold-change are represented at the gene-level (left of the panels) and for each SNP located in the haplotype used by phASER (right). Boxplot of the read number associated to each SNP are represented (bottom), in purple for the SNP located in exons and in gray for those in introns. FC, fold-change; chr, chromosome.

Cis-Regulated Genes in the Liver of Two Populations

To provide an estimation of the number of *cis*-regulated genes in one tissue, we performed an ASE analysis of the liver samples of the RpRm and the FLLL populations using phASER and its downstream tool, phASER Gene AE, that phase SNPs at the gene level (see also section "Materials and Methods"). Using exonic and intronic SNPs and selecting genes having one haplotype with at least 10 reads, we found for genes with an hepatic expression ≥ 1 TPM, that in average 29%

of the expressed PCG or lncRNA genes were *cis*-regulated (∼34% for RpRm and ∼23% for FLLL) (**Figure 7C**). For lncRNA with hepatic expression ≥ 0.1 TPM which represents most of this biotype, we found a lower percentage of *cis*-regulated genes (21%) because they are less expressed and some of them did not have more than 10 reads for at least one "super-allele" analyzed by phASER (see section "Materials and Methods"). Interestingly, among these *cis*-regulated genes, ∼50% and 37% are shared by both populations for the protein-coding genes and long non-coding genes, respectively (**Figure 7C**). Two examples of *cis*-regulated genes are provided in **Figure 7D** with a PCG, ACOT1L (ENSGALG00000008752), and a lncRNA, INRAGALG000000089295. Overall, these numbers are consistent with the literature: Zhuo et al. (2017) found that 15% of the genes were *cis*-regulated in chicken embryo liver, and Lagarrigue et al. (2013b) found a similar number in mice liver. In humans, the GTEx consortium (The GTEx Consortium, 2020) found that 26% (4,415) of the expressed genes (17,243) were *cis*-regulated in the liver.

Diversity Exploration Using RNA-Seq Variants

Finally, we explored genetic links between populations using the genotypic frequencies of SNPs detected by RNA-seq, which represent a set of SNPs, which may be under a larger selective pressure than those used in genotyping SNP chips. Indeed, the latter are considered as having a neutral effect, while most the SNPs present in our data are located in expressed regions and affect proteins to some extent (from almost neutral synonymous to deleterious stop gained).

The classification in **Figure 8** was produced using the intersection of SNPs with GT of the 10 populations with a liver presented in **Table 1** (67,341 SNP set). This classification is consistent with the known chicken population history, indicating that these SNPs detected by RNA-seq and their associated genotypes allow distinguishing different populations. The classification separated clearly the RJFh (red circle arc with a Red Jungle Fowl population, used here to represent the "ancestral" population), then the broilers (blue circle arc), the brown-egg layers (dark green circle arc), and the cream- or white-egg layers (brown circle arc with Fayoumi breed and Fr-Ag population which is an experimental leghorn line). We also observed the expected sub-groups within these 3 types of populations: the commercial lines (Novo1 and Novo2 for the layers, Cobb and HerX for the broilers) separated from the experimental lines (RpRm for the brown-egg layers, FLLL for the broilers). Interestingly for these 2 last populations, this SNP set shows a clear distinction between two subpopulations that have been divergently selected for a specific trait: Rp and Rm divergent for the residual feed intake and FL and LL divergent for body fat whereas the two Novogen populations (Novo1 and Novo2) are not distinct. We can note that the SNPs predicted as "missense" by VEP and "deleterious" by SIFT provide the same classification between the populations as the one shown in **Figure 8** (data not shown).

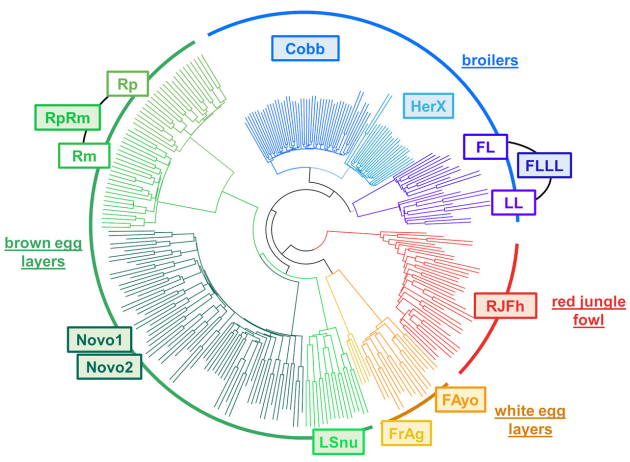


FIGURE 8 | Hierarchical clustering of 10 chicken populations using the 67,341 SNP intersection set with GT obtained using liver RNA-seq data. The hierarchical clustering was performed using the "snpgdsHCluster" from the package SNPRelate v1.8.0 (see also section "Materials and Methods").

CONCLUSION

We show here that RNA-seq data, which are cheaper to generate and store compared to DNA-seq data, can be a reliable resource for performing different analyses based on polymorphism detection. By comparing DNA-seq and RNA–seq results generated from the same animals in two independent chicken populations, this study provides a workflow to produce reliable SNPs and genotypes from RNA–seq data. We ran through this pipeline 767 RNA–seq of 382 birds from 11 populations and provided a per-population estimation of the average genotyped SNPs count per tissue (more than 550,000) and an overview of the predicted consequences of SNPs located in coding regions. In particular, thanks to this large RNA-seq dataset, we identified 440 genes containing a stop-gained impact, known to be rare because of their potentially severe impact, especially when located in the first third of the coding sequences (133 genes). In a companion study (Degalez et al., submitted), we checked the possible existence of more than one SNP in a given codon, that could "rescue" a stop-gained situation. We then gave an overview across 11 populations of genes that could be analyzed for ASE, i.e., having at least one SNP allowing to distinguish expression from both chromosomes. We applied phASER on liver RNA-seq data of two populations and identified around 21 to 30% of *cis*-regulated genes depending on the analyzed population and the gene biotype (PCG versus lncRNA), these results were consistent with other studies conducted in other species.

This study represents a first step to more ambitious projects that could analyze tens of thousands of available RNA-seq datasets to build a GTEx-like atlas reporting *cis*- and *trans*- genetic associations with gene expression, as previously performed in human (The GTEx Consortium, 2020) and more recently in cattle (Liu et al., 2020).

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

Ethical review and approval was not required for the animal study because all data have been obtained from public databases as described in the materials and methods, section "DNA-Seq and RNA-Seq Data."

AUTHOR CONTRIBUTIONS

CK and SLa conceived the study and coordinated the study. FJ, LL, CD, BA, MT-B, BB, TB, DG, HA, SF, SD, EG, FP, TZ, and SLa participated to the set-up of the experimental design and sample collection. LL, CD, and SLe carried out all RNA extractions. OB generated RNA-seq libraries and sequencing. FJ, FD, MB, CK, and SLa performed bioinformatic processing of the RNA-seq data and carried out the whole analyses. FL, PB, and MC participated to bioinformatic analyses. FJ, FD, and SLa drafted the manuscript. MB, FL, BA, MT-B, BB, HA, SF, SD, EG, TZ, FP, and CK helped to improve the manuscript. All authors read and approved the final version.

FUNDING

The sample and/or data were collected in the frame of projects that received financial support from the European Union's H2020 Program under Grant Agreement No. 633531 (Feed-a-Gene project), the French National Agency of Research (FatInteger project ANR-11-SVS7; ChickStress project, ANR-13-ADAP; EpiBird ANR project PCS-09-GENM-010), from French institutions as Institut Agro-AGROCAMPUS OUEST and INRAE [Fr-AgENCODE project (2015-2017), ELASETIC project (2012)]. This work was also supported by France Génomique National infrastructure, funded as part of "Investissement d'avenir" program managed by Agence Nationale pour la Recherche (contrat ANR-10-INBS-09). FJ and FD are Ph.D. fellows supported by the Brittany region (France) and the INRAE (Animal Genetics Division). These funding bodies had no role in the design of the study, in the collection, analysis, and interpretation of data, or in writing the manuscript.

ACKNOWLEDGMENTS

We are grateful to the GeT-PlaGe platform (doi: 10.15454/1.5572370921303193E12) for RNA-seq libraries and sequencing and the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, doi: 10.15454/1.5572369328961167E12) for providing help and/or computing and/or storage resources. A warm thanks to Morgane Boutin who passed away this year and who did a lot in RNA extraction.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.655707/full#supplementary-material

Additional File 1 | Detail per population of the number of birds, tissues and samples.

Additional File 2 | Boxplot of the exon expression in log10 (RpKb+1) for both RpRm and FLLL populations compared to noise. Noise corresponds to the expression of a set of randomly selected loci in the genome.

Additional File 3 | Effect of GATK filters on SNP numbers with RNA-seq and DNA-seq data. For RNA-seq, the analyses were performed in different tissues and populations. For DNA-seq, the analyses were performed for the two populations, RpRm with 15 individuals and FLLL with 8 individuals. Abbreviations for the tissue: adip: abdominal adipose tissue, blod: blood, embr: embryos, hypt: hypothalamus, livr: liver, spln: spleen. "rawSNP": SNP detected before the filters' application. "filtered biallelic SNP": SNP detected after the filters' application.

Additional File 4 | SNP counts per population with each genotype filters independently.

Additional File 5 | Overview of the analyzable genes for allele-specific expression in the liver of various populations for Protein Coding Gene (PCG) and Long Non-Coding gene (lncRNA), at two gene expression thresholds (0.1 TPM and 1 TPM) with the mean across the populations ("mean") and the coefficient variation ("CV"), i.e., standard deviation divided by mean. "expr. gene": expressed genes, "gene": genes bearing the SNP, "S/g": number of SNPs divided by the number of gene, "%": proportion of genes bearing the SNPs relative to the expressed genes. The 4 filters: no filter, GT satisfying our criteria ("GT filter"), with MAF $\geq$ 10% ("MAF $\geq$ 10%") and heterozygosity rate $\geq$ 25% ("Het. $\geq$ 25%").

REFERENCES

Adetunji, M. O., Lamont, S. J., Abasht, B., and Schmidt, C. J. (2019). Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data. *PLoS One* 14:e0216838. doi: 10.1371/journal.pone.0216838

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., et al. (2013). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 24, 14–24. doi: 10.1101/gr.155192.113

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* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bordas, A., Tixier-Boichard, M., and Merat, P. (1992). Direct and correlated responses to divergent selection for residual food intake in Rhode island red laying hens. *Br. Poult. Sci.* 33, 741–754. doi: 10.1080/00071669208417515

Carmi, S., Borukhov, I., and Levanon, E. Y. (2011). Identification of widespread ultra-edited human RNAs. *PLoS Genet.* 7:e1002317. doi: 10.1371/journal.pgen.1002317

Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y., and Lappalainen, T. (2016). Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun.* 7:12817.

Chamberlain, A. J., Vander Jagt, C. J., Hayes, B. J., Khansefid, M., Marett, L. C., Millen, C. A., et al. (2015). Extensive variation between tissues in allele specific

expression in an outbred mammal. *BMC Genomics* 16:993. doi: 10.1186/s12864-015-2174-0

Cornwell, M., Vangala, M., Taing, L., Herbert, Z., Köster, J., Li, B., et al. (2018). VIPER: visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinformatics* 19:135. doi: 10.1186/s12859-018-2139-9

Deelen, P., Zhernakova, D. V., de Haan, M., van der Sijde, M., Bonder, M. J., Karjalainen, J., et al. (2015). Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* 7:30.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Ensembl (2018). *dbSNP - Gallus Gallus 5 - V94*.

Frésard, L., Leroux, S., Roux, P.-F., Klopp, C., Fabre, S., Esquerré, D., et al. (2015). Genome-wide characterization of RNA editing in chicken embryos reveals common features among vertebrates. *PLoS One* 10:e0126776. doi: 10.1371/journal.pone.0126776

Frésard, L., Leroux, S., Servin, B., Gourichon, D., Dehais, P., Cristobal, M. S., et al. (2014). Transcriptome-wide investigation of genomic imprinting in chicken. *Nucleic Acids Res.* 42, 3768–3782. doi: 10.1093/nar/gkt1390

Fu, W., Dekkers, J. C., Lee, W. R., and Abasht, B. (2015). Linkage disequilibrium in crossbred and pure line chickens. *Genet. Select. Evol.* 47:11. doi: 10.1186/s12711-015-0098-4

Gallego, Romero I, Pai, A. A., Tung, J., and Gilad, Y. (2014). RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.* 12:42. doi: 10.1186/1741-7007-12-42

Genome Aggregation Database Consortium Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443.

GitLab (2019). *Snakemake/1000RNASeq_chicken/calling · master · bios4biol / workflows*. Available online at: https://forgemia.inra.fr/bios4biol/workflows/tree/master/Snakemake/1000RNASeq_chicken/calling (accessed August 28, 2019).

Glimcher, L. H., Lee, A.-H., and Iwakoshi, N. N. (2020). XBP-1 and the unfolded protein response (UPR). *Nat. Immunol.* 21, 963–965. doi: 10.1038/s41590-020-0708-3

Gondret, F., Vincent, A., Houée-Bigot, M., Siegel, A., Lagarrigue, S., Causeur, D., et al. (2017). A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. *BMC Genom.* 18:244. doi: 10.1186/s12864-017-3639-0

Guo, Y., Zhao, S., Sheng, Q., Samuels, D. C., and Shyr, Y. (2017). The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. *BMC Genom.* 18:690. doi: 10.1186/s12864-017-4022-x

Jehl, F., Désert, C., Klopp, C., Brenet, M., Rau, A., Leroux, S., et al. (2019). Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach. *BMC Genom.* 20:1033. doi: 10.1186/s12864-019-6384-8

Jehl, F., Muret, K., Bernard, M., Boutin, M., Lagoutte, L., Désert, C., et al. (2020). An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. *Sci. Rep.* 10:20457.

Kleinman, C. L., Adoue, V., and Majewski, J. (2012). RNA editing of protein sequences: a rare event in human transcriptomes. *RNA* 18, 1586–1596. doi: 10.1261/rna.033233.112

Koster, J., and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. doi: 10.1093/bioinformatics/bts480

Kranis, A., Gheyas, A. A., Boschiero, C., Turner, F., Yu, L., Smith, S., et al. (2013). Development of a high density 600K SNP genotyping array for chicken. *BMC Genom.* 14:59. doi: 10.1186/1471-2164-14-59

Krueger, F. (2021). *FelixKrueger/TrimGalore*. Available online at: https://github.com/FelixKrueger/TrimGalore (accessed March 15, 2021).

Lagarrigue, S., Hormozdiari, F., Martin, L. J., Lecerf, F., Hasin, Y., Rau, C., et al. (2013a). Limited RNA editing in exons of mouse liver and adipose. *Genetics* 193, 1107–1115. doi: 10.1534/genetics.112.149054

Lagarrigue, S., Martin, L., Hormozdiari, F., Roux, P.-F., Pan, C., van Nas, A., et al. (2013b). Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis -eQTL identified using genetic linkage. *Genetics* 195, 1157–1166. doi: 10.1534/genetics.113.153882

Le Béguec, C., Wucher, V., Lagoutte, L., Cadieu, E., Botherel, N., Hédan, B., et al. (2018). Characterisation and functional predictions of canine long non-coding RNAs. *Sci. Rep.* 8:13444.

Lee, A.-H., Iwakoshi, N. N., and Glimcher, L. H. (2003). XBP-1 regulates a subset of endoplasmic reticulum resident chaperone genes in the unfolded protein response. *Mol. Cell Biol.* 23, 7448–7459. doi: 10.1128/mcb.23.21.7448-7459.2003

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291.

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12:323. doi: 10.1186/1471-2105-12-323

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [Preprint]. Available online at: http://arxiv.org/abs/1303.3997 (accessed April 6, 2021).

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Liu, S., Gao, Y., Canela-Xandri, O., Wang, S., Yu, Y., Cai, W., et al. (2020). A comprehensive catalogue of regulatory variants in the cattle transcriptome. *bioRxiv* [Preprint]. doi: 10.1101/2020.12.01.406280v1

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

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.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., et al. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–777. doi: 10.1038/nature08903

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Muret, K., Klopp, C., Wucher, V., Esquerré, D., Legeai, F., Lecerf, F., et al. (2017). Long noncoding RNA repertoire in chicken liver and adipose tissue. *Genet. Sel. Evol.* 49:6.

Ng, P. C. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Oikkonen, L., and Lise, S. (2017). Making the most of RNA-seq: pre-processing sequencing data with Opossum for reliable SNP variant detection. *Wellcome Open Res.* 2:6. doi: 10.12688/wellcomeopenres.10501.2

Olofsson, B., and Bernardi, G. (1983). The distribution of CR1, an Alu-like family of interspersed repeats, in the chicken genome. *Biochim. Biophys. Acta Gene Struct. Express.* 740, 339–341. doi: 10.1016/0167-4781(83)90143-4

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259

Peng, Z., Cheng, Y., Tan, B. C.-M., Kang, L., Tian, Z., Zhu, Y., et al. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.* 30, 253–260. doi: 10.1038/nbt.2122

Picardi, E., D'Erchia, A. M., Lo Giudice, C., and Pesole, G. (2017). REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* 45, D750–D757.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–772. doi: 10.1038/nature08872

Piskol, R., Ramaswami, G., and Li, J. B. (2013). Reliable identification of genomic variants from RNA-Seq data. *Am. J. Hum. Genet.* 93, 641–651. doi: 10.1016/j.ajhg.2013.08.008

Porath, H. T., Carmi, S., and Levanon, E. Y. (2014). A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat. Commun.* 5:4726.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Quinn, E. M., Cormican, P., Kenny, E. M., Hill, M., Anney, R., Gill, M., et al. (2013). Development of strategies for SNP detection in RNA-Seq data: application to lymphoblastoid cell lines and evaluation using 1000 genomes data. *PLoS One* 8:e58815. doi: 10.1371/journal.pone.0058815

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Roux, P.-F., Boitard, S., Blum, Y., Parks, B., Montagner, A., Mouisel, E., et al. (2015). Combined QTL and selective sweep mappings with coding SNP annotation and cis-eQTL analysis revealed PARK2 and JAG2 as new candidate genes for adiposity regulation. *G3 Genes Genomes Genet.* 5, 517–529. doi: 10.1534/g3.115.016865

Roux, P.-F., Frésard, L., Boutin, M., Leroux, S., Klopp, C., Djari, A., et al. (2016). The extent of mRNA editing is limited in chicken liver and adipose, but impacted by tissular context, genotype, age, and feeding as exemplified with a conserved edited site in COG3. *G3* 6, 321–335. doi: 10.1534/g3.115.022251

Savary, C., Kim, A., Lespagnol, A., Gandemer, V., Pellier, I., Andrieu, C., et al. (2020). Depicting the genetic architecture of pediatric cancers through an integrative gene network approach. *Sci. Rep.* 10:1224.

Shafiei, H., Bakhtiarizadeh, M. R., and Salehi, A. (2019). Large-scale potential RNA editing profiling in different adult chicken tissues. *Anim. Genet.* 50, 460–474. doi: 10.1111/age.12818

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40, W452–W457.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15, 121–132. doi: 10.1038/nrg3642

Tan, M. H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A. N., et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. *Nature* 550, 249–254.

Tang, X., Baheti, S., Shameer, K., Thompson, K. J., Wills, Q., Niu, N., et al. (2014). The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. *Nucleic Acids Res.* 42:e172. doi: 10.1093/nar/gku1005

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

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* 43, 11.10.1–11.10.33.

Wang, C., Davila, J. I., Baheti, S., Bhagwate, A. V., Wang, X., Kocher, J.-P. A., et al. (2014). RVboost: RNA-seq variants prioritization using a boosting method. *Bioinformatics* 30, 3414–3416. doi: 10.1093/bioinformatics/btu577

Wolfien, M., Rimmbach, C., Schmitz, U., Jung, J. J., Krebs, S., Steinhoff, G., et al. (2016). TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. *BMC Bioinformatics* 17:21. doi: 10.1186/s12859-015-0873-9

Zhao, Z., Fu, Y.-X., Hewett-Emmett, D., and Boerwinkle, E. (2003). Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene* 312, 207–213. doi: 10.1016/s0378-1119(03)00670-x

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi: 10.1093/bioinformatics/bts606

Zhou, Y., Lee, J., Reno, C. M., Sun, C., Park, S. W., Chung, J., et al. (2011). Regulation of glucose homeostasis through a XBP-1-FoxO1 interaction. *Nat. Med.* 17, 356–365. doi: 10.1038/nm.2293

Zhuo, Z., Lamont, S. J., and Abasht, B. (2017). RNA seq analyses identify frequent allele specific expression and no evidence of genomic imprinting in specific embryonic tissues of chicken. *Sci. Rep.* 7:11944.

Ensembl genome browser 94 (2020). rs738179721 (SNP) - Explore this variant - Gallus gallus 5 - Archive Ensembl. Available online at: http://oct2018.archive.ensembl.org/Gallus_gallus/Variation/Explore?r=15:7763034-7764034;v=rs738179721;vdb=variation;vf=16501295 (accessed January 5, 2021).

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.

Copyright © 2021 Jehl, Degalez, Bernard, Lecerf, Lagoutte, Désert, Coulée, Bouchez, Leroux, Abasht, Tixier-Boichard, Bed'hom, Burlot, Gourichon, Bardou, Acloque, Foissac, Djebali, Giuffra, Zerjal, Pitel, Klopp and Lagarrigue. 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: 30 June 2021
doi: 10.3389/fgene.2021.664424

Transcriptomes of an Array of Chicken Ovary, Intestinal, and Immune Cells and Tissues

Eliah G. Overbey[1†], Theros T. Ng[2†], Pietro Catini[2], Lisa M. Griggs[2], Paul Stewart[2], Suzana Tkalcic[2], R. David Hawkins[3] and Yvonne Drechsler[2*]*

[1] Department of Genome Sciences, Interdepartmental Astrobiology Program, University of Washington, Seattle, WA, United States, [2] College of Veterinary Medicine, Western University of Health Sciences, Pomona, CA, United States, [3] Department of Genome Sciences, Department of Medicine, University of Washington, Seattle, WA, United States

OPEN ACCESS

Edited by:
Elisabetta Giuffra,
Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), France

Reviewed by:
Ole Madsen,
Wageningen University and Research, Netherlands
Herve Acloque,
Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), France

***Correspondence:**
R. David Hawkins
rdhawk@uw.edu
Yvonne Drechsler
ydrechsler@westernu.edu

[†] 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: 05 February 2021
Accepted: 27 May 2021
Published: 30 June 2021

Citation:
Overbey EG, Ng TT, Catini P, Griggs LM, Stewart P, Tkalcic S, Hawkins RD and Drechsler Y (2021) Transcriptomes of an Array of Chicken Ovary, Intestinal, and Immune Cells and Tissues. Front. Genet. 12:664424. doi: 10.3389/fgene.2021.664424

While the chicken (*Gallus gallus*) is the most consumed agricultural animal worldwide, the chicken transcriptome remains understudied. We have characterized the transcriptome of 10 cell and tissue types from the chicken using RNA-seq, spanning intestinal tissues (ileum, jejunum, proximal cecum), immune cells (B cells, bursa, macrophages, monocytes, spleen T cells, thymus), and reproductive tissue (ovary). We detected 17,872 genes and 24,812 transcripts across all cell and tissue types, representing 73% and 63% of the current gene annotation, respectively. Further quantification of RNA transcript biotypes revealed protein-coding and lncRNAs specific to an individual cell/tissue type. Each cell/tissue type also has an average of around 1.2 isoforms per gene, however, they all have at least one gene with at least 11 isoforms. Differential expression analysis revealed a large number of differentially expressed genes between tissues of the same category (immune and intestinal). Many of these differentially expressed genes in immune cells were involved in cellular processes relating to differentiation and cell metabolism as well as basic functions of immune cells such as cell adhesion and signal transduction. The differential expressed genes of the different segments of the chicken intestine (jejunum, ileum, proximal cecum) correlated to the metabolic processes in nutrient digestion and absorption. These data should provide a valuable resource in understanding the chicken genome.

Keywords: transcriptome, chicken, reproduction, intestinal cells, immunology

INTRODUCTION

In the United States, over nine billion broiler chickens, which is estimated to be about 19 billion kilograms of chicken products, are produced per year (NCC, 2019). Egg production totaled about 99.1 billion in 2019 in the United States (UEP, 2019). Apart from the important role in food production, the chicken has been used as an animal model to benefit key areas in functional human research including immunology (Glick et al., 1956), vaccine development (Matthews, 2006), reproduction (Nap et al., 2003, 2004; Bédécarrats et al., 2016), and nutrition (Klasing, 1984; Shang et al., 2018). The process to improve the annotation of the chicken is ongoing since it was first

sequenced in 2004. As sequencing and data science technologies rapidly evolve, new tools allow for a more accurate representation of the chicken genome. The Functional Annotation of Animal Genomes (FAANG) project was launched to comprehensively characterize the genome of farm animals to address the sustainable agriculture of farmed animals (Giuffra et al., 2019). The current study under the FAANG project focuses on the accurate annotation of the coding and long non-coding (LNC) RNA transcripts of various cells and tissues.

The chicken karyotype consists of 38 autosomes and 2 sex chromosomes (Z and W). The first drafted chicken genome was sequenced using whole-genome shotgun sequencing of a female Red Jungle Fowl, which is the closest wild variant of the domestic chicken and was 1.05 Gb in length (Hillier LaDeana, 2004; Schmid et al., 2000). The current version of the chicken genome (Gallus_gallus-6.0; GCCA_000002315.5) was sequenced using the combined long single-molecule sequencing technology, and improved BAC and physical maps (Warren et al., 2017). This resulted in the increase of genome size to 1.21 Gb, accounting for micro-chromosomes that were not accounted for or incorrectly assembled in the previous version (Cheng and Burt, 2018). The coding and non-coding regions, as well as the regulatory elements, of the chicken genome is the current focus in annotation studies. Annotation of chicken genes is performed computationally from reference genomes of species that are better annotated. This method is successful in identifying conserved genes across species. However, it is challenging for non-conserved genes because of the relative physiology of the chicken compared to other species, in addition to different genome size, and differences in intron/exon organization between species (Shepard et al., 2009). Our annotation of the chicken genome has 16,779 protein-coding genes (28,345 transcripts) and 7,577 lncRNA and other RNA biotypes (10,943 transcripts). Of the 39,288 unique transcripts, 72.1% are protein-coding, 22.6% are lncRNAs, 2.9% are miRNAs and 2.4% are other RNA biotypes.

While the central dogma has established that coding RNAs are translated into proteins, there continues to be a growing interest in the function of ncRNAs, some of which are not transcribed by RNA polymerase II (Mattick and Makunin, 2006). Recently, it was discovered that ncRNA plays a regulatory role in many biological processes (Zhang et al., 2009). Long non-coding RNAs (lncRNAs), which are non-protein-coding RNAs more than 200 nucleotides in length, play a role in post-transcriptional epigenetic regulation (Quinn and Chang, 2016). In chickens, lncRNA regulates a host of biological functions, including intramuscular adipogenesis (Zhang et al., 2017a,b), sperm motility (Liu et al., 2017), cholesterol synthesis (Muret et al., 2017), and embryonic development (Roeszler et al., 2012). In Avian leukovirus-J (ALV-J) infection, lncRNA regulates macrophages by targeting genes involved in apoptosis, inflammation, and cytokine-cytokine interactions (Dai et al., 2019). A subtype of lncRNA, named long intergenic non-coding RNA, has been implicated in Marek's disease (Han et al., 2017). Therefore, a comprehensive annotation of lncRNA expression in the chicken will reveal regulatory processes relevant to health and disease in an agriculturally important species.

In this study, we aimed to contribute to the catalog of transcriptomic differences of relevant chicken cells and tissues. We focused on multiple immune, intestinal, and reproduction-related tissues and cells. Specifically, tissue-specific immune cells (lung macrophage, spleen T cells, peripheral monocytes, and B-cells), immune organs (bursa and thymus), intestinal sections (jejunum, ileum, and proximal cecum), and ovary of the female reproductive tract were analyzed. The primary immune organs, the bursa, and thymus, are the origin of B cells and T cells in chickens, respectively (Cooper et al., 1966). The proximal cecum is located in the intestine at the ileocecal junction between the ileum and colon, is also the secondary immune organ in chickens due to the presence of mucosal-associated lymphoid tissues (MALT), such as the cecal tonsils. The findings described here will be useful toward a complete annotation of chicken tissue and cellular transcriptomes.

MATERIALS AND METHODS

Experimental Animals

The animal procedure was approved and conducted according to guidelines established by the Western University of Health Sciences, Pomona, California (WesternU) Institutional Animal Care and Use Committee, protocol R17/IACUC/058. The F1 crosses of Line 6 and Line 7 from the Avian Disease and Oncology Laboratory (ADOL) were used in this study (Stone, 1975; Briles et al., 1977; Bacon et al., 2000). The two lines have identical major histocompatibility complex (MHC) B*2 haplotype, but present different disease susceptibility to Marek's Disease Virus (Line 6_3: MDV-resistant and Line 7_2: MDV-susceptible) (Liu et al., 2001). The F1 crosses of these lines have been used in other annotation studies by the FAANG consortium; therefore, it is used in this study to allow for a better comparison of the data. The chickens were held in open cages in the vivarium of the University Research Center at Western University. In addition to daily health monitoring, fresh food and water were provided *ad libitum*. Room temperature was adjusted to and maintained at 32°C until 3 weeks of age. To minimize the risk of pecking disorders, chicks were kept under restricted lighting conditions throughout the study. Peripheral blood was collected from jugular or wing web veins. Experimental animals were euthanized by insufflation of isoflurane.

Sample Collection

All assays were performed in at least duplicates.

Immune tissue (thymus and bursa), intestinal tissues (jejunum, ileum, and proximal cecum), and reproductive tissue (ovary) were collected and flash-frozen in liquid nitrogen for later use. Tissue immune cells (lung macrophage, and spleen CD3+ T cells) were collected from the organs homogenized and filtered through 70 μm nylon cell strainers.

Tissue macrophages and T cells were extracted using magnetic beads (Dynabeads FlowComp Flexi, Invitrogen, Carlsbad, CA,

United States) coated with biotinylated-mouse-anti-chicken-monocyte/macrophage-monoclonal antibodies (Clone KUL-1, Cat. No. 8420-08, SouthernBiotech) and biotinylated-mouse-anti-chicken-CD3-monoclonal antibodies (Clone AV-20, Cat. No. 8200-08, SouthernBiotech, Birmingham, AL, United States), respectively.

Peripheral blood B cells were collected from the blood (Collisson et al., 2017). Briefly, the blood was diluted in an equal volume of PBS and layered slowly over Ficoll-Histopaque (1.083 g/mL) (Sigma-Aldrich, St. Louis, MO, United States), and then centrifuged for 35 min (400 × *g* at 23°C with the brake off). The interface containing the peripheral blood mononuclear cells (PBMCs) and B cells were collected. Peripheral blood B cells were extracted using magnetic beads (Dynabeads Pan Mouse IgG, Invitrogen, Waltham, MA, United States) coated with unlabeled-mouse-anti-chicken-Bu-1a/b-monoclonal antibodies (Clone AV-20, Cat. No. MCA5764, Bio-Rad, Hercules, CA, United States). Peripheral blood monocytes were collected from PBMCs. Briefly, after density gradient separation using Ficoll-Histopaque as described above, the monocytes were extracted from the PBMC using magnetic beads coated with unlabeled-mouse-anti-chicken-KUL01 monoclonal antibodies (SouthernBiotech).

The metadata and associated protocols concerning the 20 tissues have been deposited in the Biosamples database with the identifiers SAMA8868413 to SAMA8868433.

RNA Extraction and Library Construction

Total RNA from tissues and immune cells was collected using a modified Trizol/Chloroform method. Briefly, a second chloroform phase extraction and a second ethanol wash were included in the modified method. The total RNA from tissues was purified, and DNase treated using the Direct-zol RNA Miniprep Plus (Zymo Research, Irvine, CA, United States). The total RNA from immune cells (lung macrophage, spleen T cells, peripheral blood B cells, and peripheral blood monocytes) were not purified using the Direct-zol RNA miniprep Plus due to the lower concentration of the immune cell RNA compare to the tissue RNA. The total RNA from immune cells was DNase treated after extraction. Quality control of the total RNA was performed fluorometrically using the Qubit RNA HS Assay Kit and Qubit 3 (Thermo Fisher Scientific, Waltham, MA, United States) and RNA 6000 Nano Kit and Bioanalyzer 2100 (Agilent, Santa Clara, CA, United States). Total RNA with RNA integrity number (RIN) above 8.0 were used in the stranded library generation process using the Zymo-Seq RiboFree Total RNA Library Kit (Zymo Research). ERCC RNA Spike-In Controls (Invitrogen) were used to create a standard baseline measurement of RNA. Ribosomal-RNA (rRNA), globin, and overrepresented transcripts were removed, and sequencing adaptor ligation of the cDNA was removed by size selection and PCR enrichment. Libraries were barcoded with P5 and P7 index sequences according to the manufacturer's protocol.

RNA-Sequencing

Libraries were pooled and sequenced on HiseqX-PE150 by Novogene Bioinformatics Technology Co. (Beijing, China). Libraries were sequenced to an average depth of 43.7 million paired reads per library.

Bioinformatics Analyses of RNA-Sequencing Data

Raw reads were trimmed with TrimGalore (v0.4.1, parameters: –clip_R2 2) (Martin, 2011). Trimmed reads were mapped and quantified using STAR (v2.6.1c) and RSEM (v1.3.1) using the function rsem-calculate-expression (parameters: –star –sort-bam-by-coordinate) and the reference file Ensembl annotation release GRCg6a, Ensemble annotation release 98, genome-build-accession NCBI:GCA_000002315.5 (Li and Dewey, 2011). Read counts (raw, trimmed, aligned) can be found in **Supplementary Table 1**. Transcriptomes were assembled using StringTie (v2.1.4) and gffCompare (v0.11.6, parameters -R -r) (Pertea and Pertea, 2020). Counts of genes and transcripts from **Figures 2**, **3B,C** were obtained from the output of gffCompare.

Euclidean distance, pairwise correlations, and PCA plots were generated by pcaExplorer (Marini and Binder, 2019). PCA was performed using all expressed genes, used the gene counts from the RSEM quantification, and the gene counts were first normalized with DESeq2 (v1.30.0) (Love et al., 2014). Heatmaps were generated with Morpheus[1] (Gould, 2016). Shannon's entropy calculations were performed with the BioQC function entropyDiversity (Zhang et al., 2017a). Count matrices inputted to BioQC were normalized with DESeq2 and used counts from the RSEM output. Isoform entropy had an additional filter, requiring that the isoform's gene be expressed in at least two cell types. For all analyses, isoforms were considered expressed if they had an average TPM greater than 0.5 across replicates from the RSEM quantified counts were included. Sashimi plots were generated with ggsashimi (parameters: -M 10 -C 3 -O 3 –shrink –alpha 0.25 –base-size = 20 –ann-height = 4 –height = 3 –width = 18) (Garrido-Martín et al., 2018). Browser shots were generated using the UCSC genome browser (Kent et al., 2002). BigWig files for the UCSC genome browser were generated from the mapped bam files using deepTools bamCoverage (v3.5.0) (Ramírez et al., 2014). Transcription start site (TSS) annotations for head-to-head (H2H) detection was obtained from the UCSC table browser using the settings "clade: Vertebrate," "genome: Chicken," "assembly: Mar. 2018 GRCg6a/galGal6," "group: Genes and Gene Predictions," "track: Ensembl Genes," and "table: ensGene."

Extended lncRNA analysis was performed using the annotation from Jehl et al. (2020) [2](LNCextendedEns101.gtf.gz). Reads were pseuo-aligned to this reference first be converting the reference to a fasta file with gffread. Then a kallisto index was generated with kallisto index (parameter: –make-unique) and sample TPMs were obtained with kallisto quant. A TPM > 0.5 was used for an expression threshold. BioQC entropyDiversity was used to calculate the most specific lncRNAs by tissue type.

Differential gene expression was calculated using DESeq2 (v1.30.0) (Love et al., 2014). Genes with an adjusted *p*-value less than 0.05 were considered differentially expressed. GO

[1]https://software.broadinstitute.org/morpheus

[2]http://www.fragencode.org/lnchickenatlas.html

biological processes were calculated using WebGestalt (Liao et al., 2019) with an FDR threshold of 0.05 for determining GO category overrepresentation. WebGestalt was run with the basic parameters "Gallus gallus," "Over-Representation Analysis (ORA)," "Gene Ontology, and "Biological Process." "genome" was selected as the reference set. **Figures 5C**, **6C** and **Supplementary Figure 3B** directly use these GO terms. **Figures 5B**, **6B** display the weighted set cover, which reduces redundancy of the categories displayed. Full GO categories corresponding to the weighted set covers are provided in **Supplementary Tables 5, 6**. Venn diagrams were generated with Intervene (Khan and Mathelier, 2017). All tools used the default parameters unless otherwise indicated.

RESULTS

Sample Clustering and PCA

Ten cell and tissue types were profiled with RNA-sequencing with the goal of determining coding and primarily lncRNA expression, as well as isoform usage. All samples were compared to one another using Euclidean distance (**Figure 1A**) and principal component analysis (PCA) (**Figure 1B**) using the R package pcaExplorer (Marini, 2016). Replicates of the same tissue had the smallest Euclidean distance between one another (**Figure 1A**) and the highest Pearson correlation scores, except for the macrophages that seem to be somewhat distant in the second PCA dimension, and the highest Pearson correlation scores (**Supplementary Figure 1**). All expressed genes (**Figure 1B**) were used for PCA. Samples appear to form three distinct clusters based on functional category: immune system [B cells, bursa, macrophage (lung), monocytes (blood), T cells (spleen), thymus], reproductive tissue (ovary), and intestinal tissue (jejunum, ileum, proximal cecum). To identify genes highly specific to tissue or cell types, Shannon's entropy was calculated for each gene across all cell types, obtaining a specificity score for each gene. The expression of the 2000 most specific genes was visualized in a heatmap (**Figure 1C**), revealing that macrophage cells have the most specific gene expression, while ileum tissue and monocytes have the least. When the next 2000 most specific genes are visualized (**Supplementary Figure 2A**) we begin to see less tissue-specific expression and see genes that are expressed in a small subset of cell types, compared to the 1000 least specific genes (**Supplementary Figure 2B**), which show more uniform gene expression across all tissue and cell types. A UCSC browser shot of gene expression across all cell and tissue types shows the uniformity of expression among some genes and variable expression among others (**Figure 1D**).

Transcriptome Coverage and Biotype Detection

Among all samples, 73.4% (17,872) of all known chicken genes and 63.2% (24,812) of all known transcripts from Ensembl annotations were detected (genome build GRCg6a) (**Figure 2A**). Tissue and cell type-specific gene, transcript, and lncRNA counts are provided in **Table 1**. Between 9,839 (monocyte) – 14,418 (thymus) genes and 11,522 (monocyte) – 17,794 (proximal cecum) transcripts were detected in each sample (**Figure 2B**). Out of the fifteen transcript biotypes (protein_coding, lncRNA, miRNA, pseudogene, misc_RNA, snoRNA, snRNA, scaRNA, rRNA, processed_pseudogene, IG_V_gene, Mt_rRNA, Mt_tRNA, ribozyme, sRNA) in the *Gallus gallus* reference annotation, fourteen were found in each of the sample types. The largest number of transcripts detected was from protein-coding RNA and lncRNA (**Figure 2C**). Among all samples, 28,345 (90.0%) protein-coding transcripts were detected. More recently, an extended lncRNA annotation was released (Jehl et al., 2020). For this extended analysis, we used the genome annotation file for LNC-enriched Ensembl RNAs, which showed that 3,723 lncRNAs were identified among all cell and tissue types. Even though our library preparation method did not enrich for small RNAs, a low level of these transcripts was detected (**Figure 2D**). Additionally, protein-coding and lncRNA expression unique to each cell or tissue type was detected (**Figure 2E** and **Supplementary Table 2**). All cell and tissue types had a greater number of unique protein-coding genes, except for the ovary tissue, which had a higher number of unique lncRNAs. Lung macrophage expressed the most unique protein-coding genes (653), whereas jejunum tissue (28) and monocytes (19) expressed the fewest. For jejunum tissue, this may be attributable to the fact that other intestinal tissues, the proximal cecum and the ileum, were included in the analysis and may have more similar gene expression profiles than other tissues included in this study. The number of lncRNAs per tissue ranged from 464 [monocyte (blood)] to 2,179 [macrophage (lung)] (**Supplementary Figure 3A** and **Table 1**). Many of these lncRNAs were specific to a single tissue, with tissue-specific lncRNAs ranging from 4 [monocyte (blood)] to 408 [macrophage (lung)] (**Supplementary Figures 3B,C** and **Supplementary Table 2**). Since we did not sequence samples to a depth of 100 million aligned reads as recommended by FAANG for novel gene annotation, we did not attempt to discover new genes.

Isoform Characterization

Alternative splicing is a primary mechanism for diversifying protein expression. After constructing transcript isoforms from short-read sequencing, Shannon's entropy calculations revealed unique isoforms to each cell and tissue type were found among a set of 500 isoforms (**Figure 3A** and **Supplementary Table 3**). The highest number of unique isoforms was found in T cells. The lowest was in monocytes, B cells, and ileum tissue. When expanded to view expressions of the top 1,000 isoforms with the highest specificity, isoforms are less specific to a single cell or tissue type (**Supplementary Figure 2C**). In contrast, when the 1,000 isoforms with the least entropy are observed, we see uniform expression among most cell and tissue types (**Supplementary Figure 2D**). Each cell and sample type has an average of 1.14 (ovary) – 1.24 (spleen T cell) isoforms per gene (**Figure 3B**). Histograms allow us to further visualize the distribution of isoform counts per gene in each tissue (**Figure 3C** and **Supplementary Figure 4A**). Most genes express only a single isoform of around 10,000 for each cell and tissue type. Between 1396 (blood monocyte) – 2667 (proximal cecum) genes

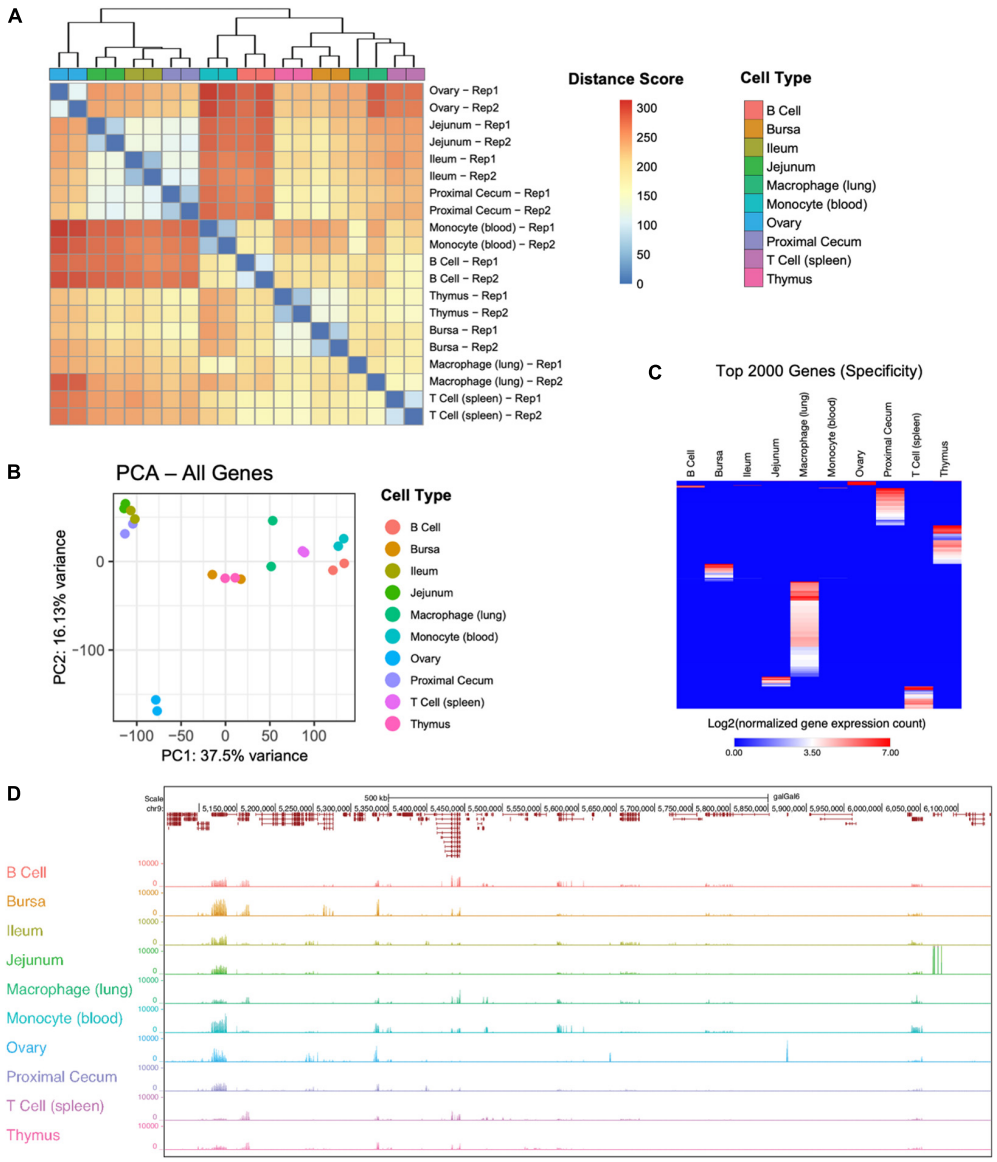


FIGURE 1 | Overview of tissue RNA-sequencing results. **(A)** Sample to sample distance heatmap quantifying the Euclidean distance between each sample. **(B)** Principal component analysis using all genes for all samples. **(C)** Expression of the 2000 genes with the highest Shannon's entropy values. Rows sorted using Euclidean distance. **(D)** UCSC browser shot of RNA-seq data showing variable expression among samples.

per cell type express two isoforms. A small subset of genes expressed more than four isoforms of a gene (**Figure 3C**, **Supplementary Figure 4A** insets, and **Supplementary Table 3**). There were 204 genes with four or more isoforms expressed among all cell and tissue types. The gene with the most isoforms is ST6GAL1, which has 10 isoforms in spleen T cell tissue. They fall into the GO biological process categories "localization within membrane," "activated T cell proliferation," and "cell migration" and the GO molecular function category "kinase binding" (**Supplementary Figure 4B**). To visualize differential splice junctions, a sashimi plot was generated for each sample (**Figure 3D** and **Supplementary Figure 5A**) for the gene *PDGFRB* (ENSGAL00000030613). Bursa, ileum, jejunum, ovary, proximal cecum, and thymus tissue express nearly all exons, whereas B cells, macrophages, monocytes, and T cells express a subset of exons. A UCSC browser shot of the gene *PDGFRB* (ENSGAL00000030613) also assists visualization of these differences in isoform expression of a single gene among different tissue and cell types (**Supplementary Figure 5B**).

Co-expression and Mono-Expression on Forward and Reverse Strands

A subset of expression will occur within the same genomic coordinate range on strands opposite to one another. Co-expression of this kind can serve as a feedback mechanism to

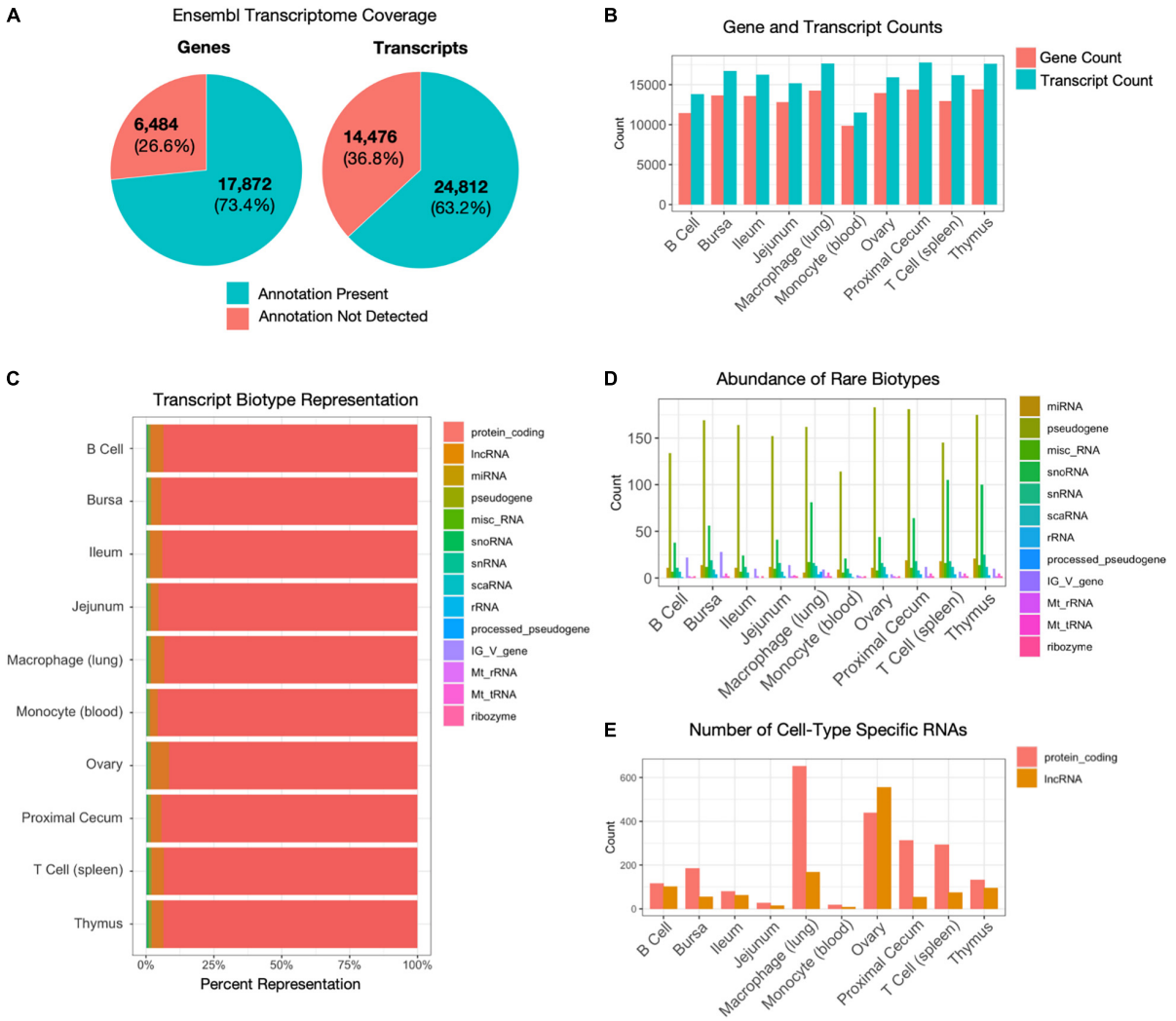


FIGURE 2 | Gene and transcript characterization. **(A)** Percentage of annotated chicken genes and transcripts detected across all samples. There are 24,356 genes and 39,288 transcripts in the *Gallus gallus* GRCg6a Ensembl annotation. **(B)** The number of genes and transcripts detected per cell type. Gene counts range between 9,839 [monocyte (blood)] – 14,418 [thymus]. Transcript counts range between 11,522 [monocyte (blood)] – 17,794 [proximal cecum]. **(C)** Breakdown of transcript types detected per cell type, by percentage. **(D)** Counts of low abundance transcript biotypes with less than 3% representation (all transcript biotypes, except lncRNA and protein-coding RNA). **(E)** The number of protein-coding RNAs and lncRNAs unique to each sample type.

regulate the expression of one another, particularly between lncRNAs and protein-coding transcripts. An example of this is the expression of the protein-coding gene *FRMPD4* (ENSGALT00000049598) occurring on the strand opposite to the lncRNA gene ENSGALT00000098634 (**Figure 4A**). Co-expression was determined by locating genes whose 5′UTR-3′UTR sequences were overlapping by at least one base pair on opposite strands of one another. The number of co-expressed pairs ranged from 371 (monocyte) to 621 (thymus) (**Figure 4B** and **Supplementary Table 4**). The majority of pairs were both protein-coding genes for all cell and tissue types (range: 307–454) (**Figure 4C**). The next most common pairing was protein coding-lncRNA co-expression (range: 10–51). A small number of instances were lncRNA-lncRNA co-expression (range: 0–6). Also present were interactions between other biotypes (miRNA, pseudogene, misc_RNA, snoRNA, snRNA, scaRNA, rRNA, processed_pseudogene, IG_V_gene, Mt_rRNA, Mt_tRNA, ribozyme, sRNA) (range: 11–19).

Also of interest is mono-expression: when two genes occur within the same genomic coordinates range on opposite strands of one another, but only one of the genes is expressed. The number of mono-expressed pairs ranged from 145 (lung macrophage) to 435 (ovary) (**Figure 4D**). Similar to co-expressed genes, the most common pairing were pairs of protein-coding genes (range: 102–278), followed by protein coding-lncRNA mono-expression (range: 9–63), then lncRNA-lncRNA mono-expression (range: 1–15) (**Figure 4E**). There were also instances of mono-expression between other biotypes (miRNA, pseudogene, misc_RNA, snoRNA, snRNA, scaRNA, rRNA, processed_pseudogene, IG_V_gene, Mt_rRNA, Mt_tRNA, ribozyme, sRNA) (range: 8–16) (**Supplementary Table 4**).

TABLE 1 | The number of transcripts, genes, and lncRNA by tissue.

Type	Tissue	# of transcripts	# of genes	# of lncRNA	Extended lncRNA analysis #
Immune	B cell	16,170	12,945	648	958
	Bursa	13,825	11,442	608	1,433
	Monocyte (blood)	17,794	14,380	316	464
	Macrophage (lung)	11,522	9,839	868	2,179
	T cell (spleen)	16,703	13,642	731	1,630
	Thymus	17,639	14,240	757	1,718
Intestinal	Jejunum	15,172	12.820	398	1,311
	Ileum	15,924	13,940	710	1,007
	Proximal cecum	17,619	14,418	690	1,746
Reproductive	Ovary	16,255	13,585	1,063	925

Similar to co-expressed genes are head-to-head (H2H) genes. These genes are located on opposite strands and their TSSs are within 1 kb of each other. We detected 2,628 H2H genes in the *Gallus gallus* genome annotation. Out of these, 1,590 were detected within our cell/tissue samples (**Supplementary Figure 6A**). All of the H2H genes are between protein-coding genes. At the cell/tissue level, we detected between 812 [monocyte (blood)] and 1,146 (bursa) total H2H genes expressed. A small subset of these is unique to a single cell/tissue type, with a range between 2 [monocyte (blood)] and 44 [macrophage (lung)] (**Supplementary Figures 6B,C** and **Supplementary Table 4**). We also examined mono-expressed H2H genes (**Supplementary Figures 6B,C**) and detected between 1020 (bursa) and 1114 (ovary) H2H expressed genes at the cell/tissue level. Similar to co-expressed H2H genes, mono-expressed H2H genes have a small subset that is unique to each cell/tissue type, ranging between 3 (ileum) and 54 [T cell (spleen)] (**Supplementary Table 4**).

DEG Analysis on Immune and Intestinal Samples

In addition to determining genes and isoforms highly enriched for cell- or tissue-specific expression, we identified genes differentially expressed between related cells or tissues. Differentially expressed genes (DEGs) were computed for six cell/tissue type comparisons using Deseq2. Three of these comparisons were among immune cell samples. There were 4911, 5907, and 3951 DEGs for the comparisons B cells vs. monocytes, B cells vs. bursa tissue, and bursa tissue vs. thymus tissue, respectively (**Figure 5A** and **Supplementary Table 5**). A weighted set cover analysis in WebGestalt (Liao et al., 2019) was performed to reduce redundancy and find the most representative GO biological process categories among sample comparisons (**Figure 5B**). The GO category "response to stress" was the only category shared among all three comparisons. When we compare the DEGs across all three comparisons, we find that there is a subset of genes that are shared across multiple sample comparisons, however, there is a sizable number of genes unique to each tissue comparison (**Figure 5C** and **Supplementary Table 5**). This was also reflected in similarities between enriched GO categories, which shared 33 categories between all three comparisons. Additionally, we see unique sets of genes among the top 10 DEGs for each comparison (**Figure 5D**).

Differentially expressed gene comparisons were also performed for three comparisons among intestinal samples. There were 3903, 2306, and 4270 DEGs for the comparisons of jejunum tissue vs. ileum tissue, jejunum tissue vs. proximal cecum tissue, and ileum tissue vs. proximal tissue, respectively (**Figure 6A** and **Supplementary Table 6**). A weighted set cover analysis was again performed (**Figure 6B**). There were no overlaps of enriched GO categories in the weighted set cover or among the sets of all GO terms enriched for each cell type, despite seeing 332 differentially expressed genes shared between all tissue comparisons (**Figures 6B,C** and **Supplementary Table 6**). Separating both DEG GO analyses, immune and intestinal, by upregulated and downregulated genes yields similar results (**Supplementary Tables 7, 8**). Among the sets of the top 10 differentially expressed genes for each tissue comparison, we observe the genes *APOA4* and *LCT* are present for tissue comparisons of jejunum vs. ileum and jejunum vs. proximal cecum (**Figure 6D**). Additionally, the tissue comparisons jejunum vs. ileum and ileum vs. proximal cecum share the five differentially expressed genes *MT-ND2, ND1, ND4, ND6,* and *SNORA73*.

Overrepresented KEGG pathways were also identified using WebGestalt for both of these immune and intestinal tissue comparisons. Each set of DEGs has a unique set of modified pathways, however, there are some overlaps between comparisons (**Supplementary Figures 7A–C**). In particular, in the immune system comparisons, the pathways "cell cycle" and "DNA replication" are enriched in the DEG sets for both the B cell vs. bursa and bursa vs. thymus comparisons (**Supplementary Table 9**). In the intestinal system comparisons, the "peroxisome" pathway is enriched in the DEG sets for jejunum vs. ileum and jejunum vs. proximal cecum. Additionally, the "spliceosome" pathway is enriched in the DEG sets for the jejunum vs. ileum and the ileum vs. proximal cecum. Out of all comparisons, the jejunum vs. proximal cecum has the most enriched pathways, many of which are involved in various metabolism and biosynthetic functions (**Supplementary Figures 7D–F** and **Supplementary Table 10**).

DISCUSSION

Side-by-side comparisons of transcriptomes were made for some of the immune cells and tissues, as well as intestinal tissues, to gain additional biological insight. B cells were compared to monocytes from peripheral blood, B cells with bursa, bursa with the thymus. The most significant (P-value) differentially expressed genes were highlighted in the results (**Figure 5D**). In the comparison between the monocytes and B cells, *CSF1R*, *GSTA3, LY86, S100A6, TGFβ1*, and *VCAN* were highly expressed in monocytes. Colony-stimulating factor-1 receptor (*CSF1R*) is a major stimulator of macrophage maturation from monocytes (Gan et al., 2020; Peng et al., 2020;

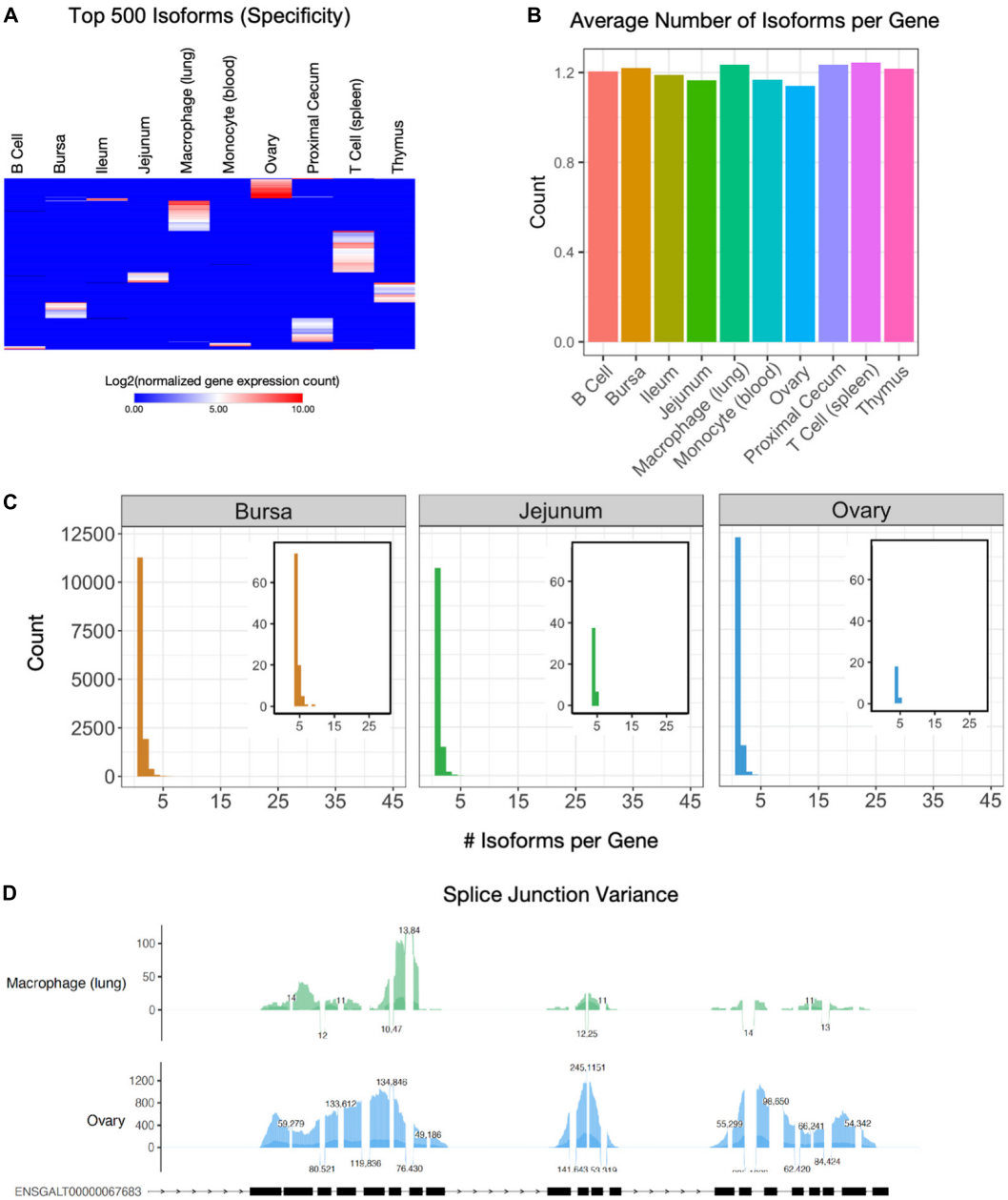


FIGURE 3 | Isoform characterization. **(A)** Expression of the 500 isoforms with the highest Shannon's entropy values. Rows sorted using Euclidean distance. Isoforms have been filtered for genes that have a TPM of at least 0.5 in at least two cell types. **(B)** The average number of isoforms per gene for each cell type. **(C)** Histogram of isoform counts per gene. The cutout plot in the upper-right corner is a zoomed-in section for 4+ isoforms per gene. **(D)** Sashimi plots of splice junction variance for macrophage (lung) cells and ovary tissue for gene *PDGFRB* (ENSGAL00000030613), which has a single annotated transcript ENSGALT00000067683.

Wu et al., 2020). Glutathione S-transferase α3 (*GSTA3*), for glutathione metabolism, is expressed in the macrophages against reactive oxygen species (ROSs) (McNeill et al., 2015). After phagocytosis of antigen or dead cells, macrophages release ROSs to destroy the ingested molecules through respiratory burst. Therefore, it is logical that monocytes have a higher expression of *GSTA3* to control the over-production of ROSs. Lymphocyte antigen 86 (*LY86*), also known as Myeloid Differentiating Protein-1 (*MD1*), activates toll-like receptors in innate immune cells (Candel et al., 2015). *S100A6* (calcyclin) has been implicated in cell differentiation and apoptosis (Donato et al., 2017). Transforming growth factor-β1 (*TGFβ1*) is produced by monocytes to regulate chemotaxis (McCartney-Francis et al., 1990; Sato et al., 2000). Versican (*VCAN*) is a chondroitin sulfate proteoglycan involved in cell proliferation (Zhang et al., 1998) and is produced by leukocytes to regulate inflammation (Wight et al., 2014). Due to the constant flux in monocyte development in the peripheral blood, it

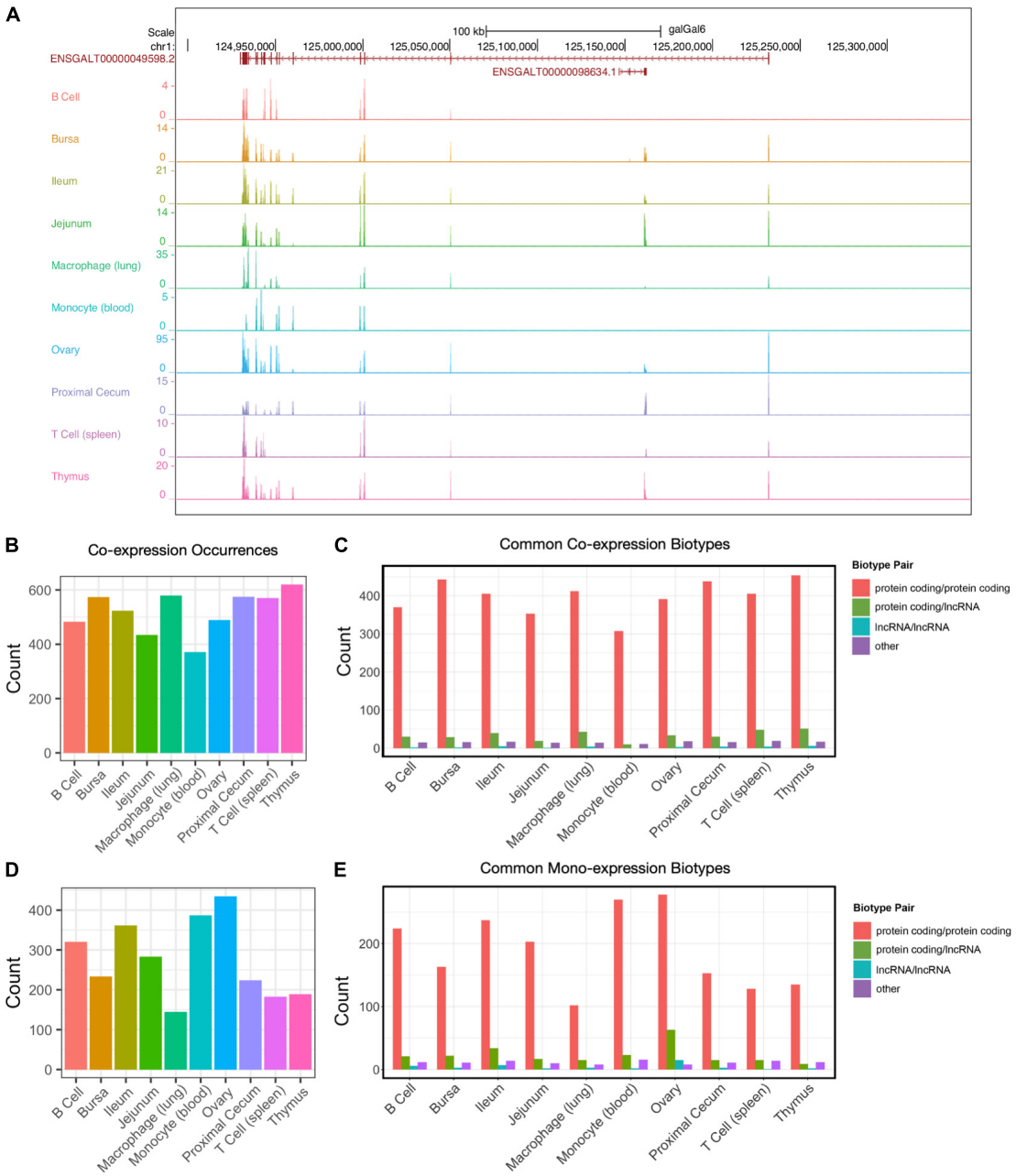


FIGURE 4 | Forward–reverse strand co-expression. **(A)** Example of co-expressed transcripts on the forward and reverse strands. ENSGALT00000049598/*FRMPD4* (reverse strand) and ENSGALG00000098634 (forward strand) are overlapping in their genomic coordinates. **(B)** The number of co-expression occurrences in each tissue type. **(C)** Co-expression counts of protein-coding RNAs and lncRNAs. **(D)** The number of mono-expression occurrences in each tissue type. **(E)** Mono-expression of protein-coding RNAs and lncRNAs.

explains the higher expressions of *Ly86, S100A6, TGFβ1*, and *VCAN* in monocytes.

DENND5B, HVCN1, and *IKZF3*, and *POU2AF1, BACH2, and IRF4* expression were significantly upregulated in the B cells compared to monocytes. The role of DENN Domain Containing 5B (*DENND5B*) on B cells is unclear. B cell antigen receptor (*BCR*) signaling requires the internalization of BCR with Hydrogen Voltage-Gated Channel 1 (*HVCN1*) to regulate ROS production (Capasso et al., 2010). The Ikaros Family of Zinc-finger Protein-3 (*IKZF3*) is involved in early B cell development and its expression is increased progressively throughout B cell development (Ferreirós-Vidal et al., 2013). The POU Class 2 Homeobox Associating Factor 1 (*POU2AF1*) promotes B cell development and maturation (Zhao et al., 2008). BACH2 is involved in proliferation of B cells (Miura et al., 2018) and IRF4 is essential for

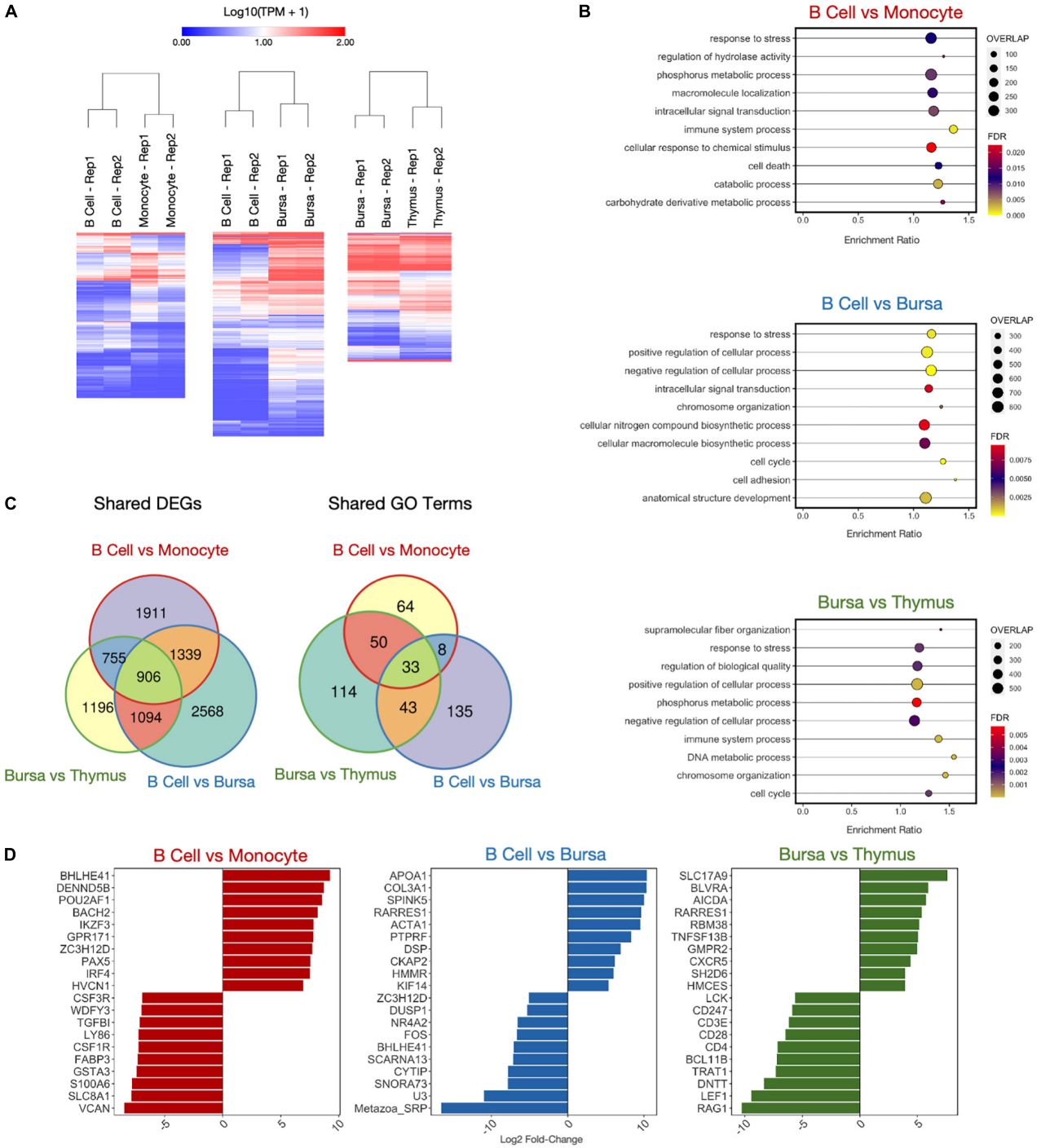


FIGURE 5 | Differentially expressed gene (DEG) analysis on immune samples. **(A)** Heatmaps of DEGs in three cell type comparisons: B cells vs. monocytes (4911 DEGs), B cells vs. bursa tissue (5907 DEGs), bursa tissue vs. thymus tissue (3951 DEGs). Samples were clustered both by column and by row using Euclidean distance based on log-transformed TPM value. **(B)** Enriched GO biological process categories for each sample using weighted set cover filtering in WebGestalt. "Overlap" quantifies the number of DEGs present in that GO set. **(C)** The numbers of DEGs overlapping between-sample comparisons. **(D)** Log_2 fold-change of top 10 upregulated and downregulated DEGs for each sample comparison.

lymphocyte function and involved in the development, affinity maturation, and terminal differentiation of B cells (Mittrücker et al., 1997).

APOA1, *PTPRF*, and *RARRES1* had higher expression in the B cells compared to bursa in this study. The bursa of Fabricius is a unique organ near the cloaca of the birds for B

cell development and production (Glick et al., 1956). *APOA1*, short for Apolipoprotein A-1, is a major component in high-density lipoprotein (HDL) for lipid transport in the plasma. Interestingly, *APOA1* was one of the most abundant proteins identified in the bursa in early embryonic development (Korte et al., 2013). However, the bursas sampled for this study were more mature, which might explain that the gene expression was lower. It is not clear what the role of *APOA1* in B cells might be. *PTPRF*, short for Protein Tyrosine Phosphatase Receptor Type F, regulates Wnt signaling, which mediates B cell differentiation (Qiang et al., 2003; Gan et al., 2020). RARRES1 (Retinoic Acid Receptor Responder 1), also known as Tazarotene-induced gene 1 protein/RAR-responsive protein TIG1, facilitates retinoic acid synthesis from β-carotene (precursor of vitamin A) (Chung and Lo, 2007; Mihály et al., 2011). Vitamin A and retinoic acid are essential for B cell development and antibody production (Ross et al., 2011), as well as monocyte differentiation into macrophages (Gundra et al., 2017).

BHLHE41, short for Basic Helix-Loop-Helix Family Member E41, is a regulator of B cell development, which is consistent with our data showing that BHLHE41 is more expressed in bursa than in more mature peripheral B cells (Kreslavsky et al., 2017). The Cytohesin 1 interacting protein (*CYTIP*) regulates lymphocyte cell adhesion (Boehm et al., 2003), an important function of B cells. cFos is involved in immune receptor interaction (Bush and Bishop, 2008). The transcription factor, *NR4A2*, limits B cell activation when the secondary T cell signaling is absent (Tan et al., 2020).

We identified several non-coding RNAs with higher expression in the bursa than B cells, particularly *Metazoa_SRP, SCARNA13*, and *U3*. *Metazoa_SRP* encodes for a signal recognition particle RNA that is predominantly studied in archaea, bacteria, fungi, and protozoa species (Rosenblad et al., 2004; Dumesic et al., 2015). Little is known about the *Metazoa_SRP* gene in animals but it is thought to be involved in the translocation of RNA between the endoplasmic reticulo-membrane and cytosol (Shan and Walter, 2005) and post-translational transport of proteins to the ER (Abell et al., 2004). *SCARNA-13*, (small Cajal body-specific RNA-13), is a regulatory RNA. These small RNAs regulate gene expressions in the Cajal bodies by controlling small nucleolar RNA such as the *U3* (Richard et al., 2003; Allantaz et al., 2012).

In the comparison of the bursa and thymus DEGs, higher expressed genes in the thymus are essential genes for T cell and thymic development, such as *CD247* (Lundholm et al., 2010), *CD28* (Lenschow et al., 1996), *CD3E* (Call et al., 2002), *CD4* (Zhang et al., 2009; Zhu et al., 2009), *DNTT* (Su et al., 2004, 2005), *LCK* (Van Laethem et al., 2013), *LEF1* (Xing et al., 2019), *RAG1* (Xing et al., 2019), *TRAT1* (Mijušković et al., 2015), while the *BCL11B* transcription factor is involved in both B and T cell (Avram and Califano, 2014). Genes higher expressed in bursa included CXCR5, TNFSF13B, AICDA, and SH2D6 (or BLNK). CXCR5 plays an important role in the migration of B and T cells to secondary lymphoid organs (Legler et al., 1998) and has previously been shown to be highly expressed in bursa (Annamalai and Selvaraj, 2011). TNFSF13B is a cytokine that belongs to the tumor necrosis factor (TNF) ligand family and is also known as B cell-activating factor (BAFF). It is expressed in B cell lineage cells and has been shown to play an important role in the proliferation and differentiation of B cells (Mackay et al., 1999). AICDA, the gene coding for AID (activation-induced cytidine deaminase), is essential for immunoglobulin (Ig) gene somatic hypermutation (SHM) and class switch DNA recombination (CSR). AID expression is induced by activated B-cell CD40 signaling, critical for germinal center reaction (Park et al., 2009). Finally, SH2D6 or BLNK, functions as a central linker protein, downstream of the B-cell receptor (BCR). Activation leads to a multitude of signaling pathways and regulating biological outcomes of B-cell function and development (Ishiai et al., 1999).

In summary, many of these genes were mostly involved in cellular processes relating to differentiation and cell metabolism as well as basic functions of immune cells such as cell adhesion and signal transduction. This was to be expected, as there was no explicit immunological stimulus involved, the transcriptome rather represents the baseline activity at the time sampled. Nevertheless, it was notable that DEGs in the comparison between bursa and thymus that were upregulated in the thymus were related to T cell differentiation and maturation. On the other hand, genes differentially upregulated in B cell vs. bursa or bursa vs. thymus, are mostly involved in B cell development and differentiation, or activation. Genes differentially regulated in B cells and monocytes are involved in specific functions of the cell types.

While the chicken ileum was previously profiled (Kuo et al., 2017), the jejunum and cecum were not studied previously. We included the top ten genes of the differential expression analyses between tissue types based on levels of significance. Hierarchical clustering showed clear discrimination between the different parts of the intestine (**Figure 6A**). Of the 3,903 DEGs of the jejunal and ileal cells, the number of genes involved in steroid metabolism is the most different between jejunal and ileal tissues. Lipid metabolism of fat in the diet requires steroid biosynthesis of molecules such as bile acid from the pancreas into the small intestine (Dawson and Karpen, 2015). The bile acid emulsifies lipid molecules, which travel through the small intestine and allow fatty acids to be absorbed. Consistent with the observation in rats, absorption of steroidal hormones decreases throughout the small intestine (Nakayama et al., 1999). Unsurprisingly, bile acid absorption can be twice as high in the jejunum than in the ileum (Krag and Phillips, 1974; Aldini et al., 1996). This further confirms the higher lipid metabolism of the jejunum than the ileum in chickens (Tancharoenrat et al., 2014). Of the a06 DEGs of the jejunum and proximal cecum, the number of genes involved in the oxidation-reduction process, lipid metabolic process, and cell adhesion were the most different. The primary role of the jejunum is the digestion and absorption of nutrients. In contrast, the ceca are blind-sacs in the chicken intestine that play multiple roles in nutrients absorption including bacterial fermentation of small molecules and biosynthesis of short-chain fatty acids (propionic and butyric acids) (Clench, 1999). The proximal cecum contains the cecal tonsils, which are the largest gut-associated lymphoid tissues (GALT) in chickens that demonstrate protective immune responses in the intestinal tract

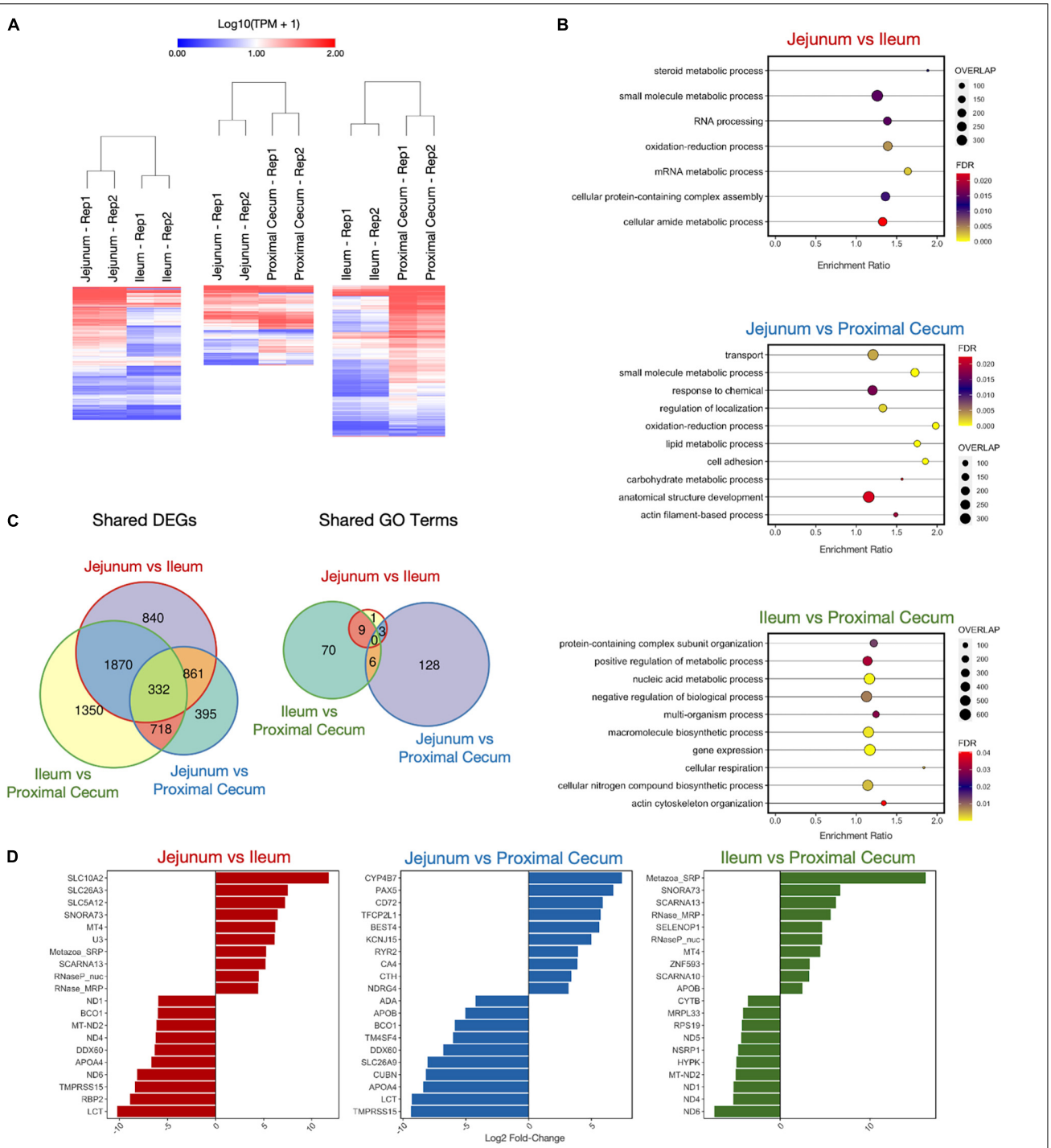


FIGURE 6 | Differentially Expressed Gene (DEG) Analysis on Intestinal Samples. **(A)** Heatmaps of DEGs in three cell type comparisons: Jejunum vs. Ileum (3903 DEGs), Jejunum vs. Proximal Cecum (2306 DEGs), Ileum vs. Proximal Cecum (4270 DEGs). Samples were clustered both by column and by row using Euclidean distance based on log transformed TPM value. **(B)** Enriched GO Biological Process categories for each sample using weighted set cover filtering in WebGestalt. "Overlap" quantities the number of DEGs present in that GO set. **(C)** The numbers of DEGs overlapping between sample comparisons. **(D)** Log_2 fold-change of top 10 upregulated and downregulated DEGs for each sample comparison.

(Heidari et al., 2015). Therefore, it is logical that the DEGs of these metabolic functions are more pronounced in the jejunum than in the cecum. Of the 4,270 DEGs between the ileum and cecum, the DEGs corresponding to cellular respiration were the most different. This could be expected as bacteria fermentation produces high levels of short-chain fatty acids in the cecum,

which can be used as energy by the intestinal cells (Murugesan et al., 2014). Due to the relative size and metabolic demands of the ileum compared to the cecum, much energy is needed from aerobic respiration and mitochondrial electron transport to produce adequate energy in the ileum.

Interestingly, the comparisons of the jejunum to the ileum and the proximal cecum revealed differential expression of *LCT*, the gene encoded for lactase production, which is lower in the jejunum compared to that in the ileum and proximal cecum (**Figure 6D**). Since chickens are not mammals, the expression of the lactase gene is perplexing. The expression of the lactase gene in chickens has been debated in the past (Hamilton and Mitchell, 1924). Several hypotheses had been proposed about the presence of the lactase gene in chickens. The presence of the lactase gene could be due to (1) bacterial fermentation of lactase in the intestine, (2) evolutionary artifacts, or (3) improper annotation of the gene in chickens that could have the same sequence but functionally different in the chicken compared to mammals. An early study using based on the disappearance of lactase *in vitro* showed that lactase was assimilated in the crop but not in the proventriculus or the intestine (Plimmer and Rosedale, 1922). However, the assumption of the disappearance of lactase as evident of lactose digestion is flawed because it does not account for the microbial degradation of lactose. Later, molecular cloning confirmed lactase expression in the chicken intestinal tract as well as in mussel (Freund et al., 1997). Based on our sequencing results, we cannot conclude whether this is due to evolutionary artifacts, evolutionary converged traits with separate lineages, or genes with the same sequence but with completely different functions, or otherwise.

Among the DEGs from the intestinal tract, *APOA4* is responsible for lipid metabolism (Tso et al., 2004; Wang et al., 2020) and lipid-soluble vitamin metabolism such as retinoic acid (vitamin A) (Hebiguchi et al., 2015). Coincidentally, *APOA4* and retinoic acid-binding protein-2 (*RBP2*), and beta-carotene oxygenase 1 (*BCO1*) had higher expressions in the ileum compared to the jejunum. Mitochondrial NADH dehydrogenase (*MT-ND2*), NADH dehydrogenase-1 (*ND1*), NADH dehydrogenase-4 (*ND4*), NADH dehydrogenase-5 (*ND5*), and NADH dehydrogenase-6 (*ND6*) relate to the electron transport chain that generates cellular energy in the form of ATP through oxidative respiration (Weiss et al., 1991). These energy metabolic genes had higher expression in the ileum compared to the jejunum and in the proximal cecum compared to the ileum, suggesting higher energy production through aerobic respiration in these tissues. The Transmembrane Serine Protease 15 (*TMPRSS15*) is an enteropeptidase secreted from the pancreas that catabolizes trypsinogen to trypsin and chymotrypsinogen to procarboxypeptidase for protein digestion in the intestine (Zhang et al., 2009). Expression of *TMPRSS15* was higher in the ileum than the jejunum, suggesting the increasing rate of protein digestion throughout the small intestine. Consistent with a previous study on ion transport in the intestine (Wingate et al., 1973), several ion transporter genes, Solute Carrier Family 5 Member 12 (SLC5A12) for sodium and glucose co-transport, Solute Carrier Family 26 Member for chloride transport, and Solute Carrier Family 10 Member A2 for sodium and bile acid co-transport had higher expression in the jejunum than the ileum. Metallothionein-4 (MT4) is a tissue-specific binding protein for zinc and copper for sequestering the trace minerals from pathogens and regulating the intra- and extra-cellular concentrations (Sakulsak, 2012). Expression of *MT4* was higher in the jejunum in the current study.

Several of the DEGs highly expressed in the proximal cecum are involved in lipid metabolism, including *APOA4* and *APOB* (major components of lipoproteins) (Schianca et al., 2011), Beta-carotene Oxygenase-1 (*BCO1*) (lipid-soluble vitamin A metabolism), *CUBN* (lipoprotein endocytosis) (Christensen and Birn, 2002), and *SLC26A9* (bile metabolism) (Li et al., 2016). Surprisingly, *APOB* is higher expressed in the ileum than the proximal cecum. Adenosine deaminase (*ADA*) is involved in purine metabolism for nucleotide synthesis (Ikehara and Fukui, 1974) and is abundant in lymphocytes (Sakumi and Sekiguchi, 1989). Bacteria in the intestine are essential for vitamin absorption for the host (Ikehara and Fukui, 1974; LeBlanc et al., 2013). Therefore, it is consistent that *TM4SF4*, which is involved in thiamine (vitamin B1) metabolism, displayed higher levels of expression in the proximal cecum. The DEAD/DEAD-Box Helicase-60 is involved in innate immunity (Perčulija and Ouyang, 2019). The mucosal-associated lymphoid tissue (MALT) in the proximal cecum is the secondary lymphoid organ of the chicken, and the cecum houses microbiota that regulates metabolism (Polansky et al., 2016). This could explain the higher expression of the DEAD/DEAD-Box Helicase-60 (*DDX60*), Adenosine deaminase (*ADA*), and Liver Enriched Antimicrobial Peptide 2 (LEAP2) in the proximal cecum.

The potassium inwardly rectifying channel subfamily J member 15 (*KCNJ15*) (Yuan et al., 2015), ryanodine receptor 2 (*RYR2*) (Jiang et al., 2004), and the bestrophin family anion channel (*BEST4*) (Fischmeister and Hartzell, 2005) for ion exchanges were upregulated in the jejunum compared to the proximal cecum. The carbonic anhydrase (*CA4*) utilizes zinc to produce carbonic acid for maintaining acid-base balance (Sly and Hu, 1995). Glutathionase (*CTH)* utilizes glutathione for antioxidant production against reactive oxygen species (ROS) (McBean, 2017). Since the jejunum is responsible for nutrient absorption, whereas the cecum is a blind sac that is involved in immunity, it is conceivable that these genes are higher expressed in the jejunum compared to the proximal cecum. In addition, two transcription factors were upregulated in the jejunum: transcription factor CP2 like 1 (*TFCP2L1*) and paired box family of transcription factor (*PAX5*). The former is involved in epithelial cells' development consistent with the high turnover of intestinal epithelial cells (Werth et al., 2017). However, the latter is involved in B cell development (Nutt et al., 1999). *CD72* regulates B cell development and signaling and it showed higher levels of expression in the jejunum compared to the proximal cecum (Kumanogoh et al., 2000). *CYP4B7* belongs to the cytochrome P450 family detoxification enzyme (Alber et al., 2020). The higher expression in the jejunum is consistent with its digestive functions. Three trace mineral-related genes had higher expression in the ileum than the proximal cecum:

selenoprotein (*SELENOP1*), metallothionein-4 (*MT4*), and zinc finger protein 593 (*ZNF593*). N-myc downregulated gene family (*NDRG4*) regulates smooth muscle cells (Qu et al., 2016). Similar to *NDRG4*, *ZNF593* regulates muscle cell differentiation (Lynch et al., 2019). Consistently, *NDRG4* and *ZNG593* are less expressed in the proximal cecum because the primary function of the cecum is thought to be related to modulate immunity and metabolism through the microbiota; whereas the jejunum and ileum are primarily responsible for digestion and absorption of nutrients that require contraction of smooth muscles during peristalsis.

In summary, we were able to correlate most of the differential expressed genes in the intestine to mostly metabolic processes related to nutrient digestion and absorption. Several genes in the distal part of the intestine were particularly implicated in vitamin metabolism. This was not surprising because vitamin metabolism requires the microbiota, which is more abundant in the distal intestines. Genes involved in energy metabolism are also abundant in the cecum, which suggests that microbial contribution of energy production in the intestine is especially important.

In the current study, whole transcriptome RNA-seq of immune, intestinal, and reproductive cells and tissues were sequenced. The Ensembl chicken annotation release 98 (GRCg6a, genome-build-accession NCBI:GCA_000002315.5), contains 16,779 protein-coding genes, 7,577 non-coding genes, and 39,288 gene transcripts. Of the non-coding genes, 5,504 were long non-coding genes; 10,301 lncRNAs are annotated when considering Jehl et al. (2020). From 10 diverse cell and tissue types, we recovered 73% of annotated genes and 63% of known transcripts. Of annotated genes, 90% of coding genes are expressed in the 10 cell and tissue types studied here, while only 36% of annotated lncRNAs are expressed. The potential regulatory role of lncRNAs may explain the limited expression, and suggest a more cell- or tissue-specific role. We found that biosamples often expressed hundreds of cell- or tissue-specific coding genes and lncRNAs. While many genes are commonly expressed in multiple samples, we also determined that over 500 isoforms of genes are uniquely expressed. Each cell and tissue type only expressed an average of 1–2 gene isoforms; however, each biosample type had at least one gene with 11 or more isoforms expressed in the cell or tissue type. We did not attempt to annotate novel genes base due to our current sequence depth per sample. Analysis of differentially expressed genes revealed biological processes that are consistent with a function in the cells or tissues of interest. Continued investigation of these genes should further our understanding of disease susceptibility/resistance, feed conversion, and egg production. Collectively, these data provide a deeper understanding of the chicken transcriptome in a cell- and tissue-specific manner. We have provided lists of unique transcripts, genes with high isoform count, sense-antisense co-expression pairs, and differentially expressed genes in our Supplementary Tables as a resource to the community. Additional samples from the FAANG and greater community will continue to advance efforts toward a comprehensive catalog of the chicken transcriptome.

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: Gene Expression Omnibus (GEO) – GSE166257.

ETHICS STATEMENT

The animal study was reviewed and approved by Western University of Health Sciences, Pomona, California (WesternU) Institutional Animal Care and Use Committee.

AUTHOR CONTRIBUTIONS

EO performed data analysis, figure generation, and manuscript writing. TN performed sample isolation and preparation, data interpretation, and manuscript writing. PC, LG, and PS contributed to sample isolation and preparation. ST led the tissue collection. YD and RH conceived the study, guided data analysis, contributed to data interpretation, and manuscript writing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the USDA NIFA Grant 2018-67015-27499.

ACKNOWLEDGMENTS

We would like to thank Hans Cheng at ADOL and Raj Kulkarni for supplying F1 6x7 chicks, and additional members of the Drechsler and Hawkins labs for their advice and insight.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.664424/full#supplementary-material

Supplementary Figure 1 | Pairwise correlations. Pearson correlation plot of all genes in a sample for each pairwise comparison generated with pcaExplorer. Plots use a subset of 1000 genes and use $\log_2$ normalized gene counts for plot axes and values, respectively.

Supplementary Figure 2 | Gene and isoform expression specificity. **(A)** Top 2000–4000 most specific genes. **(B)** The 1000 least specific genes. **(C)** The top 1000 most specific isoforms. **(D)** 1000 least specific isoforms. Rows sorted by Euclidean distance. **(C,D)** Isoforms have been filtered for genes that have a TPM of at least 0.5 in at least two cell types. Matrix entries that have no expression of that isoform's gene are colored black.

Supplementary Figure 3 | Extended lncRNA analysis. **(A)** Total lncRNA counts using the extended lncRNA annotation. **(B)** Cell/tissue-type specific lncRNA counts using the extended lncRNA annotation. **(C)** Top 1000 most specific lncRNAs for the extended lncRNA annotation.

Supplementary Figure 4 | Isoform histograms. **(A)** Histogram of isoform counts per gene for tissues not included in Figure 3B. The cutout plot in the upper-right corner is a zoomed-in section for 4+ isoforms per gene. **(B)** GO biological process and molecular function analysis for genes with 4+ isoforms.

Supplementary Figure 5 | Extended isoform visualization. **(A)** Additional sashimi plots from *PDGFRB* in Figure 3D. **(B)** Browser shot of the same annotated isoform for the gene *PDGFRB*.

Supplementary Figure 6 | H2H transcripts. **(A)** Percent of total co-expressed H2H transcripts (2,628) detected across all cell/tissue types. **(B)** Total H2H transcript counts by cell/tissue type. **(C)** Cell/tissue type-specific H2H transcript count.

Supplementary Figure 7 | Enriched pathways. **(A–C)** Enriched KEGG pathways for the immune comparisons in **Figure 5** (FDR < 0.05). **(D–F)** Enriched KEGG pathways for the intestinal comparisons in **Figure 6** (FDR < 0.05).

Supplementary Table 1 | Sequencing read counts.

Supplementary Table 2 | Cell type specific transcripts and lncRNAs (related to **Figure 2**). (Sheet 1) Protein coding RNAs. (Sheet 2) lncRNAs. (Sheet 3) Extended lncRNA analysis transcripts.

Supplementary Table 3 | Isoforms (related to **Figure 3**). (Sheet 1) Isoforms from specificity plot (**Figure 3A**). (Sheet 2) Genes with 4+ isoforms per tissue/cell type (cutouts in **Figure 3C** and **Supplementary Figure 4A**).

Supplementary Table 4 | Co-expressed and H2H pairs (related to Figure 4 and **Supplementary Figure 6**). (Sheet 1) Co-expressed transcript pairs. (Sheet 2) Mono-expressed transcript pairs. (Sheet 3) Co-expressed H2H transcript pairs. (Sheet 4) Mono-expressed H2H transcript pairs.

Supplementary Table 5 | Immune differential expression and GO results (related to **Figure 5**). (Sheet 1) B cell vs. monocyte DESeq2 results. (Sheet 2) B cell vs. bursa DESeq2 results. (Sheet 3) Bursa vs. thymus DESeq2 results. (Sheet 4) B cell vs. monocyte WebGestalt (GO) results. (Sheet 5) B Cell vs. bursa WebGestalt (GO) results. (Sheet 6) Bursa vs. thymus WebGestalt (GO) results.

Supplementary Table 6 | Intestinal differential expression and GO results (related to **Figure 6**). (Sheet 1) Jejunum vs. ileum DESeq2 results. (Sheet 2) Jejunum vs. proximal cecum DESeq2 results. (Sheet 3) Ileum vs. proximal cecum DESeq2 results. (Sheet 4) Jejunum vs. ileum WebGestalt (GO) results. (Sheet 5) Jejunum vs. proximal cecum WebGestalt (GO) results. (Sheet 6) Ileum vs. proximal cecum WebGestalt (GO) results

Supplementary Table 7 | Separate up/down regulated immune DEGs. (Sheet 1) B cell vs. monocyte WebGestalt (GO) results – UPREGULATED DEGs ONLY. (Sheet 2) B cell vs. monocyte WebGestalt (GO) results – DOWNREGULATED DEGs ONLY. (Sheet 3) B cell vs. bursa WebGestalt (GO) results – UPREGULATED DEGs ONLY. (Sheet 4) B cell vs. bursa WebGestalt (GO) results – DOWNREGULATED DEGs ONLY. (Sheet 5) Bursa vs. thymus WebGestalt (GO) results – UPREGULATED DEGs ONLY. (Sheet 6) Bursa vs. thymus WebGestalt (GO) results – DOWNREGULATED DEGs ONLY.

Supplementary Table 8 | Separate up/down regulated intestinal DEGs. (Sheet 1) Jejunum vs. ileum WebGestalt (GO) results – DOWNREGULATED DEGs ONLY. (Sheet 2) Jejunum vs. proximal cecum WebGestalt (GO) results – UPREGULATED DEGs ONLY. (Sheet 3) Jejunum vs. proximal cecum WebGestalt (GO) results – DOWNREGULATED DEGs ONLY. (Sheet 4) Ileum vs. proximal cecum WebGestalt (GO) results – UPREGULATED DEGs ONLY. (Sheet 5) Ileum vs. proximal cecum WebGestalt (GO) results – DOWNREGULATED DEGs ONLY.

Supplementary Table 9 | Immune pathway results. (Sheet 1) B cell vs. monocyte WebGestalt (pathway) results. (Sheet 2) B Cell vs. bursa WebGestalt (pathway) results. (Sheet 3) Bursa vs. thymus WebGestalt (pathway) results.

Supplementary Table 10 | Intestinal pathway results. (Sheet 1) Jejunum vs. ileum WebGestalt (pathway) results. (Sheet 2) Jejunum vs. proximal cecum WebGestalt (pathway) results. (Sheet 3) Ileum vs. proximal cecum WebGestalt (pathway) results.

REFERENCES

Abell, B. M., Pool, M. R., Schlenker, O., Sinning, I., and High, S. (2004). Signal recognition particle mediates post-translational targeting in eukaryotes. *EMBO J.* 23, 2755–2764. doi: 10.1038/sj.emboj.7600281

Alber, A., Morris, K. M., Bryson, K. J., Sutton, K. M., Monson, M. S., Chintoan-Uta, C., et al. (2020). Avian pathogenic *Escherichia coli* (APEC) strain-dependent immunomodulation of respiratory granulocytes and mononuclear phagocytes in CSF1R-reporter transgenic chickens. *Front. Immunol.* 10:3055. doi: 10.3389/fimmu.2019.03055

Aldini, R., Montagnani, M., Roda, A., Hrelia, S., Biagi, P., and Roda, E. (1996). Intestinal absorption of bile acids in the rabbit: different transport rates in jejunum and ileum. *Gastroenterology* 110, 459–468. doi: 10.1053/gast.1996.v110.pm8566593

Allantaz, F., Cheng, D. T., Bergauer, T., Ravindran, P., Rossier, M. F., Ebeling, M., et al. (2012). Expression profiling of human immune cell subsets identifies miRNA-mRNA regulatory relationships correlated with cell type specific expression. *PLoS One* 7:e29979. doi: 10.1371/journal.pone.0029979

Annamalai, T., and Selvaraj, R. K. (2011). Chemokine receptor CCR7 and CXCR5 mRNA in chickens following inflammation or vaccination. *Poultry Sci.* 90, 1695–1700. doi: 10.3382/ps.2011-01377

Avram, D., and Califano, D. (2014). The multifaceted roles of Bcl11b in thymic and peripheral T cells: impact on immune diseases. *J. Immunol.* 193, 2059–2065. doi: 10.4049/jimmunol.1400930

Bacon, L., Hunt, H., and Cheng, H. (2000). A review of the development of chicken lines to resolve genes determining resistance to diseases. *Poultry Sci.* 79, 1082–1093. doi: 10.1093/ps/79.8.1082

Bédécarrats, G. Y., Baxter, M., and Sparling, B. (2016). An updated model to describe the neuroendocrine control of reproduction in chickens. *General Comparat. Endocrinol.* 227, 58–63. doi: 10.1016/j.ygcen.2015.09.023

Boehm, T., Hofer, S., Winklehner, P., Kellersch, B., Geiger, C., Trockenbacher, A., et al. (2003). Attenuation of cell adhesion in lymphocytes is regulated by CYTIP, a protein which mediates signal complex sequestration. *EMBO J.* 22, 1014–1024. doi: 10.1093/emboj/cdg101

Briles, W. E., Stone, H. A., and Cole, R. (1977). Marek's disease: effects of B histocompatibility alloalleles in resistant and susceptible chicken lines. *Science* 195, 193–195. doi: 10.1126/science.831269

Bush, T. J. V., and Bishop, G. A. (2008). TLR7 and CD40 cooperate in IL-6 production via enhanced JNK and AP-1 activation. *Eur. J. Immunol.* 38, 400–409. doi: 10.1002/eji.200737602

Call, M. E., Pyrdol, J., Wiedmann, M., and Wucherpfennig, K. W. (2002). The organizing principle in the formation of the T cell receptor-CD3 complex. *Cell* 111, 967–979. doi: 10.1016/s0092-8674(02)01194-7

Candel, S., Sepulcre, M. P., Espín-Palazón, R., Tyrkalska, S. D., De Oliveira, S., Meseguer, J., et al. (2015). Md1 and Rp105 regulate innate immunity and viral resistance in zebrafish. *Dev. Comparat. Immunol.* 50, 155–165. doi: 10.1016/j.dci.2015.01.005

Capasso, M., Bhamrah, M. K., Henley, T., Boyd, R. S., Langlais, C., Cain, K., et al. (2010). HVCN1 modulates BCR signal strength via regulation of BCR-dependent generation of reactive oxygen species. *Nat. Immunol.* 11:265. doi: 10.1038/ni.1843

Cheng, Y., and Burt, D. W. (2018). Chicken genomics. *Int. J. Dev. Biol.* 62, 265–271.

Christensen, E. I., and Birn, H. (2002). Megalin and cubilin: multifunctional endocytic receptors. *Nat. Rev. Mol. Cell Biol.* 3, 258–267. doi: 10.1038/nrm778

Chung, G., and Lo, K. (2007). RARRES1 (Retinoic Acid Receptor Responder (Tazarotene Induced) 1). Atlas of Genetics and Cytogenetics in Oncology and Haematology. Vanddœuvre-lès-Nancy, France: INIST-CNRS Institute for scientific and technical information.

Clench, M. H. (1999). The avian cecum: update and motility review. *J. Exp. Zool.* 283, 441–447. doi: 10.1002/(sici)1097-010x(19990301/01)283:4/5<441::aid-jez13>3.0.co;2-8

Collisson, E., Griggs, L., and Drechsler, Y. (2017). Macrophages from disease resistant B2 haplotype chickens activate T lymphocytes more effectively than macrophages from disease susceptible B19 birds. *Dev. Comparat. Immunol.* 67, 249–256. doi: 10.1016/j.dci.2016.09.013

Cooper, M. D., Peterson, R. D., South, M. A., and Good, R. A. (1966). The functions of the thymus system and the bursa system in the chicken. *J. Exp. Med.* 123, 75–102. doi: 10.1084/jem.123.1.75

Dai, M., Feng, M., Xie, T., and Zhang, X. (2019). Long non-coding RNA and MicroRNA profiling provides comprehensive insight into non-coding RNA involved host immune responses in ALV-J-infected chicken primary macrophage. *Dev. Comparat. Immunol.* 100:103414. doi: 10.1016/j.dci.2019.103414

Dawson, P. A., and Karpen, S. J. (2015). Intestinal transport and metabolism of bile acids. *J. Lip. Res.* 56, 1085–1099. doi: 10.1194/jlr.r054114

Donato, R., Sorci, G., and Giambanco, I. (2017). S100A6 protein: functional roles. *Cell. Mol. Life Sci.* 74, 2749–2760. doi: 10.1007/s00018-017-2526-9

Dumesic, P. A., Rosenblad, M. A., Samuelsson, T., Nguyen, T., Moresco, J. J., Yates Iii, J. R., et al. (2015). Noncanoncial signal recognition particle RNAs in a major eukaryotic phylum revealed by purification of SRP from the human pathogen *Cryptococcus neoformans*. *Nucl. Acids Res.* 43, 9017–9027. doi: 10.1093/nar/gkv819

Ferreirós-Vidal, I., Carroll, T., Taylor, B., Terry, A., Liang, Z., Bruno, L., et al. (2013). Genome-wide identification of Ikaros targets elucidates its contribution to mouse B-cell lineage specification and pre-B-cell differentiation. *Blood J. Am. Soc. Hematol.* 121, 1769–1782. doi: 10.1182/blood-2012-08-450114

Fischmeister, R., and Hartzell, H. C. (2005). Volume sensitivity of the bestrophin family of chloride channels. *J. Physiol.* 562, 477–491. doi: 10.1113/jphysiol.2004.075622

Freund, J.-N., Jost, B., Loreztz, O., and Duluc, I. (1997). Identification of homologues of the mammalian intestinal lactase gene in non-mammals (birds and molluscs). *Biochem. J.* 322, 491–498. doi: 10.1042/bj3220491

Gan, T., Stevens, A. T., Xiong, X., Wen, Y.-A., Farmer, T. N., Li, A. T., et al. (2020). Inhibition of protein tyrosine phosphatase receptor type F suppresses Wnt signaling in colorectal cancer. *Oncogene* 39, 6789–6801. doi: 10.1038/s41388-020-01472-z

Garrido-Martín, D., Palumbo, E., Guigó, R., and Breschi, A. (2018). ggsashimi: Sashimi plot revised for browser-and annotation-independent splicing visualization. *PLoS Comput. Biol.* 14:e1006360. doi: 10.1371/journal.pcbi.1006360

Giuffra, E., Tuggle, C. K., and Consortium, F. (2019). Functional annotation of animal genomes (FAANG): current achievements and roadmap. *Ann. Rev. Animal Biosci.* 7, 65–88. doi: 10.1146/annurev-animal-020518-114913

Glick, B., Chang, T. S., and Jaap, R. G. (1956). The bursa of Fabricius and antibody production. *Poultry Sci.* 35, 224–225. doi: 10.3382/ps.0350224

Gould, J. (2016). *Morpheus: Versatile Matrix Visualization and Analysis Software*. Cambridge, MA, USA: Broad Institute.

Gundra, U. M., Girgis, N. M., Gonzalez, M. A., San Tang, M., Van Der Zande, H. J., Lin, J.-D., et al. (2017). Vitamin a mediates conversion of monocyte-derived macrophages into tissue-resident macrophages during alternative activation. *Nat. Immunol.* 18, 642–653. doi: 10.1038/ni.3734

Hamilton, T., and Mitchell, H. (1924). The occurrence of lactase in the alimentary tract of the chicken. *J. Agricul. Res.* 27, 605–608.

Han, B., He, Y., Zhang, L., Ding, Y., Lian, L., Zhao, C., et al. (2017). Long intergenic non-coding RNA GALMD3 in chicken Marek's disease. *Sci. Rep.* 7, 1–13.

Hebiguchi, T., Mezaki, Y., Morii, M., Watanabe, R., Yoshikawa, K., Miura, M., et al. (2015). Massivnts of genes ane bowel resection upregulates the intestinal mRNA expression levels of cellular retinol-binding protein II and apolipoprotein A-IV and alters the intestinal vitamin A status in rats. *Int. J. Mol. Med.* 35, 724–730. doi: 10.3892/ijmm.2015.2066

Heidari, M., Fitzgerald, S. D., and Zhang, H. (2015). Immune responses in cecal tonsils of Marek's disease Virus–infected chickens. *Avian Dis.* 59, 213–226. doi: 10.1637/10950-093014-reg.1

Hillier LaDeana, W. (2004). International chicken genome sequencing consortium. *Nature* 432, 695–716.

Ikehara, M., and Fukui, T. (1974). Studies of nucleosides and nucleotides: LVIII. Deamination of adenosine analogs with calf intestine adenosine deaminase. *Biochim. Biophys. Acta (BBA) General Subjects* 338, 512–519.

Ishiai, M., Kurosaki, M., Pappu, R., Okawa, K., Ronko, I., Fu, C., et al. (1999). BLNK required for coupling Syk to PLCγ2 and Rac1-JNK in B cells. *Immunity* 10, 117–125. doi: 10.1016/s1074-7613(00)80012-6

Jehl, F., Muret, K., Bernard, M., Boutin, M., Lagoutte, L., Désert, C., et al. (2020). An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. *Sci. Rep.* 10, 1–17.

Jiang, D., Xiao, B., Yang, D., Wang, R., Choi, P., Zhang, L., et al. (2004). RyR2 mutations linked to ventricular tachycardia and sudden death reduce the threshold for store-overload-induced Ca2+ release (SOICR). *Proc. Natl. Acad. Sci.U.S.A.* 101, 13062–13067. doi: 10.1073/pnas.0402388101

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006. doi: 10.1101/gr.229102.

Khan, A., and Mathelier, A. (2017). Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinform.* 18:287. doi: 10.1186/s12859-017-1708-7

Klasing, K. C. (1984). Effect of inflammatory agents and interleukin 1 on iron and zinc metabolism. *Am. J. Phys. Regul. Integrat. Comparat. Physiol.* 247, R901–R904.

Korte, J., Fröhlich, T., Kohn, M., Kaspers, B., Arnold, G. J., and Härtle, S. (2013). 2D DIGE analysis of the bursa of Fabricius reveals characteristic proteome profiles for different stages of chicken B−cell development. *Proteomics* 13, 119–133. doi: 10.1002/pmic.201200177

Krag, E., and Phillips, S. F. (1974). Active and passive bile acid absorption in man. Perfusion studies of the ileum and jejunum. *J. Clin. Investigat.* 53, 1686–1694. doi: 10.1172/jci107720

Kreslavsky, T., Vilagos, B., Tagoh, H., Poliakova, D. K., Schwickert, T. A., Wöhner, M., et al. (2017). Essential role for the transcription factor Bhlhe41 in regulating the development, self-renewal and BCR repertoire of B-1a cells. *Nat. Immunol.* 18, 442–455. doi: 10.1038/ni.3694

Kumanogoh, A., Watanabe, C., Lee, I., Wang, X., Shi, W., Araki, H., et al. (2000). Identification of CD72 as a lymphocyte receptor for the class IV semaphorin CD100: a novel mechanism for regulating B cell signaling. *Immunity* 13, 621–631. doi: 10.1016/s1074-7613(00)00062-5

Kuo, R. I., Tseng, E., Eory, L., Paton, I. R., Archibald, A. L., and Burt, D. W. (2017). Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* 18:1–19. doi: 10.1101/780015

LeBlanc, J. G., Milani, C., De Giori, G. S., Sesma, F., Van Sinderen, D., and Ventura, M. (2013). Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr. Opin. Biotechnol.* 24, 160–168. doi: 10.1016/j.copbio.2012.08.005

Legler, D. F., Loetscher, M., Roos, R. S., Clark-Lewis, I., Baggiolini, M., and Moser, B. (1998). B cell–attracting chemokine 1, a human CXC chemokine expressed in lymphoid tissues, selectively attracts B lymphocytes via BLR1/CXCR5. *J. Exp. Med.* 187, 655–660. doi: 10.1084/jem.187.4.655

Lenschow, D. J., Walunas, T. L., and Bluestone, J. A. (1996). CD28/B7 system of T cell costimulation. *Ann. Rev. Immunol.* 14, 233–258. doi: 10.1146/annurev.immunol.14.1.233

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12:323. doi: 10.1186/1471-2105-12-323

Li, T., Riederer, B., Liu, X., Pallagi, P., Singh, A. K., Soleimani, M., et al. (2016). Tu1491 loss of Slc26a9 anion transporter results in reduced pancreatic fluid secretion in young female mice. *Gastroenterology* 150:S915.

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucl. Acids Res.* 47, W199–W205.

Liu, H. C., Cheng, H., Tirunagaru, V., Sofer, L., and Burnside, J. (2001). A strategy to identify positional candidate genes conferring Marek's disease resistance by integrating DNA microarrays and genetic mapping. *Animal Genet.* 32, 351–359.

Liu, Y., Sun, Y., Li, Y., Bai, H., Xue, F., Xu, S., et al. (2017). Analyses of long non-coding RNA and mRNA profiling using RNA sequencing in chicken testis with extreme sperm motility. *Sci. Rep.* 7, 1–8. doi: 10.1002/9783527678679.dg06939

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.

Lundholm, M., Mayans, S., Motta, V., Löfgren-Burström, A., Danska, J., and Holmberg, D. (2010). Variation in the CD3ζ (Cd247) gene correlates with altered T cell activation and is associated with autoimmune diabetes. *J. Immunol.* 184, 5537–5544. doi: 10.4049/jimmunol.0904012

Lynch, S. A., Mcleod, M. A., Orsech, H. C., Cirelli, A. M., and Waddell, D. S. (2019). Zinc finger protein 593 is upregulated during skeletal muscle atrophy and modulates muscle cell differentiation. *Exp. Cell Res.* 383:111563. doi: 10.1016/j.yexcr.2019.111563

Mackay, F., Woodcock, S. A., Lawton, P., Ambrose, C., Baetscher, M., Schneider, P., et al. (1999). Mice transgenic for BAFF develop lymphocytic disorders along with autoimmune manifestations. *J. Exp. Med.* 190, 1697–1710. doi: 10.1084/jem.190.11.1697

Marini, F. (2016). *pcaExplorer: Interactive Visualization of RNA-Seq Data Using a Principal Components Approach*. Available online at: https://bioconductor.org/packages/release/bioc/html/pcaExplorer.html

Marini, F., and Binder, H. (2019). pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinform.* 20:331. doi: 10.1186/s12859-019-2879-1

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *J. EMBnet* 17, 10–12. doi: 10.14806/ej.17.1.200

Matthews, J. T. (2006). Egg-based production of influenza vaccine: 30 years of commercial experience. *Br. Wash. Natl. Acad. Eng.* 36:17.

Mattick, J. S., and Makunin, I. V. (2006). Non-coding RNA. *Hum. Mol. Genet.* 15, R17–R29.

McBean, G. J. (2017). Cysteine, glutathione, and thiol redox balance in astrocytes. *Antioxidants* 6:62. doi: 10.3390/antiox6030062

McCartney-Francis, N., Mizel, D., Wong, H., Wahl, L., and Wahl, S. (1990). TGF-β regulates production of growth factors and TGF-β by human peripheral blood monocytes. *Growth Factors* 4, 27–35. doi: 10.3109/08977199009011007

McNeill, E., Crabtree, M. J., Sahgal, N., Patel, J., Chuaiphichai, S., Iqbal, A. J., et al. (2015). Regulation of iNOS function and cellular redox state by macrophage Gch1 reveals specific requirements for tetrahydrobiopterin in NRF2 activation. *Free Radical Biol. Med.* 79, 206–216. doi: 10.1016/j.freeradbiomed.2014.10.575

Mihály, J., Gamlieli, A., Worm, M., and Rühl, R. (2011). Decreased retinoid concentration and retinoid signalling pathways in human atopic dermatitis. *Exp. Dermatol.* 20, 326–330. doi: 10.1111/j.1600-0625.2010.01225.x

Mijušković, M., Chou, Y.-F., Gigi, V., Lindsay, C. R., Shestova, O., Lewis, S. M., et al. (2015). Off-target V (D) J recombination drives lymphomagenesis and is escalated by loss of the Rag2 C terminus. *Cell Rep.* 12, 1842–1852. doi: 10.1016/j.celrep.2015.08.034

Mittrücker, H.-W., Matsuyama, T., Grossman, A., Kündig, T. M., Potter, J., Shahinian, A., et al. (1997). Requirement for the transcription factor LSIRF/IRF4 for mature B and T lymphocyte function. *Science* 275, 540–543. doi: 10.1126/science.275.5299.540

Miura, Y., Morooka, M., Sax, N., Roychoudhuri, R., Itoh-Nakadai, A., Brydun, A., et al. (2018). Bach2 promotes B cell receptor–induced proliferation of B lymphocytes and represses cyclin-dependent kinase inhibitors. *J. Immunol.* 200, 2882–2893. doi: 10.4049/jimmunol.1601863

Muret, K., Klopp, C., Wucher, V., Esquerré, D., Legeai, F., Lecerf, F., et al. (2017). Long noncoding RNA repertoire in chicken liver and adipose tissue. *Genet. Selection Evol.* 49, 1–17. doi: 10.1155/2019/3945475

Murugesan, G. R., Romero, L. F., and Persia, M. E. (2014). Effects of protease, phytase and a *Bacillus* sp. direct-fed microbial on nutrient and energy digestibility, ileal brush border digestive enzyme activity and cecal short-chain fatty acid concentration in broiler chickens. *PLoS One* 9:e101888. doi: 10.1371/journal.pone.0101888

Nakayama, A., Eguchi, O., Hatakeyama, M., Saitoh, H., and Takada, M. (1999). Different absorption behaviors among steriod hormones due to possible interaction with P-glycoprotein in the rat small intestine. *Biol. Pharmaceutical Bull.* 22, 535–538. doi: 10.1248/bpb.22.535

Nap, A. W., Dunselman, G. A., De Goeij, A. F., Evers, J. L., and Groothuis, P. G. (2004). Inhibiting MMP activity prevents the development of endometriosis in the chicken chorioallantoic membrane model. *Hum. Reprod.* 19, 2180–2187. doi: 10.1093/humrep/deh408

Nap, A. W., Groothuis, P. G., Demir, A. Y., Maas, J. W., Dunselman, G. A., De Goeij, A. F., et al. (2003). Tissue integrity is essential for ectopic implantation of human endometrium in the chicken chorioallantoic membrane. *Hum. Reprod.* 18, 30–34. doi: 10.1093/humrep/deg033

NCC. (2019). *Broiler Chicken Industry Key Facts 2019*. Washington, DC: National Chicken Council.

Nutt, S. L., Heavey, B., Rolink, A. G., and Busslinger, M. (1999). Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature* 401, 556–562. doi: 10.1038/44076

Park, S.-R., Zan, H., Pal, Z., Zhang, J., Al-Qahtani, A., Pone, E. J., et al. (2009). HoxC4 binds to the promoter of the cytidine deaminase AID gene to induce AID expression, class-switch DNA recombination and somatic hypermutation. *Nat. Immunol.* 10:540. doi: 10.1038/ni.1725

Peng, L., Van Den Biggelaar, R. H., Jansen, C. A., Haagsman, H. P., and Veldhuizen, E. J. (2020). A method to differentiate chicken monocytes into macrophages with proinflammatory properties. *Immunobiology* 225:152004. doi: 10.1016/j.imbio.2020.152004

Perčulija, V., and Ouyang, S. (2019). "Diverse roles of dead/deah-box helicases in innate immunity and diseases," in *Helicases From All Domains of Life*, ed. R. Tuteja (Amsterdam: Elsevier), 141–171. doi: 10.1016/b978-0-12-814685-9.00009-9

Pertea, G., and Pertea, M. (2020). GFF utilities: GffRead and GffCompare. *Research* 9:304. doi: 10.12688/f1000research.23297.2

Plimmer, R. H. A., and Rosedale, J. L. (1922). Distribution of enzymes in the alimentary canal of the chicken. *Biochem. J.* 16, 23–26. doi: 10.1042/bj0160023

Polansky, O., Sekelova, Z., Faldynova, M., Sebkova, A., Sisak, F., and Rychlik, I. (2016). Important metabolic pathways and biological processes expressed by chicken cecal microbiota. *Appl. Environ. Microbiol.* 82, 1569–1576. doi: 10.1128/aem.03473-15

Qiang, Y.-W., Endo, Y., Rubin, J. S., and Rudikoff, S. (2003). Wnt signaling in B-cell neoplasia. *Oncogene* 22, 1536–1545. doi: 10.1038/sj.onc.1206239

Qu, X., Li, J., and Baldwin, H. S. (2016). Postnatal lethality and abnormal development of foregut and spleen in Ndrg4 mutant mice. *Biochem. Biophys. Res. Commun.* 470, 613–619. doi: 10.1016/j.bbrc.2016.01.096

Quinn, J. J., and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 17:47. doi: 10.1038/nrg.2015.10

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., and Manke, T. (2014). deeptools: a flexible platform for exploring deep-sequencing data. *Nucl. Acids Res.* 42, W187–W191.

Richard, P., Darzacq, X., Bertrand, E., Jády, B. E., Verheggen, C., and Kiss, T. (2003). A common sequence motif determines the Cajal body-specific localization of box H/ACA scaRNAs. *EMBO J.* 22, 4283–4293. doi: 10.1093/emboj/cdg394

Roeszler, K. N., Itman, C., Sinclair, A. H., and Smith, C. A. (2012). The long non-coding RNA, MHM, plays a role in chicken embryonic development, including gonadogenesis. *Dev. Biol.* 366, 317–326. doi: 10.1016/j.ydbio.2012.03.025

Rosenblad, M. A., Zwieb, C., and Samuelsson, T. (2004). Identification and comparative analysis of components from the signal recognition particle in protozoa and fungi. *BMC Genom.* 5:5. doi: 10.1186/1471-2164-5-5

Ross, A. C., Chen, Q., and Ma, Y. (2011). Vitamin A and retinoic acid in the regulation of B-cell development and antibody production. *Vitamins Hormones* 86, 103–126. doi: 10.1016/b978-0-12-386960-9.00005-8

Sakulsak, N. (2012). Metallothionein: an overview on its metal homeostatic regulation in mammals. *Int. J. Morphol* 30, 1007–1012. doi: 10.4067/s0717-95022012000300039

Sakumi, K., and Sekiguchi, M. (1989). Regulation of expression of the ada gene controlling the adaptive response: interactions with the ada promoter of the Ada protein and RNA polymerase. *J. Mol. Biol.* 205, 373–385. doi: 10.1016/0022-2836(89)90348-3

Sato, K., Kawasaki, H., Nagayama, H., Enomoto, M., Morimoto, C., Tadokoro, K., et al. (2000). TGF-β1 reciprocally controls chemotaxis of human peripheral blood monocyte-derived dendritic cells via chemokine receptors. *J. Immunol.* 164, 2285–2295. doi: 10.4049/jimmunol.164.5.2285

Schianca, G. C., Pedrazzoli, R., Onolfo, S., Colli, E., Cornetti, E., Bergamasco, L., et al. (2011). ApoB/apoA-I ratio is better than LDL-C in detecting cardiovascular risk. *Nut. Metabol. Cardiovascular Dis.* 21, 406–411. doi: 10.1016/j.numecd.2009.11.002

Schmid, M., Nanda, I., Guttenbach, M., Steinlein, C., Hoehn, M., Schartl, M., et al. (2000). First report on chicken genes and chromosomes 2000. *Cytogenetic Genome Res.* 90, 169–218.

Shan, S.-O., and Walter, P. (2005). Co-translational protein targeting by the signal recognition particle. *FEBS Lett.* 579, 921–926. doi: 10.1016/j.febslet.2004.11.049

Shang, Y., Kumar, S., Oakley, B., and Kim, W. K. (2018). Chicken gut microbiota: importance and detection technology. *Front. Veterinary Sci.* 5:254. doi: 10.3389/fvets.2018.00254

Shepard, S., Mccreary, M., and Fedorov, A. (2009). The Peculiarities of large intron splicing in animals. *PLoS One* 4:e7853. doi: 10.1371/journal.pone.0007853

Sly, W. S., and Hu, P. Y. (1995). Human carbonic anhydrases and carbonic anhydrase deficiencies. *Ann. Rev. Biochem.* 64, 375–401. doi: 10.1146/annurev.bi.64.070195.002111

Stone, H. A. (1975). Use of highly inbred chickens in research. *Agricul. Res. Service* 21, 1–22.

Su, R.-C., Brown, K. E., Saaber, S., Fisher, A. G., Merkenschlager, M., and Smale, S. T. (2004). Dynamic assembly of silent chromatin during thymocyte maturation. *Nat. Genet.* 36, 502–506. doi: 10.1038/ng1351

Su, R.-C., Sridharan, R., and Smale, S. T. (2005). Assembly of silent chromatin during thymocyte development. *Seminars Immunol.* 17, 129–140. doi: 10.1016/j.smim.2005.01.003

Tan, C., Hiwa, R., Mueller, J. L., Vykunta, V., Hibiya, K., Noviski, M., et al. (2020). NR4A nuclear receptors restrain B cell responses to antigen when second signals are absent or limiting. *Nat. Immunol.* 21, 1267–1279. doi: 10.1038/s41590-020-0765-7

Tancharoenrat, P., Ravindran, V., Zaefarian, F., and Ravindran, G. (2014). Digestion of fat and fatty acids along the gastrointestinal tract of broiler chickens. *Poultry Sci.* 93, 371–379. doi: 10.3382/ps.2013-03344

Tso, P., Sun, W., and Liu, M. (2004). Gastrointestinal satiety signals IV. *Apolipoprotein* A-IV. *Am. J. Physiol. Gastrointestinal Liver Physiol.* 286, G885–G890.

UEP (2019). *Facts and Stats [Online]. United Egg Producers*. Georgia. Available online at: https://unitedegg.com/facts-stats/#:~:text=U.S.%20table%20egg%20production%20totaled,produces%20294%20eggs%20per%20year (accessed May, 2, 2021).

Van Laethem, F., Tikhonova, A. N., Pobezinsky, L. A., Tai, X., Kimura, M. Y., Le Saout, C., et al. (2013). Lck availability during thymic selection determines the recognition specificity of the T cell repertoire. *Cell* 154, 1326–1341. doi: 10.1016/j.cell.2013.08.009

Wang, M., Mao, Y., Wang, B., Wang, S., Lu, H., Ying, L., et al. (2020). Quercetin improving lipid metabolism by regulating lipid metabolism pathway of ileum mucosa in broilers. *Oxidative Med. Cell. Longevity* 2020, 1–17.

Warren, W. C., Hillier, L. W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., et al. (2017). A new chicken genome assembly provides insight into avian genome structure. *G3 Genes Geno. Genet.* 7, 109–117.

Weiss, H., Friedrich, T., Hofhaus, G., and Preis, D. (1991). The respiratory-chain NADH dehydrogenase (complex I) of mitochondria. *Eur. J. Biochem.* 197, 563–576. doi: 10.1111/j.1432-1033.1991.tb15945.x

Werth, M., Schmidt-Ott, K. M., Leete, T., Qiu, A., Hinze, C., Viltard, M., et al. (2017). Transcription factor TFCP2L1 patterns cells in the mouse kidney collecting ducts. *Elife* 6:e24265.

Wight, T. N., Kang, I., and Merrilees, M. J. (2014). Versican and the control of inflammation. *Matrix Biol.* 35, 152–161. doi: 10.1016/j.matbio.2014.01.015

Wingate, D., Krag, E., Mekhjian, H. S., and Phillips, S. (1973). Relationships between ion and water movement in the human jejunum, ileum and colon during perfusion with bile acids. *Clin. Sci. Mol. Med.* 45, 593–606. doi: 10.1042/cs0450593

Wu, Z., Harne, R., Chintoan-Uta, C., Hu, T.-J., Wallace, R., Maccallum, A., et al. (2020). Regulation and function of macrophage colony-stimulating factor (CSF1) in the chicken immune system. *Dev. Comparat. Immunol.* 105:103586. doi: 10.1016/j.dci.2019.103586

Xing, S., Gai, K., Li, X., Shao, P., Zeng, Z., Zhao, X., et al. (2019). Tcf1 and Lef1 are required for the immunosuppressive function of regulatory T cells. *J. Exp. Med.* 216, 847–866. doi: 10.1084/jem.20182010

Yuan, J., Liu, W., Karvar, S., Baker, S. S., He, W., Baker, R. D., et al. (2015). Potassium channel KCNJ15 is required for histamine-stimulated gastric acid secretion. *Am. J. Physiol. Cell Physiol.* 309, C264–C270.

Zhang, J. D., Hatje, K., Sturm, G., Broger, C., Ebeling, M., Burtin, M., et al. (2017a). Detect tissue heterogeneity in gene expression data with BioQC. *BMC Genom.* 18:277. doi: 10.1186/s12864-017-3661-2

Zhang, T., Zhang, X., Han, K., Zhang, G., Wang, J., Xie, K., et al. (2017b). Analysis of long noncoding RNA and mRNA using RNA sequencing during the differentiation of intramuscular preadipocytes in chicken. *PLoS One* 12:e0172389. doi: 10.1371/journal.pone.0172389

Zhang, Y., Cao, L., Yang, B. L., and Yang, B. B. (1998). The G3 domain of versican enhances cell proliferation via epidermial growth factor-like motifs. *J. Biol. Chem.* 273, 21342–21351. doi: 10.1074/jbc.273.33.21342

Zhang, Y., Wang, J., Huang, S., Zhu, X., Liu, J., Yang, N., et al. (2009). Systematic identification and characterization of chicken (*Gallus gallus*) ncRNAs. *Nucl. Acids Res.* 37, 6562–6574. doi: 10.1093/nar/gkp704

Zhao, C., Inoue, J., Imoto, I., Otsuki, T., Iida, S., Ueda, R., et al. (2008). POU2AF1, an amplification target at 11q23, promotes growth of multiple myeloma cells by directly regulating expression of a B-cell maturation factor, TNFRSF17. *Oncogene* 27, 63–75. doi: 10.1038/sj.onc.1210637

Zhu, J., Yamane, H., and Paul, W. E. (2009). Differentiation of effector CD4 T cell populations. *Ann. Rev. Immunol.* 28, 445–489. doi: 10.1146/annurev-immunol-030409-101212

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.

Copyright © 2021 Overbey, Ng, Catini, Griggs, Stewart, Tkalcic, Hawkins and Drechsler. 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: 07 July 2021
doi: 10.3389/fgene.2021.659287

Watch Out for a Second SNP: Focus on Multi-Nucleotide Variants in Coding Regions and Rescued Stop-Gained

*Fabien Degalez[1†], Frédéric Jehl[1†], Kévin Muret[1], Maria Bernard[2,3], Frédéric Lecerf[1], Laetitia Lagoutte[1], Colette Désert[1], Frédérique Pitel[4], Christophe Klopp[2] and Sandrine Lagarrigue[1]**

[1] INRAE, INSTITUT AGRO, PEGASE UMR 1348, Saint-Gilles, France, [2] INRAE, SIGENAE, Genotoul Bioinfo MIAT, Castanet-Tolosan, France, [3] INRAE, AgroParisTech, Université Paris-Saclay, GABI UMR 1313, Jouy-en-Josas, France, [4] INRAE, INPT, ENVT, Université de Toulouse, GenPhySE UMR 1388, Castanet-Tolosan, France

OPEN ACCESS

Edited by:
Hans Cheng,
Agricultural Research Service (USDA), United States

Reviewed by:
John B. Cole,
United States Department of Agriculture (USDA), United States
Robert D. Schnabel,
University of Missouri, United States

***Correspondence:**
Sandrine Lagarrigue
sandrine.lagarrigue@agrocampus-ouest.fr

[†]These authors share first authorship

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

Received: 27 January 2021
Accepted: 27 May 2021
Published: 07 July 2021

Citation:
Degalez F, Jehl F, Muret K, Bernard M, Lecerf F, Lagoutte L, Désert C, Pitel F, Klopp C and Lagarrigue S (2021) Watch Out for a Second SNP: Focus on Multi-Nucleotide Variants in Coding Regions and Rescued Stop-Gained. Front. Genet. 12:659287. doi: 10.3389/fgene.2021.659287

Most single-nucleotide polymorphisms (SNPs) are located in non-coding regions, but the fraction usually studied is harbored in protein-coding regions because potential impacts on proteins are relatively easy to predict by popular tools such as the Variant Effect Predictor. These tools annotate variants independently without considering the potential effect of grouped or haplotypic variations, often called "multi-nucleotide variants" (MNVs). Here, we used a large RNA-seq dataset to survey MNVs, comprising 382 chicken samples originating from 11 populations analyzed in the companion paper in which 9.5M SNPs— including 3.3M SNPs with reliable genotypes—were detected. We focused our study on in-codon MNVs and evaluate their potential mis-annotation. Using GATK HaplotypeCaller read-based phasing results, we identified 2,965 MNVs observed in at least five individuals located in 1,792 genes. We found 41.1% of them showing a novel impact when compared to the effect of their constituent SNPs analyzed separately. The biggest impact variation flux concerns the originally annotated stop-gained consequences, for which around 95% were rescued; this flux is followed by the missense consequences for which 37% were reannotated with a different amino acid. We then present in more depth the rescued stop-gained MNVs and give an illustration in the *SLC27A4* gene. As previously shown in human datasets, our results in chicken demonstrate the value of haplotype-aware variant annotation, and the interest to consider MNVs in the coding region, particularly when searching for severe functional consequence such as stop-gained variants.

Keywords: MNV, SNP, variation, rescued stop-gained, *SLC27A4*, FATP4

INTRODUCTION

Next-generation sequencing has given access to genomes at the nucleotide level through DNA-seq but also specifically to expressed regions by whole-exome sequencing (WES, originally focusing on exonic parts of the genome) or RNA-seq. These data enable us to call genetic variations by spotting differences between aligned reads and the species reference genome or among aligned

reads. Among these genetic variations, single-nucleotide polymorphisms (SNPs) are the most frequent and most studied variations. Although most variations are located in non-coding regions, the most analyzed lie in protein-coding regions where their potential impact(s) on the protein are relatively easy to predict. For example, in a study using 60,706 human exomes, the Exome Aggregation Consortium (ExAC) identified 3,230 genes with near-complete depletion of predicted protein-truncating variants. Of these genes, 72% have not been related to any known human disease phenotype (Lek et al., 2016). Different popular tools have been developed this last decade to predict SNPs' effects on proteins such as *Variant Effect Predictor* (VEP) (McLaren et al., 2016), SnpEff (Cingolani et al., 2012), or *ANNOtate VARiation* (ANNOVAR) (Wang et al., 2010). But these tools consider each variation location individually, as if it they were specific to "reference" nucleotides. However, SNPs can be grouped by two or more coexisting variants present in the same haplotype (in the same individual), in which case they are called "multi-nucleotide variants" (MNVs). An example of MNV in one individual (with two nearby SNPs) is given in **Figure 1**. When such MNVs occur within a codon, the amino acid modification caused by this MNV may be different from protein change resulting from each constituent SNPs taken individually, leading to a risk of erroneous functional consequence prediction, as depicted in **Figure 1**. MNV identification tools have been developed using different methods for phasing SNPs [MAC (Wei et al., 2015), varDic (Lai et al., 2016), COPE (Cheng et al., 2017), BCFtools (Danecek and McCarthy, 2017), and MACARON (Khan et al., 2018)] and have been applied to different human genetic variant datasets [1,000 Genomes Project dataset (Cheng et al., 2017; Danecek and McCarthy, 2017; Khan et al., 2018; Wang et al., 2020), ExAC (Lek et al., 2016), The Cancer Genome Atlas (Lai et al., 2016), or gnomAD consortium (Wang et al., 2020)], mainly based on exomes.

To our knowledge, no study has been conducted on MNVs in livestock species. The aim of this paper is to focus on MNVs occurring in protein-coding regions to provide examples and evaluate the functional consequences of resulting mis-annotations. Considering this aim, we used 9.5M SNPs recently detected in 382 chickens from 767 multi-tissue RNA-seq, enriched by construction in expressed regions and therefore in protein-coding regions. From this 9.5M SNPs, we focused on the 3.3M SNPs with reliable genotypes [see the companion paper (Jehl et al., 2021)]. MNV identification requires properly phased variants, i.e., to be located either on the same haplotype (called therefore MNV) or on two different haplotypes (a case of individual SNPs) (see **Figures 1A,B**). Different SNP phasing strategies exist: (*i*) population-based phasing, using statistical inference of phase from haplotypes shared among individuals of a large genotyped population; (*ii*) family-based phasing, which analyzes the co-transmission of variants between parents and offsprings; and (*iii*) read-based phasing, which evaluates whether close variants are present on the same reads in the DNA-seq or RNA-seq data. Read-based phasing is particularly relevant for close variants, making this method appropriate for MNV analysis in codons, in which variants fall within a maximum distance of 2 bp from one another. Therefore, in this study, we have chosen to identify MNVs by using the read-based phasing provided by the HaplotypeCaller tool of the GATK toolkit, in the VCF file through additional fields (**Figure 1C**, with the PID and PGT fields) recently added by the common variant caller (DePristo et al., 2011; Van der Auwera et al., 2013; McKenna et al., 2020).

MATERIALS AND METHODS

SNP Dataset

The 3,276,615 SNPs analyzed in this study have been detected following the method presented in the companion paper (Jehl et al., 2021) using 767 multi-tissue RNA-seq of 382 birds from 11 chicken populations (see **Additional File 1**). This SNP set corresponds to the union of the SNPs with reliable genotypes found in each population (list available on http://www.fragencode.org/lnchickenatlas.html). Briefly, variant detection was performed for each sample using the HaplotypeCaller tool of GATK toolkit (DePristo et al., 2011; Van der Auwera et al., 2013; McKenna et al., 2020) 3.7.0 with options "--stand_call_conf 20.0," "--min_base_quality_score 10," and "--min_mapping_quality_score 20" (which are the defaults values). The "GenotypeGVCFs" function was then used with the option "--stand_call_conf 20.0," to jointly genotype all these samples into one VCF per tissue. First, biallelic SNPs were then extracted using the "SelectVariant" function with option "--selectType SNP –restrictAllelesTo BIALLELIC." Variants were filtered using "VariantFiltration" with "QD $<$ 2" and "FS $>$ 30." Considering genotypes, variants were selected with a "(5.reads.DP) genotype CR $\geq$ 20%" and a "CR $\geq$ 50%." The 11 populations include a red jungle fowl population (RJFh), three broiler populations with one experimental line (FLLL) and two commercial ones (Cobb, HerX), six layer populations with two brown-egg commercial lines (Novo1 and Novo2), two brown-egg experimental populations (RpRm and LSnu), two white-egg or cream-egg experimental populations (FrAg and FAyo), and finally a cross between white- and brown-egg experimental lines (Rmx6).

Analysis of the Functional Impact of Each Individual SNP in the Coding Regions

VEP v92 (McLaren et al., 2016) with a GTF file enriched in long non-coding genes ("--gtf") was used for effect prediction of each SNP, with "--everything" and "--total_length" options to respectively, obtain SIFT score predictions and lengths of cDNA, CDS, and protein positions (Ng and Henikoff, 2003; Sim et al., 2012).

MNV Calling and Recalculation of Consequences

The script to detect the MNV and to calculate the consequences is available in **Additional File 2**.

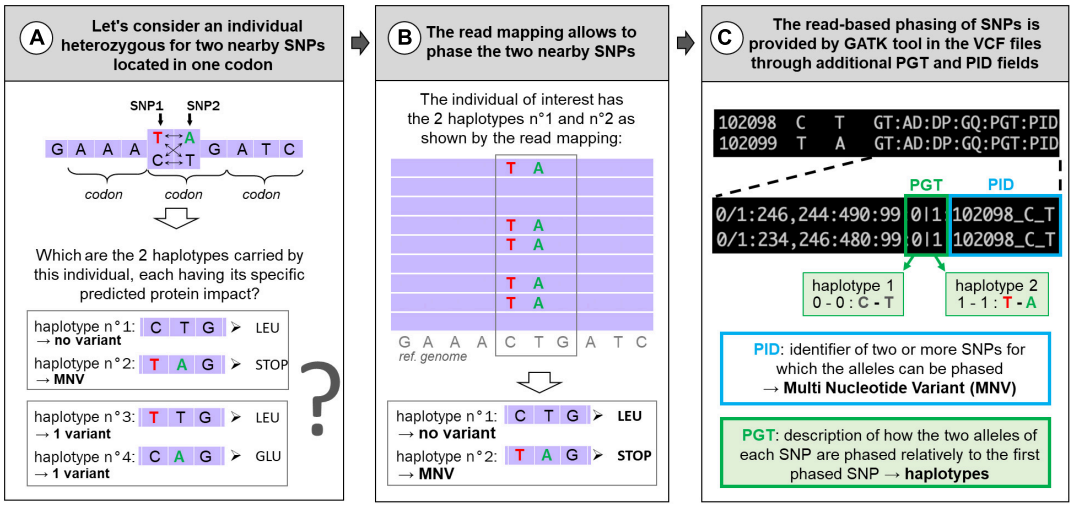


FIGURE 1 | Example of MNVs: predicted impact on the associated protein **(A)** and how to identify them **(B,C)**. **(A)** Example of an MNV composed of two nearby SNPs in one codon and its four potential haplotypes in the population and their predicted impact on the associated protein. In contrast to other haplotypes, haplotype no. 2 contains two variants (T and A) and corresponds to an MNV. **(B)** The IGV (Integrated Genome Browser; Robinson et al., 2011) screen shot indicates the principle of read-based phasing of SNPs: short read mapping against the reference genome of the heterozygous individual allows us to phase both SNPs giving two haplotypes: C with T (reference alleles) on one side and T with A (alternative alleles) on the other side. When translated, these two haplotypes correspond to a leucine or a STOP codon and not to a simple amino acid change (LEU → GLU) if the two haplotypes had been composed by only one reference and one variant as shown in **(A)**. **(C)** Information (PID and PGT) provided by GATK in the VCF files about the phased SNPs according to the read-based phasing shown in **(B)**.

Detection of SNPs Located in the Same Codon

With the information produced by VEP, an ID composed of the "transcriptID" and the "position of the SNP in the coding sequence (expressed in codon number)" was created for each consequence. Through this approach, the same codon of the same transcript supporting at least two different SNPs will have the same ID. Thus, only duplicated IDs were kept as they correspond to those containing two or more SNPs.

Detection of Co-located SNPs Carried by the Same Haplotype (MNV)

To test if the SNPs located in the same codon were also present in the same haplotype, the VEP file generated in the previous step and the VCF file were joined by the "SNPid" key, equivalent to "CHR_POS_REF/ALT." The resulting file (VEP merged to VCF information) contained SNPs on the same codon with additional information about their phase (PID and PGT). Finally, the SNPs which were phased (i.e., same PID) and co-located in a codon were extracted: they correspond to MNVs containing two or three phased SNPs in the same codon.

Recalculation of the Consequences

With the R package Biostrings v2.50.2 (Pagès et al., 2021), the associated amino acids were produced for each MNV, and with the same strategy as VEP being adopted, the MNV consequences were established.

Analysis of MNV Functional Impacts and Comparison With the Constituent SNP Impacts

To compare MNV and independent SNP consequences, we selected only the most impactful consequence per codon for these constituent SNPs using the following order of priority from severe to weak consequences: (1) stop-gained, (2) stop-lost, (3) start-lost, (4) missense_variant, (5) stop-retained_variant, and (6) synonymous_variant.

For MNVs with a missense annotation corresponding to a missense annotation for both constituent SNPs, we distinguished two cases:

- missense MNV with an amino acid different from those predicted by the constituent SNPs (SNP1: Missense A; SNP2: Missense B → MNV: Missense C) and
- missense MNV with an amino acid common to one of two amino acids predicted by the constituent SNPs (SNP1: Missense A; SNP2: Missense B → MNV: Missense A or B).

In order to visualize the results, we produced an alluvial plot using the R "alluvial" package v0.1-2 (Bojanowski, 2020).

GO or KEGG Term Enrichment Analysis

The enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) terms in the gene set of interest was performed using the STRING v11.0. tool (Szklarczyk et al., 2019), and a GO or KEGG term was found significantly enriched if the BH-adjusted $p \leq 5\%$.

DNA Sequencing of *SLC27A4*

Five microliters of DNA samples was mixed with 5 μl of GoTaq Flexi Buffer 5 × , 2 μl of $MgCl_2$ solution (25 mM), 0.125 μl of GoTaq DNA polymerase (5 U/μl) (Promega, catalog number: M891), 0.5 μl of dNTPs 10 mM, 12.5 μl H_2O, and 1.25 μl of specific reverse (CATTCCCGTAGTGCCAGAGG) and forward primers (GCACTTTCTGGTGCAAAGCA) at 10 μM. Reaction mixtures were then incubated in a T100 thermal cycler (Bio-Rad, Marne la Coquette, France) for 30 cycles with 30 s at 94°C, 30 s at 60°C, and 30 s at 72°C. The amplification products were then deposited on a 2% agarose gel and sent for sequencing (Genoscreen) to verify their location on the chicken genome.

RESULTS AND DISCUSSION

Read-Based Phasing for Identification of MNVs

Using 3.3M SNPs previously detected from 767 multi-tissue RNA-seq of 382 animals from 11 chicken populations and therefore enriched in coding regions [see the companion paper (Jehl et al., 2021), section "Materials and Methods"], we identified 260,919 unique SNPs in 26,702 transcripts corresponding to 15,835 genes out of 19,545 protein-coding genes (**Figure 2**, right part—in yellow).

As shown in **Figure 2** (left part—in green), we then defined an MNV in a codon as a group of two or three phased SNPs, i.e., existing on the same haplotype in the same individual. We found 11,183 SNPs (4.3% of the SNPs in codons) as constituent variants of 5,533 MNVs, which corresponded to 4,415 transcripts and 2,916 genes. Most of them (98%: 5,416) contained two SNPs with similar proportions (1/3) by constituent SNP position in the codon (1–3, 1–2, and 2–3, **Figure 2**, left). In order to ensure the reliability of the MNVs, we selected MNVs observed in at least five individuals. Out of the 5,416 MNVs with two SNPs, 2,965 MNVs were present in at least five individuals, corresponding to 2,636 transcripts and 1,792 genes. No GO terms or KEGG terms were found as significantly enriched for this gene list, suggesting that no specific biological pathway was impacted by MNVs. **Table 1** gives the distribution of MNVs and their consequences according to the individual number supporting the MNV (ranging from 2 to 100 individuals). We can note that 31% of the 5,416 MNVs with two SNPs are observed only in a single individual and are here considered as erroneous, likely due to sequencing errors.

Functional Impact Comparison of MNVs and Their Component SNPs

Focusing on the 2,965 MNVs present in at least five individuals, we then compared their functional consequences with those of the 5,930 constituent SNPs as illustrated by the right part of the workflow provided in **Figure 2**. For such a comparison, we retained for each MNVs, the most severe consequence of the constituent SNPs according to the order indicated in **Figure 2** (bottom right). The alluvial plot in **Figure 3** depicts the consequence variations before (left) and after (right) taking the MNV impacts into account, according to the different consequence categories; the details of impact variation per MNV are given in **Additional File 3** for the whole 2,965 MNV set. We can observe in **Figure 3** that the biggest change in variant impacts concerns the originally stop-gained consequence categories, for which 95.6% were re-predicted as missense (green flux: 87 out of the 91 stop-gained initially predicted). The second and third biggest fluxes concern missense consequence categories, for which 37.3% had a different predicted amino acid (violet flux: 1,038 MNVs out of the 2,780 initial missenses), and 3.0% became synonymous variants (blue flux: 83 MNVs out of the initial missenses). The distribution of re-prediction fluxes is provided in **Table 1** as a function of the individual number supporting the MNV among the 382 individuals analyzed. Among the 87 rescued stop-gained observed in five individuals, half (47) are observed in at least 15 individuals and are present on average in five populations (see **Additional File 4**). Out of the MNVs, the proportion of rescued stop-gained MNVs (2.9%), defined as at least one of the individual SNPs creating a nonsense mutation but not the resulting MNV, is in the same order of magnitude as the one reported by the gnomAD consortium with 1,821 rescued stop-gained MNVs out of 31,575 human MNVs (5.8%) (Wang et al., 2020). Genes with a stop-gained MNV rescued in the missense variant are available in **Additional File 4** with the population affected and the individual number per population carrying these MNVs. To a lesser extent, nine missenses were re-predicted as stop-gained, which would have gone unnoticed without re-prediction. After a deeper investigation with the IGV browser, these re-predicted stop-gained variants seem to be present since they were not located in a potential exon skipping. Finally, this stop-gained category drastically declined by 86% (from 91 to 13) after considering MNVs, whereas the synonymous category was increased by twofold (from 79 to 159). These different category changes after considering MNVs have a major impact on variant interpretation and thus are critical for accurate variant annotation. More broadly, when the MNVs were considered together, the resulting functional impact differed from the independent impacts of the individual variants in 41.1% of the analyzed MNVs. This large percentage of mis-annotations is relatively consistent with ∼60% of reannotations in human MNVs recently reported by the gnomAD consortium in coding regions (Wang et al., 2020). Such results show the importance of paying attention to these MNVs as highlighted by McLaren et al. (2016): "Current annotation tools, including the VEP, annotate each input variant independently, without considering the potential compound effects of combining alternate alleles across multiple variant loci."

Example of an Erroneously Predicted Stop-Gained

As an example of erroneously predicted stop-gained, we present the case of the *SLC27A4* gene, which is located on the reverse strand of chicken chromosome 17 (ENSGALG00000004965) (**Figure 4A**). In this gene, two SNPs rs316701182 and rs15031398, already reported in the Ensembl SNP database (Ensembl, 2018), were respectively, predicted as a stop-gained variant

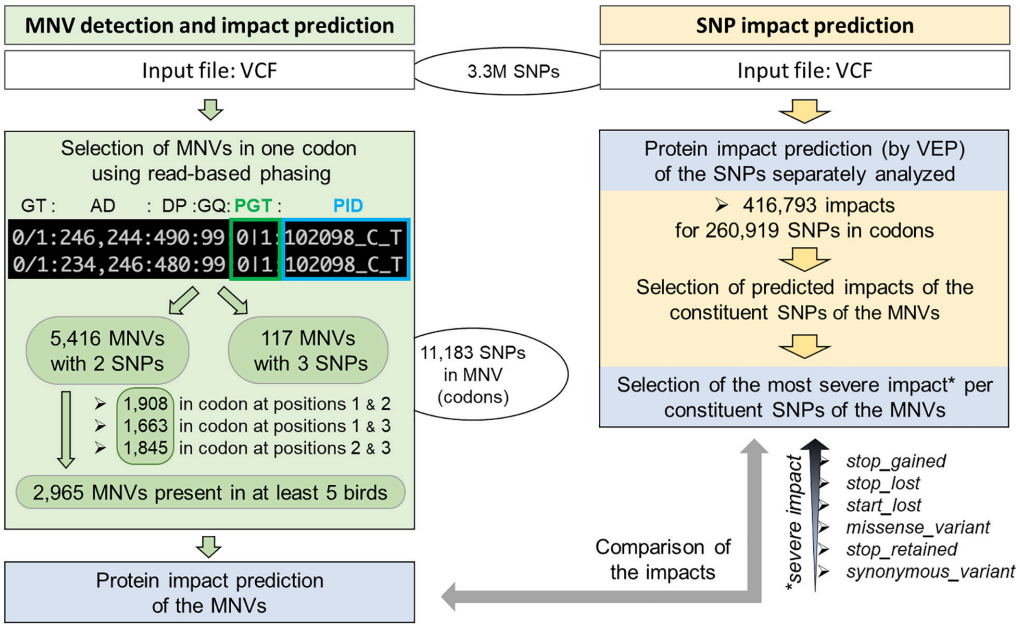


FIGURE 2 | Workflow of MNV detection in coding regions and functional consequence prediction. Left: MNV detection from 3.3M SNPs previously identified using RNA-seq of 382 chickens (companion article; Jehl et al., 2021). Right: MNV constituent SNP selection and protein impact selection of these SNPs separately analyzed by VEP.

TABLE 1 | Occurrences for each type of re-prediction according to the number of individuals carrying the MNV.

SNP annotation	→	MNV annotation	Number of individuals carrying the MNV										
			1	**2**	**3**	**4**	**5**	**10**	**15**	**20**	**30**	**50**	**100**
	Eased impact												
Missense	→	Synonymous	110	91	86	86	**83**	76	74	74	72	66	54
Stop_gained	→	Missense	194	118	102	99	**87**	63	47	39	33	27	14
	Equal impact												
Stop_lost	→	Stop_lost	8	4	4	4	**4**	2	2	2	2	1	0
Start_lost	→	Start_lost	19	11	11	11	**11**	8	7	6	5	3	1
Synonymous	→	Synonymous	95	84	81	78	**76**	69	63	58	54	45	32
Missense	→	Missense	4,932	3,425	3,107	2,854	**2,688**	2,162	1,876	1,652	1,369	1,083	716
Stop_gained	→	Stop_gained	12	6	5	5	**4**	3	2	2	2	1	0
	Aggravated impact												
Stop_retained	→	Stop_lost	1	0	0	0	**0**	0	0	0	0	0	0
Synonymous	→	Missense	15	6	6	4	**3**	2	1	1	1	1	1
Missense	→	Stop_gained	30	13	11	10	**9**	6	4	3	3	2	1
		Total MNVs	5,416	3,758	3,413	3,151	**2,965**	2,391	2,076	1,837	1,541	1,229	819

*The column in bold corresponds to MNVs observed in at least five individuals (MNV set used in **Figure 3**).*

(TGA; stop-gained) and a synonymous variant (CGC; arginine) when compared to the reference haplotype (CGA; arginine) (**Figure 4B**). These SNPs were present in the FLLL population with frequencies > 20% and interestingly with contrasted frequencies between FL (fat line) and LL (lean line), two subpopulations divergently selected for adipose tissue weight (Leclercq et al., 1980). The rs15031398 SNP is absent in FL (**Figure 4B**); in the LL population in which we observed both SNPs (**Figure 4C**), we did not find any animal with the TGA (stop-gained) haplotype (composed of one variant only), with the rs316701182 T variant being always associated with the rs15031398 C variant within the "TGC" MNV. The absence of TGA (stop-gained) haplotype is consistent with several *SLC27A4*-knockout mouse studies which report prenatal lethality (Gimeno et al., 2003) or neonatal lethality (Herrmann et al., 2003; Moulson et al., 2003; Lin et al., 2010; Tao et al., 2012). The *SLC27A4* gene codes fatty acid transport protein 4 (FATP4), which is particularly involved in the uptake of long-chain fatty acids

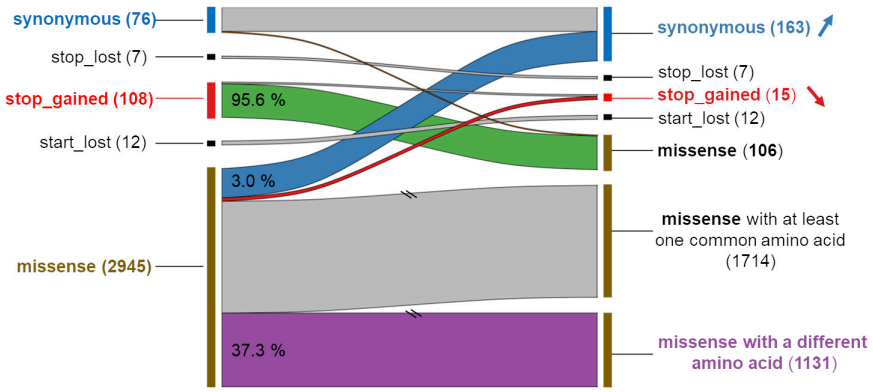
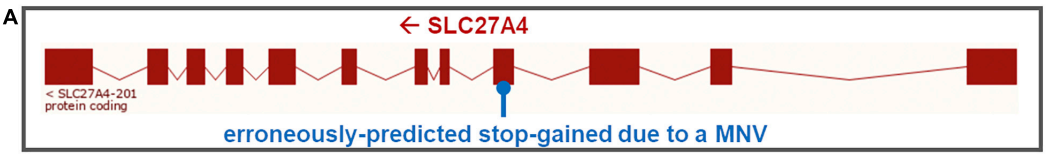


FIGURE 3 | Comparison of the functional impact of MNVs (right) and their component SNPs (left) for each of the 2,965 MNVs. Left: The consequence originally predicted for the component SNPs, the most severe impact being retained by the codon (see **Figure 2** or the text for the order). Right: The new prediction associated with the MNVs. For each category of functional predictions of the component SNPs (left), the numbers and percentages are given with new predictions due to the associated MNV. The two slashes indicate that the scale has been adapted (reduction by five times) for better readability.

A



B

SNP_ID:	rs316701182	rs15031398	population					
position:	17:5515128	17:5515126	FLLL		FL		LL	
	SNP1	SNP2	SNP1	SNP2	SNP1	SNP2	SNP1	SNP2
REF	C	A	75%	62.5%	92%	100%	58%	25%
ALT	T	C	25%	37.5%	8%	0%	42%	75%
codon	Cga -> Tga	cgA -> cgC						
impact	arg-> STOP	arg -> arg						

C

haplotype			FL	LL
haplotype with constituent variant	codon	effect	observed freq.	
haplotype with no variation	CgA	arginine	92%	25%
haplotype with SNP2	CgC	arginine	8%	33%
haplotype with SNP1	TgA	**STOP**	**0%**	**0%**
haplotype with 2 SNPs = MNV	TgC	**cysteine**	0%	**42%**

	FL	LL
codon	*theoretical freq.*	
CgA	*92%*	*15%*
CgC	*8%*	*44%*
TgA	*0%*	*10%*
TgC	*0%*	*31%*

D

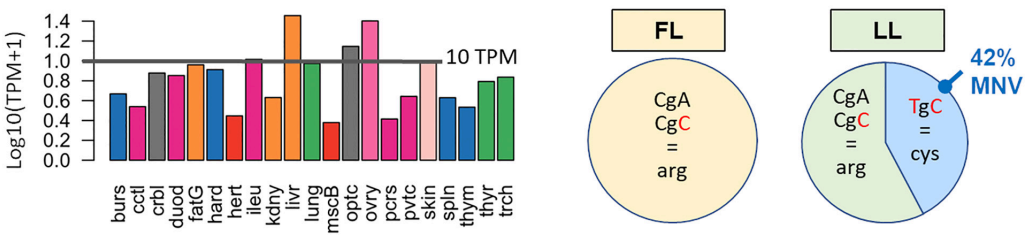


FIGURE 4 | *SLC27A4* with an MNV composed of two phased SNPs observed in the experimental divergent lean line (LL). **(A)** Exon structure of the *SLC27A4* gene and the MNV location. **(B)** For the two SNPs (SNP1: rs316701182 and SNP2: rs15031398) related to the MNV of interest, the allele position (on the Galgal5 genome), functional impact on the associated protein, and frequencies in the FLLL population are indicated, and the two FL and LL subpopulations are divergently selected on abdominal fat weight. **(C)** Effects of the four haplotypes related to SNP1 and SNP2 separately analyzed by VEP and frequencies in LL (n = 12) and FL (n = 12) subpopulations and focus on the percentage of observed haplotypes in the two FL and LL subpopulations. The haplotypes were determined through the IGV browser of mapped RNA-seq reads against the chicken genome. **(D)** Tissue expression of the gene in a chicken RNA-seq dataset composed of 21 tissues (Jehl et al., 2020).

(LCFAs); this gene is highly expressed in various chicken tissues as shown in **Figure 4D** with an expression > 10 TPM in the liver, ovary, optical system, skin, and intestine (ileum). Interestingly, FATP4 is thought to play a major role in dietary fatty acid uptake in intestinal epithelial cells (Hirsch et al., 1998) and in physiological uptake across cell membranes of LCFAs, which are key metabolites for energy generation and storage; it is viewed as a target to prevent or reverse obesity (Hirsch et al., 1998; Schaffer, 2002). FATP4 could be then related to the lean phenotype of the LL population for two reasons. First, the "TGC" (cysteine) MNV haplotype is reported as a severe change by the SIFT software package compared to the reference "CGA" (arginine) haplotype, suggesting a severe impact on the FATP4 protein function. Second, this "TGC" MNV haplotype is absent in FL birds, whereas it is frequent (42%) in 12 LL birds, with a higher frequency than expected (**Figure 4C**). We confirmed these results by extending this analysis to 58 birds (29 birds per line) using PCR amplification of the region of interest followed by Sanger sequencing. No rs15031398 was identified in the FL line. In the LL line, we observed 12 birds carrying the "TGC" MNV haplotype (three homozygous and nine heterozygous) and no bird with the TGA (stop-gained) haplotype. These results suggest a strong but not lethal impact of the MNV haplotype on the FATP4 protein function, which could then participate to the lean phenotype of the LL line. However, a genetic association study is needed to support a potential causal link between the FATP4 dysfunctional MNV and a low adiposity in the LL line compared to the FL line.

CONCLUSION

We have shown that MNVs represent an important class of genetic variations since they have a significant impact on polymorphism functional interpretation with roughly 40% of MNVs in our dataset inducing reannotation. These reannotations show a decreased impact severity of MNVs when compared to their constituent SNPs, at least for the stop-gained category. As previously demonstrated in human studies, our results in chicken demonstrate the value of haplotype-aware variant annotation and the interest to consider MNVs in coding region particularly when focusing on severe functional consequences such as stop-gained. We illustrated such a case with an erroneous stop-gained annotation found in the chicken *SLC27A4* gene.

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

FD, FJ, and SL conceived the study and drafted the manuscript. FD, FJ, FL, CK, and SL participated to in bioinformatics processing of the RNA-seq data and bioinformatic analyses. LL, CD, and KM participated in the IGV analysis, PCR amplification, and Sanger sequencing for *SLC27A4* analyses. KM, FP, FL, and CK helped to improve the manuscript. All authors read and approved the final version.

FUNDING

The sample and/or data were collected in the frame of projects that received financial support from the European Union's H2020 Program under Grant Agreement No. 633531 (Feed-a-Gene project), the French National Agency of Research (FatInteger project ANR-11-SVS7; ChickStress project, ANR-13-ADAP; EpiBird ANR project PCS-09-GENM-010), and the French institutions Institut Agro (AGROCAMPUS OUEST) and INRAE [Fr-AgENCODE project (2015–2017) and ELASTICe project (2012)]. FJ, FD, and KM were Ph.D. fellows supported by the Brittany region (France) and the INRAE Animal Genetics division. These funding bodies had no role in the design of the study; in the collection, analysis, and interpretation of data; or in the writing of the manuscript.

ACKNOWLEDGMENTS

We are grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, doi: 10.15454/1.5572369328961167E1) for providing help and/or computing and/or storage resources. We thank to Morgane Boutin who passed away this year and who did a lot in *SLC27A4* analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.659287/full#supplementary-material

Additional File 1 | Detail per population of the number of birds, tissues, and samples. The median number of sequenced fragments (× 1E6) is also indicated for each population.

Additional File 2 | Script used for recalculating the consequences for MNVs made up of two SNPs. The script is not generalized and therefore requires adaptations from the users. It is written in bash and R; the change of language is clearly indicated.

Additional File 3 | List of MNVs with two phased SNPs (position on the Galgal5 chicken genome) identified among the 382 chickens from 11 populations with their functional impacts (newConseq) and their constituent SNPs (oldConseq). **(A)** List of 2,965 MNVs identified in at least five individuals. **(B)** List of 5,416 MNVs without filtering on the number of individuals supporting the MNV. For the two tables, the population number (popNb), the total number of individuals (indNb_total), and the population name with the individual number per population (popName_indNb), in which the MNVs have been observed, are provided. AA, amino acid.

Additional File 4 | List of genes presenting a rescued stop-gained according to the number of individuals carrying the MNV. **(A)** Summary of the number of genes detected as "rescued" according to the number of individuals supporting the MNV. **(B)** For each gene, the population number (popNb) and the population name(s) with the individual number per population (popName:indNb) where the MNV is observed are indicated. The column in blue corresponds to MNVs observed in at least five individuals, the MNV set used in **Figure 3**.

REFERENCES

Bojanowski, M. (2020). *mbojan/alluvial. R*. Available online at: https://github.com/mbojan/alluvial. (accessed 8 Jan 2021).

Cheng, S.-J., Shi, F.-Y., Liu, H., Ding, Y., Jiang, S., Liang, N., et al. (2017). Accurately annotate compound effects of genetic variants using a context-sensitive framework. *Nucleic Acids Res.* 45:e82. doi: 10.1093/nar/gkx041

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92. doi: 10.4161/fly.19695

Danecek, P., and McCarthy, S. A. (2017). BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* 33, 2037–2039. doi: 10.1093/bioinformatics/btx100

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498.

Ensembl. (2018). *dbSNP - Gallus Gallus 5 - V94*. ftp://ftp.ensembl.org/pub/release-94/variation/vcf/gallus_gallus (accessed December 11, 2020).*

Gimeno, R. E., Hirsch, D. J., Punreddy, S., Sun, Y., Ortegon, A. M., Wu, H., et al. (2003). Targeted deletion of fatty acid transport protein-4 results in early embryonic lethality. *J. Biol. Chem.* 278, 49512–49516. doi: 10.1074/jbc.m309759200

Herrmann, T., van der Hoeven, F., Gröne, H.-J., Stewart, A. F., Langbein, L., Kaiser, I., et al. (2003). Mice with targeted disruption of the fatty acid transport protein 4 (Fatp 4, Slc27a4) gene show features of lethal restrictive dermopathy. *J. Cell Biol.* 161, 1105–1115. doi: 10.1083/jcb.200207080

Hirsch, D., Stahl, A., and Lodish, H. F. (1998). A family of fatty acid transporters conserved from mycobacterium to man. *Proc Natl Acad Sci U S A.* 95, 8625–8629. doi: 10.1073/pnas.95.15.8625

Jehl, F., Degalez, F., Bernard, M., Lecerf, F., Lagoutte, L., Désert, C., et al. (2021). RNA-Seq data for reliable SNP detection and genotype calling: Interest for coding variant characterization and cis-regulation analysis by allele-specific expression in livestock species. *Front. Genet.* 12:655707. doi: 10.3389/fgene.2021.655707

Jehl, F., Muret, K., Bernard, M., Boutin, M., Lagoutte, L., Désert, C., et al. (2020). An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. *Sci. Rep.* 10:20457.

Khan, W., Varma Saripella, G., Ludwig, T., Cuppens, T., Thibord, F., Génin, E., et al. (2018). MACARON: a python framework to identify and re-annotate multi-base affected codons in whole genome/exome sequence data. *Bioinformatics* 34, 3396–3398. doi: 10.1093/bioinformatics/bty382

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., et al. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* 44:e108. doi: 10.1093/nar/gkw227

Leclercq, B., Blum, J. C., and Boyer, J. P. (1980). Selecting broilers for low or high abdominal fat: Initial observations. *Br. Poult. Sci.* 21, 107–113. doi: 10.1080/00071668008416644

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291.

Lin, M.-H., Chang, K.-W., Lin, S.-C., and Miner, J. H. (2010). Epidermal hyperproliferation in mice lacking fatty acid transport protein 4 (FATP4) involves ectopic EGF receptor and STAT3 signaling. *Dev. Biol.* 344, 707–719. doi: 10.1016/j.ydbio.2010.05.503

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2020). *The Genome Analysis Toolkit: A Mapreduce Framework For Analyzing Next-Generation Dna Sequencing Data*. Available online at: https://genome.cshlp.org/content/20/9/1297.long. (accessed 18 Sep 2020).

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.

Moulson, C. L., Martin, D. R., Lugus, J. J., Schaffer, J. E., Lind, A. C., and Miner, J. H. (2003). Cloning of wrinkle-free, a previously uncharacterized mouse mutation, reveals crucial roles for fatty acid transport protein 4 in skin and hair development. *Proc. Natl. Acad. Sci. U.S.A.* 100, 5274–5279. doi: 10.1073/pnas.0431186100

Ng, P. C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2021). *Biostrings: Efficient Manipulation Of Biological Strings. Bioconductor Version: Release (3.12)*. doi: 10.18129/B9.bioc.Biostrings

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Schaffer, J. E. (2002). Fatty acid transport: the roads taken. *Am. J. Physiol. Endocrinol. Metab.* 282, E239–E246.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40, W452–W457.

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, D607–D613.

Tao, J., Koster, M. I., Harrison, W., Moran, J. L., Beier, D. R., Roop, D. R., et al. (2012). A spontaneous fatp4/scl27a4 splice site mutation in a new murine model for congenital ichthyosis. *PLoS One.* 7:e50634. doi: 10.1371/journal.pone.0050634

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43, 11.10.1–11.10.33.

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

Wang, Q., Pierce-Hoffman, E., Cummings, B. B., Alföldi, J., Francioli, L. C., Gauthier, L. D., et al. (2020). Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat. Commun.* 11:2539.

Wei, L., Liu, L. T., Conroy, J. R., Hu, Q., Conroy, J. M., Morrison, C. D., et al. (2015). MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC Genomics.* 16:569. doi: 10.1186/s12864-015-1779-7

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.

Copyright © 2021 Degalez, Jehl, Muret, Bernard, Lecerf, Lagoutte, Désert, Pitel, Klopp and Lagarrigue. 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: 15 July 2021
doi: 10.3389/fgene.2021.683408

PacBio Iso-Seq Improves the Rainbow Trout Genome Annotation and Identifies Alternative Splicing Associated With Economically Important Phenotypes

*Ali Ali[1], Gary H. Thorgaard[2] and Mohamed Salem[1]**

[1] Department of Animal and Avian Sciences, University of Maryland, College Park, College Park, MD, United States, [2] School of Biological Sciences and Center for Reproductive Biology, Washington State University, Pullman, WA, United States

Rainbow trout is an important model organism that has received concerted international efforts to study the transcriptome. For this purpose, short-read sequencing has been primarily used over the past decade. However, these sequences are too short of resolving the transcriptome complexity. This study reported a first full-length transcriptome assembly of the rainbow trout using single-molecule long-read isoform sequencing (Iso-Seq). Extensive computational approaches were used to refine and validate the reconstructed transcriptome. The study identified 10,640 high-confidence transcripts not previously annotated, in addition to 1,479 isoforms not mapped to the current Swanson reference genome. Most of the identified lncRNAs were non-coding variants of coding transcripts. The majority of genes had multiple transcript isoforms (average ∼3 isoforms/locus). Intron retention (IR) and exon skipping (ES) accounted for 56% of alternative splicing (AS) events. Iso-Seq improved the reference genome annotation, which allowed identification of characteristic AS associated with fish growth, muscle accretion, disease resistance, stress response, and fish migration. For instance, an ES in *GVIN1* gene existed in fish susceptible to bacterial cold-water disease (BCWD). Besides, under five stress conditions, there was a commonly regulated exon in prolyl 4-hydroxylase subunit alpha-2 (*P4HA2*) gene. The reconstructed gene models and their posttranscriptional processing in rainbow trout provide invaluable resources that could be further used for future genetics and genomics studies. Additionally, the study identified characteristic transcription events associated with economically important phenotypes, which could be applied in selective breeding.

Keywords: rainbow trout, transcriptome, PacBio, Iso-Seq, long reads, alternative splicing, alternative polyadenylation, exon usage

OPEN ACCESS

Edited by:
Hans Cheng, Agricultural Research Service (USDA), United States

Reviewed by:
Surendra Kumar, Memorial University of Newfoundland, Canada
Agusto R. Luzuriaga Neira, University of Nevada, Reno, United States

****Correspondence:***
Mohamed Salem
mosalem@umd.edu

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

Received: *20 March 2021*
Accepted: *14 June 2021*
Published: *15 July 2021*

Citation:
Ali A, Thorgaard GH and Salem M (2021) PacBio Iso-Seq Improves the Rainbow Trout Genome Annotation and Identifies Alternative Splicing Associated With Economically Important Phenotypes. Front. Genet. 12:683408. doi: 10.3389/fgene.2021.683408

INTRODUCTION

Rainbow trout is one of the most important fish species that significantly contributes to the aquaculture industry of the United States and has been extensively used as a model organism for biomedical research. International efforts have been ongoing over the years to develop genomic and transcriptomic resources for this species (Salem et al., 2010, 2015; Ali et al., 2014;

Berthelot et al., 2014; Al-Tobasei et al., 2016). The Sanger sequencing approach has been considered as the gold standard for sequencing full-length (FL) cDNA clones and genome annotation (Denoeud et al., 2008). This approach was previously used with the 454 pyrosequencing technology to assemble the rainbow trout transcriptome yielding transcripts with an average length below 1 kb (Salem et al., 2010). Sanger sequencing fell behind when cheaper short-read technologies came out to refine the rainbow trout transcriptome (Fox et al., 2014; Salem et al., 2015). The rainbow trout genome assembly (Berthelot et al., 2014), released in 2014, failed to completely cover and adequately anchor a high percentage of genes to chromosomes. More recently, the genome assembly and gene spaces were further refined (Pearse et al., 2020). Despite the accumulation of massive short-read data over recent years, the lack of FL transcripts has been a significant limitation to define alternatively spliced and polyadenylated transcripts leading to incorrect or incomplete gene annotations (Au et al., 2013; Abdel-Ghany et al., 2016). Transcript reconstruction methods for short reads achieved good precision at the exon level, but the accuracy was low to assemble complete transcripts even in species with simple transcript structures (Steijger et al., 2013). Short reads can accurately identify splice sites but are limited to infer splice site usage and discover transcript isoforms (Steijger et al., 2013; Wang et al., 2016).

Alternative splicing (AS) is a predominant phenomenon in eukaryotic genomes that increases the repertoire of proteins without increasing the number of genes (reviewed in Kornblihtt et al., 2013). In humans, ∼95% of the multi-exonic genes undergo AS (Pan et al., 2008; Barash et al., 2010) and, thus, facilitate the evolution of complex functional transcriptomes capable of regulating various molecular, cellular, and developmental processes (Kalsotra and Cooper, 2011; Seo et al., 2013). In *Drosophila*, the *DSCAM* gene alternatively splices to generate more than ∼38,000 isoforms equivalent to ∼2.5 the number of genes in the fly (Schmucker et al., 2000). The biological functions of multiple isoforms are poorly explored; however, some studies provided evidence for the mechanistic regulatory role of AS. For example, the *Bcl-x* gene of the fruit fly generates two transcript isoforms coding for antagonistic proteins where one isoform activates apoptosis and the other inhibits it (Li et al., 2004). In humans, AS due to skipping exon 7 of the *SMN* (survival motor neuron) has been demonstrated to directly correlate with spinal muscular atrophy (Zhou et al., 2008). Conversely, the inclusion of exon 10 in tau transcript due to abnormal splicing has been implicated in tauopathies (Zhou et al., 2008). Clinical strategies are underway to target aberrant AS associated with human diseases (Zhou et al., 2008, 2012).

In addition to AS, recent RNA sequencing studies showed that alternative cleavage and polyadenylation contribute to transcriptome complexity and diversity in higher organisms (Wu et al., 2011; Sherstnev et al., 2012). Although RNA-Seq provides massive depth and understanding of the transcriptome, RNA-Seq protocols are behind in resolving transcript termini (Steijger et al., 2013). Therefore, other methods for sequencing 3′ and 5′ ends were adopted to retrieve requisite information. Cap analysis for gene expression (CAGE) sequencing has been used to annotate transcription start sites (TSSs) (Main et al., 2013; Boley et al., 2014), whereas deep 3′-sequencing (3′-seq) was used to define transcript termini and reveal unexpected alternative polyadenylation (APA) patterns (reviewed in Miura et al., 2014). In the human complex transcriptome, 54% of genes have multiple TSSs (Tyner et al., 2017). Precise promoter annotation will help to investigate the 5′ untranslated region (UTR) differential usage and the functional impact of genetic variation on gene expression. For instance, a regulatory single nucleotide polymorphism creates a new TSS causing thalassemia (De Gobbi et al., 2006). 3′ UTRs are the major mediators for posttranscriptional regulatory mechanisms, and therefore, gain or loss of regulatory elements such as microRNA binding sites, due to APA, can affect transcript stability and translational efficiency (reviewed in Miura et al., 2014). Although specialized methods in resolving transcript termini are available, none of the technologies mentioned above provides insights into the complete transcript structure.

The single-molecule real-time (SMRT) Iso-Seq of Pacific Biosciences (PacBio) allows a comprehensive analysis of the transcriptome. Unlike short-read RNA-Seq, Iso-Seq can capture full-length sequences, thereby improves gene annotation and accurately identifies transcript isoforms and gene fusions (Nudelman et al., 2018; Feng S. et al., 2019; Tian et al., 2019). Besides, long-read sequencing provides clear evidence for posttranscriptional processes such as APA and splicing events (Treutlein et al., 2014; Abdel-Ghany et al., 2016). Thus far, long-read sequencing has not widely been used in fish, with few reports in *Danio rerio* (Nudelman et al., 2018), *Lateolabrax maculatus* (Tian et al., 2019), *Misgurnus anguillicaudatus* (Yi et al., 2018), *Gymnocypris selincuoensis* (Feng X. et al., 2019), and *Salmo salar* (Ramberg et al., 2021). Conducting similar analyses in other species will contribute to understanding AS and the regulatory roles of APA and reveal the evolutionary conservation of splice isoforms (Abdel-Ghany et al., 2016).

In this study, PacBio long-read transcriptome sequencing was applied to improve the rainbow trout transcriptome annotation and yield a catalog of high-confidence transcript isoforms. We sequenced 14 tissues from three doubled haploid YY males from the Swanson River clonal line to achieve high coverage of transcript isoforms. In parallel, short-read RNA-Seq datasets were used to validate splice sites and AS events. The study findings revealed that intron retention (IR) is the most frequent AS event. The corrected PacBio transcriptome has been used to study the plasticity in exon usage in association with several physiological conditions of the fish. This study demonstrated the utility of PacBio Iso-Seq platform to characterize FL cDNA sequences and identify novel genes/isoforms, improving genome annotation and extending our knowledge/understanding of the rainbow trout transcriptome beyond the currently available resources.

RESULTS AND DISCUSSION

Iso-Seq Analysis Pipeline

Large-scale sequencing is essential for gene discovery and genome annotation; however, the sequencing depth, sequence

completeness, and cost are the main limitations of sequencing technologies (Wang et al., 2016). EST sequences and 454 pyrosequencing were previously used to assemble the trout transcriptome (Salem et al., 2010). Sanger sequencing is relatively expensive and generated sequences shorter than 1 kb. The 454 pyrosequencing produced ∼1.3 million reads (344 bp long on average) shorter than the EST sequences (Salem et al., 2010). Recently, Illumina short-read sequencing provided high sequencing depth, which assisted in refining the transcriptome (Berthelot et al., 2014; Salem et al., 2015; Pearse et al., 2020) and providing insights into transcriptional networks (Ali et al., 2018; Paneru et al., 2018) and gene structure (Berthelot et al., 2014; Pearse et al., 2020). However, short-read RNA-Seq breaks the continuity of the transcript and, therefore, fails to reconstruct the actually expressed transcripts and impairs our understanding of the functional aspects of isoform diversity (Steijger et al., 2013; Tilgner et al., 2014). More recently, PacBio Iso-Seq has been extensively used to identify FL transcripts and improve genome annotation (Abdel-Ghany et al., 2016; Wang et al., 2016; Nudelman et al., 2018; Feng S. et al., 2019; Feng X. et al., 2019; Tian et al., 2019). To characterize the rainbow trout transcriptome using Iso-Seq, RNA samples were isolated from 14 tissues in addition to a pooled RNA sample from fertilized eggs at different embryonic developmental stages. Tissues were collected from three doubled haploid fish to reduce heterozygosity but maintain tissue specificity. Twenty samples from two fish were barcoded and sequenced on four SMRT cells. To obtain a higher yield per tissue, 15 samples from one more fish were sequenced using SMRT cell per tissue. To reconstruct a high-confidence FL transcriptome, the ToFU pipeline (Isoseq3 v3.2.2) (Gordon et al., 2015) was used as illustrated in **Figure 1**. PacBio sequencing yielded a total of 6,776,786 reads of inserts (RoIs). Circular consensus sequencing (CCS) reads were generated and classified into 5,411,377 (79.9%) full-length non-chimeric (FLnc) reads of length ranges from 50 up to 25,831 bp (avg. = 2.3 kb). FLnc reads were defined as sequences having 5′ and 3′ barcoded primers and the poly(A) tail. Reads lacking any of these requirements were classified as non-full length (nFL) and were excluded from the analyses. In sea bass (*Lateolabrax maculatus*), 42.5% of the reads were classified as FL (Tian et al., 2019). The high percentage of the FLnc reads (79.9%) indicates high integrity of the trout RNAs used in the current Iso-Seq study.

The iterative clustering for error correction (ICE) algorithm was used in the Iso-Seq pipeline to obtain clusters of FL reads and then compute FL consensus isoform sequences (**Figure 1**). High-quality consensus sequences (452,955 FLnc) were mapped to the rainbow trout genome (NCBI Omyk_1.0) (Pearse et al., 2020) using the minimap2 alignment tool. A total of 451,178 reads (99.61%) were mapped to the reference genome, suggesting that the error rate of PacBio raw data, if any, was successfully corrected by the ICE as previously reported (Gordon et al., 2015). The percentage of unmapped reads (0.4%) was lower than that (3.6%) reported for zebrafish (Nudelman et al., 2018). The mapped reads were collapsed using the Cupcake tool, yielding 108,501 non-redundant isoforms (average length ∼2.8 kb) exhibiting alignment identity $\geq$ 0.95 and alignment coverage $\geq$ 0.99. To avoid truncated transcripts, incomplete retrotranscription reads differing only in the exonic structure of the 5′ ends were considered redundant, and only the longest isoform was retained. Although the high mapping percentage was achieved, we noticed small indels accumulated in 33.6% of the collapsed transcripts (avg. ∼1.6 indels/transcript) (**Supplementary Table 1**). Small indels were previously reported in 56.2% of the FL transcripts identified from a mouse neural differentiation PacBio dataset (Tardaguila et al., 2018). Previous efforts indicated that correction of indels with matching short reads decreased the number of transcripts harboring indels to 16% but was not satisfactory for open reading frame (ORF) prediction (Tardaguila et al., 2018). Thus, in our study, we used a reference-guided error correction; all collapsed isoforms were mapped back to the genome by SQANTI2, which returned a corrected PacBio reference transcriptome. In a previous study, a hybrid error correction approach using short reads and TAPIS (reference-guided error correction) yielded a 96% mapping percentage compared with 95% using TAPIS suggesting achievement of a high alignment rate without Illumina short reads (Abdel-Ghany et al., 2016).

Filtration and Characterization of the PacBio Isoforms

Tardaguila et al. (2018) recommended applying quality filters on the PacBio sequencing data to avoid potential technical artifacts due to reverse transcriptase (RT) template switching and off-priming. RT switching is enhanced by RNA secondary structures, which allow RTs to jump without terminating cDNA synthesis leading to gaps that could be interpreted as splicing events. Additionally, the oligo(dT) primer may anneal to non-poly(A) tail in A-rich regions of the template resulting in false cDNA molecules. To investigate the possible intrapriming, the percent of genomic "A's" in downstream 20 bp from the TTS were calculated. Thus, we adopted various approaches to remove potential technical artifacts from the PacBio transcriptome, including short-read support (Accession # PRJNA389609, PRJNA380337, PRJNA227065, and PRJNA259860), intrapriming, and RT-switching activities (**Supplementary Figure 1**). Overall, quality filters removed 31,641 transcripts (**Supplementary Figure 2**). The remaining transcriptome had 76,860 transcripts encoded by 24,729 (95.9%) known genes and 1,068 (4.1%) novel genes when compared with the RefSeq annotation reference (**Supplementary File 1**). In total, 65,670 ORFs of length $\geq$ 100 amino acids were predicted (**Supplementary File 2**). The predicted ORFs were mapped to the Swiss-Prot, TrEMBL, and Pfam protein domain databases (**Supplementary Table 2**). A total of 62,951 (96%) transcripts had homology with at least one database entity, whereas 49,690 (76%) transcripts had significant matches in the three databases (E-value $< 10^{-5}$). Among all collapsed isoforms in our data, there were 2,719 (∼4%) transcripts with predicted ORFs and no matches to any of the protein databases.

Notably, 11,190 transcripts (14.6%) had no ORFs greater than 100 amino acids long, suggesting that they are non-coding transcripts. To confirm the non-coding potential of those transcripts, they were searched against rainbow trout

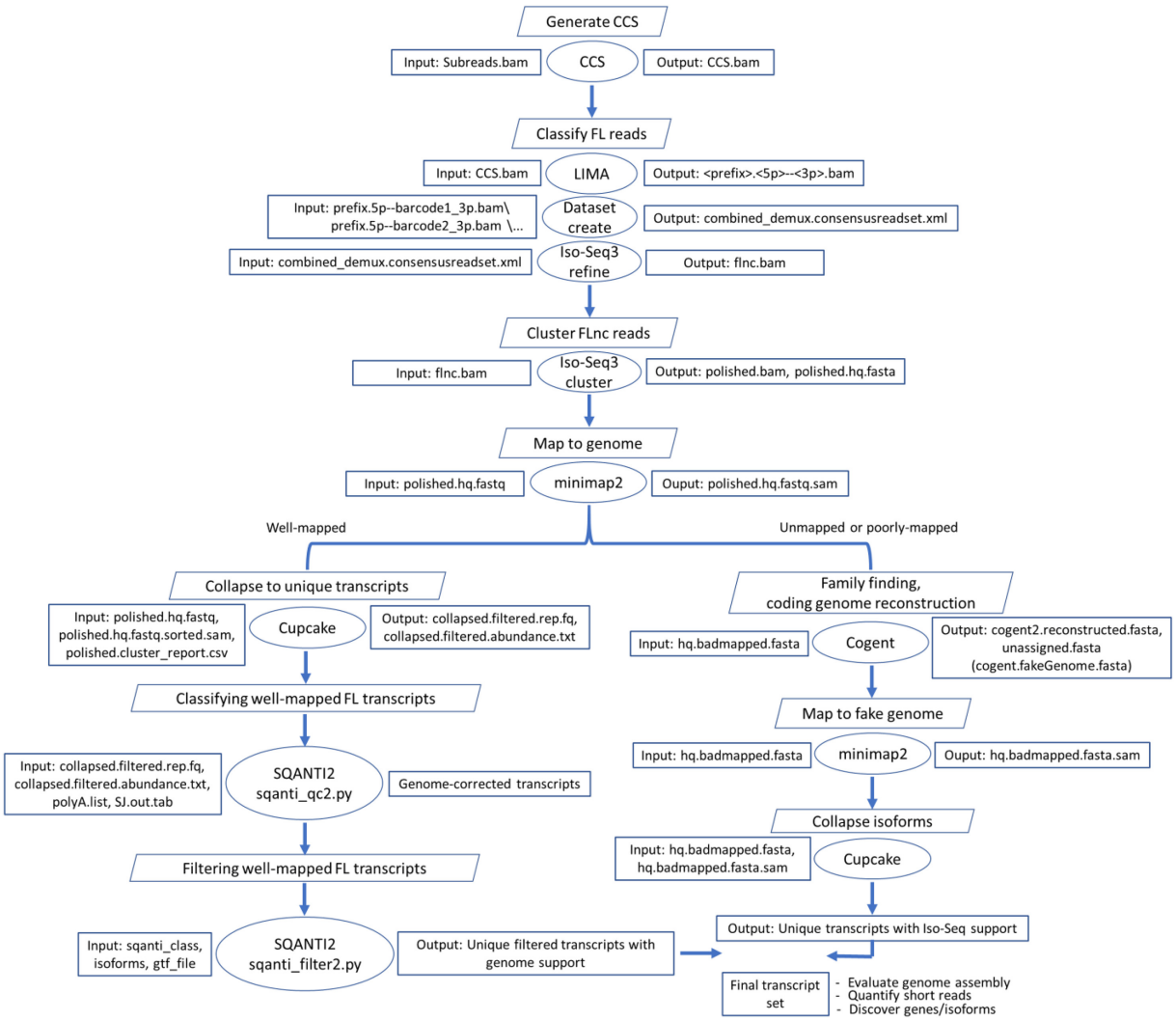


FIGURE 1 | Bioinformatics pipeline to reconstruct the rainbow trout transcriptome from the Iso-Seq dataset. CCS.bam file contains circular consensus sequence (CCS) reads, flnc.bam contains full-length non-chimeric (FLnc) reads, SJ.out.tab contains high confidence collapsed splice junctions (tab-delimited format), and polyA.list contains a list of polyA motifs to find upstream of the 3′ end site. "hq" stands for high quality.

pre-miRNAs (459 records) (Juanchich et al., 2016) and all the miRNA stem-loop sequences (38,589 sequences) available in the miRbase and Rfam (*E*-value ≤ 1e-10). A total of 489 transcripts exhibited homology with 92 miRNA precursors (**Supplementary Table 3**). There were 10,701 transcripts without homology with miRNA precursors and other non-coding RNA families in Rfam. Those transcripts were processed for lncRNA prediction as we previously described (Al-Tobasei et al., 2016). In total, 4,292 transcripts had a coding score ≤ 1 (**Supplementary Tables 4, 5**) and, therefore, were considered as lncRNAs. Interestingly, ∼59% of these lncRNAs were non-coding variants of protein-coding transcripts, missing 5′ exons and/or 3′ fragments than their coding transcript counterparts (**Supplementary Table 4**). Conversion of protein-coding RNA to non-coding RNA has been reported in some bifunctional coding genes, including activating signal cointegrator 1 complex subunit 3 (*ASCC3*) (Williamson et al., 2017), steroid receptor RNA activator 1 (*SRA1*) (Hube et al., 2006), and Protein Phosphatase 1 Nuclear Targeting Subunit (*PNUTS*) (Grelet et al., 2017). For instance, the ASCC3 mRNA switches to a shorter non-coding isoform due to alternative last exon splicing (Williamson et al., 2017). The short non-coding isoform has opposite effects on transcription recovery in response to UV-induced DNA damage (Williamson et al., 2017). *LncRNA–mRNA* hybrid genes need an in-depth investigation to unveil their biological regulatory mechanisms.

The final corrected transcripts were compared to the RefSeq genome annotation (Release 100; GCF_002163495.1) (**Figure 2** and **Supplementary Tables 1, 6**). There were 32,364 (42.1%) full splice match (FSM) isoforms that perfectly matched reference transcripts at all splice junctions (**Supplementary Table 1**). About 25.4% of the zebrafish long-read dataset showed an exact match to the RefSeq annotated transcripts

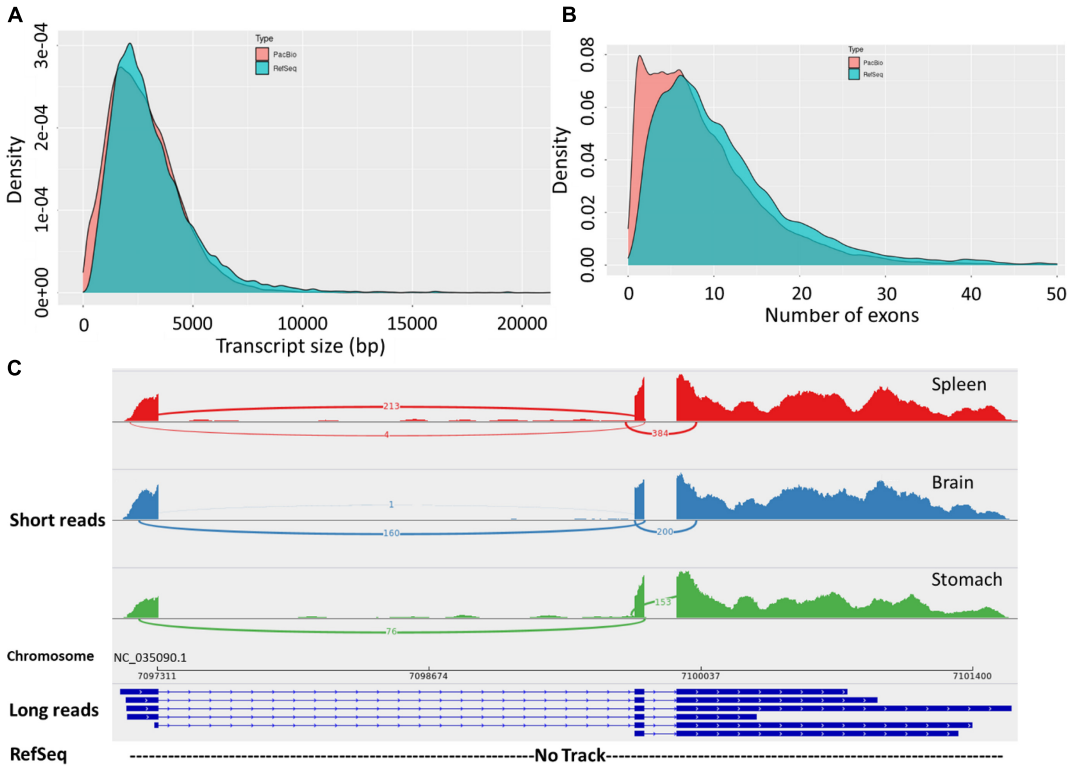


FIGURE 2 | (A) Length distribution of FL transcripts obtained from Iso-Seq data compared with RefSeq transcripts. **(B)** Distribution of the number of exons in the long-read sequences and RefSeq transcripts. **(C)** Sashimi plot showing an example of novel transcript isoforms on Omy14, detected by the PacBio Iso-Seq (long-reads track). The bottom (empty) track shows no corresponding annotation for these isoforms in the RefSeq reference. The top three tracks show short reads from three tissues precisely mapped to the exonic structures of the long-read track.

(Nudelman et al., 2018). This result suggests the presence of a significant fraction of undiscovered transcriptional diversity in the current RefSeq annotation. Also, 17.4% incomplete splice match (ISM) transcripts were identified as partially matching the reference genome. In zebrafish, 14.8% ISM transcripts lacking 5′– and 3′ end exons were identified (Nudelman et al., 2018). Furthermore, 31,125 (40.5%) novel isoforms were identified in this study. Remarkably, the PacBio isoforms had a fewer average number of exons (avg. 8.8 vs. 11.7 exons) and isoforms per gene compared with the RefSeq transcripts (avg. 3 vs. 4.7 isoforms). We also noticed that novel genes, compared with the RefSeq annotation, tend to have a single multi-exonic (avg. 3.6 exons) isoform per gene. The distribution of isoforms per gene and transcript lengths by structural categories are shown in **Supplementary Figures 3a–c**.

Notably, the PacBio isoforms (2.8 kb) are significantly shorter than the RefSeq transcripts (3.2 kb) on average (*t*-test, $p = 3.05 \times 10^{-243}$). The distances of 3′ end and 5′ end of FSM and ISM transcripts from the annotated polyadenylation and transcription start sites were calculated, respectively (**Figure 3**). Within 20 nt upstream of the annotated polyA site, only 41% of FSM transcripts had an exact or close to complete overlap with the 3′ end of matched reference transcripts (**Figure 3A**). In contrast, ∼15% of FSM transcripts showed a complete or close to complete overlap with the annotated 5′ end (**Figure 3B**). Additionally, it was obvious that more than 50% of 3′ ends of ISM sequences were falling short within 1 kb upstream of the reference annotated 3′ end (**Figure 3C**). Most of the ISM transcripts had short 5′ ends, particularly within 1 and 10 kb downstream of the reference 5′ end (**Figure 3D**). This result agrees with the notion of less control over the completeness of 5′ ends during cDNA library preparation. The imperfect matches between both ends of the PacBio FSM transcripts and reference transcripts may indicate APA/alternative TSS events (Tardaguila et al., 2018). Further investigation using specialized methods in resolving transcript termini is warranted.

Alternative Splicing and Polyadenylation Modes

Splice junctions were classified as canonical and non-canonical according to the dinucleotide pairs at the beginning and end of the encompassed intron (Tardaguila et al., 2018) (**Supplementary Table 6**). Junctions harboring GT-AG, GC-AG, and AT-AC were considered canonical, whereas other possible combinations were non-canonical splicing. Junctions in the reference were described as known junctions; otherwise, they were considered novel junctions. In total, 203,490 splice junctions from the collapsed isoforms were identified. Most of the identified splice junctions were from the known category (90.3%) (**Figure 4A**). Out of

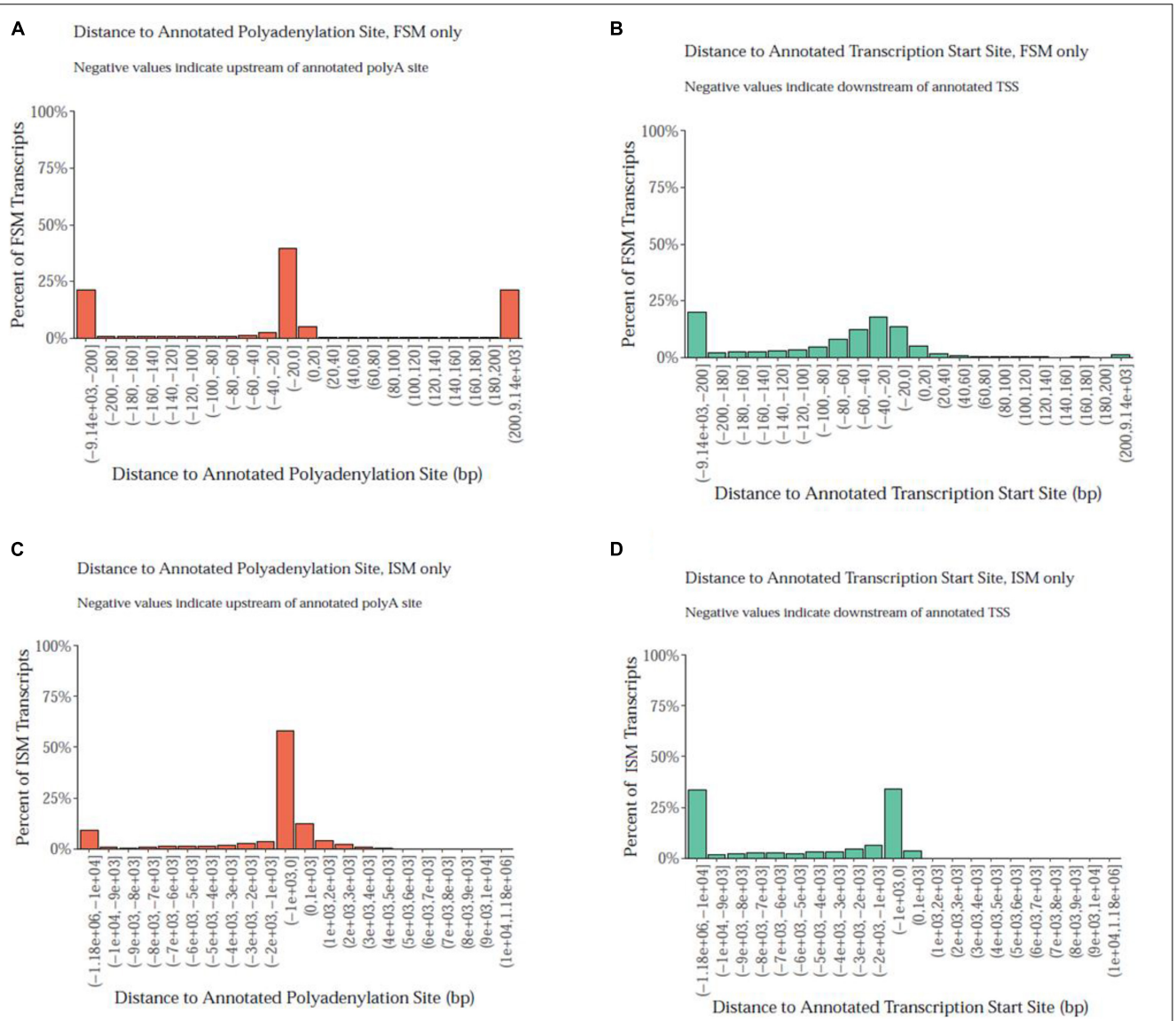


FIGURE 3 | Distance of PacBio transcript ends relative to the reference genome polyadenylation and transcript start sites. **(A)** At the 3′ end, ∼41% of FSM transcripts had an exact or close to complete overlap with the matched reference transcripts. **(B)** At the annotated 5′ end, ∼15% of FSM transcripts showed a complete or close to complete overlap with the matched reference transcripts. **(C)** More than 50% of 3′ ends of ISM sequences fell short within 1 kb upstream the reference annotated 3′ end. **(D)** More than 25% of ISM transcripts had short 5′ ends at 10 kb downstream of the reference 5′ end.

183,785 known splice junctions, 183,655 (99.9%) were canonical, and only 130 (∼0.1%) were non-canonical. In humans, canonical splice junctions were identified in more than 99.9% of all introns (Cocquet et al., 2006; Parada et al., 2014). Of note, novel junctions were found far from the TSS compared with known junctions (**Figure 4B**); ∼99.3% of the known canonical junctions were supported by short reads, whereas ∼57% of the novel canonical junctions were validated (**Supplementary Figure 4**). Notably, less than 1% of the novel non-canonical junctions were supported by short reads (**Supplementary Figure 4**). Following filtration, 96.7% of the remaining novel non-canonical junctions were supported with short reads. Splice junctions are described in detail in **Supplementary Figures 4, 5**.

Reconstruction of the rainbow trout transcriptome revealed that 20,431 loci (79.2%) are multi-exonic (avg. ∼9.6 exons). As shown in **Figure 4C**, AS events were extracted from the annotation file generated from the PacBio dataset (**Figure 4D**). A total of 33,383 AS events were identified from the PacBio dataset. IR was the most abundant AS event (34.15%) followed by exon skipping (ES) (12.34%) (**Figure 4D**). On the contrary, in the RefSeq annotation, ES was the most frequent event (24.44%), whereas IR was the least represented one (8.15%) (**Figure 4D**). Differences may be due to RefSeq annotation being combined from many tissues and different experimental conditions. A recent study reported 16–20% of IR of the genes in mouse and human cortex (Jeffries et al., 2020).

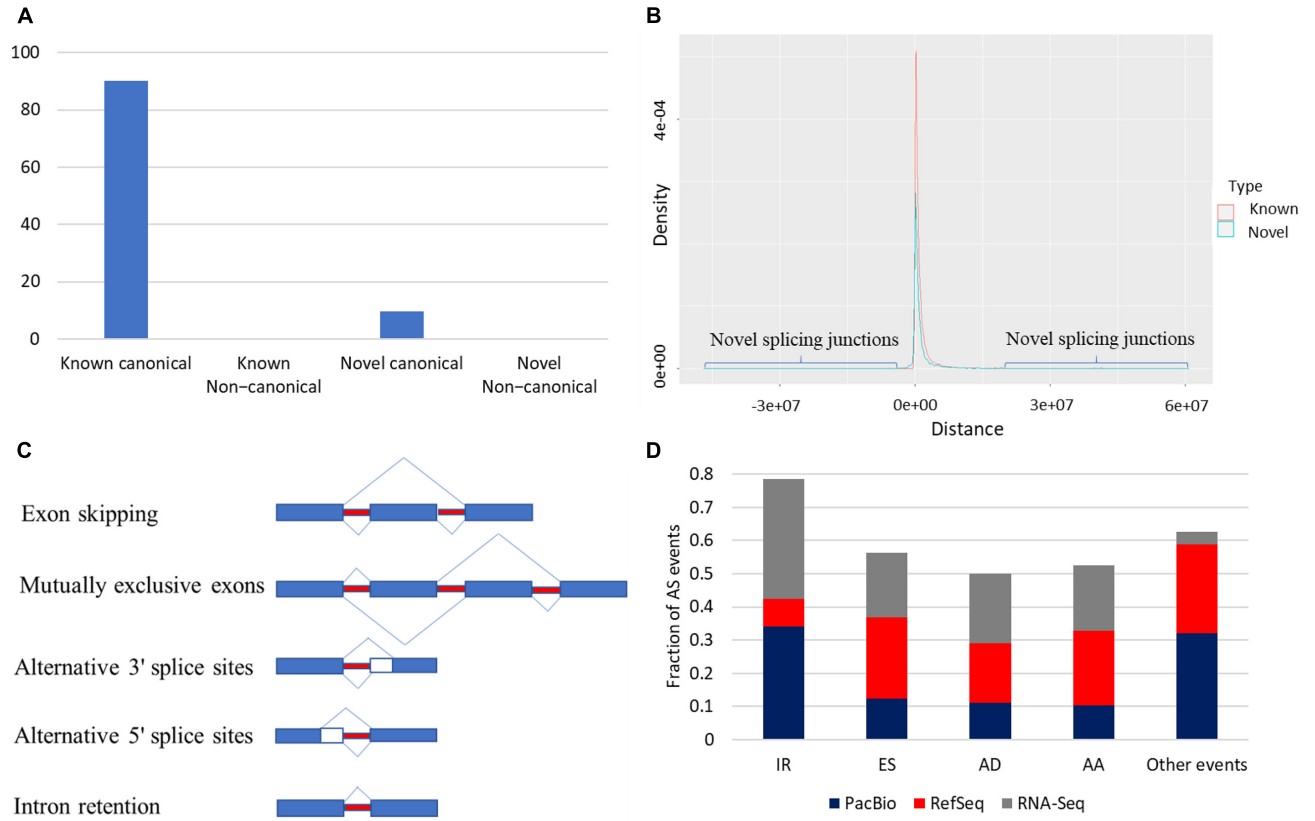
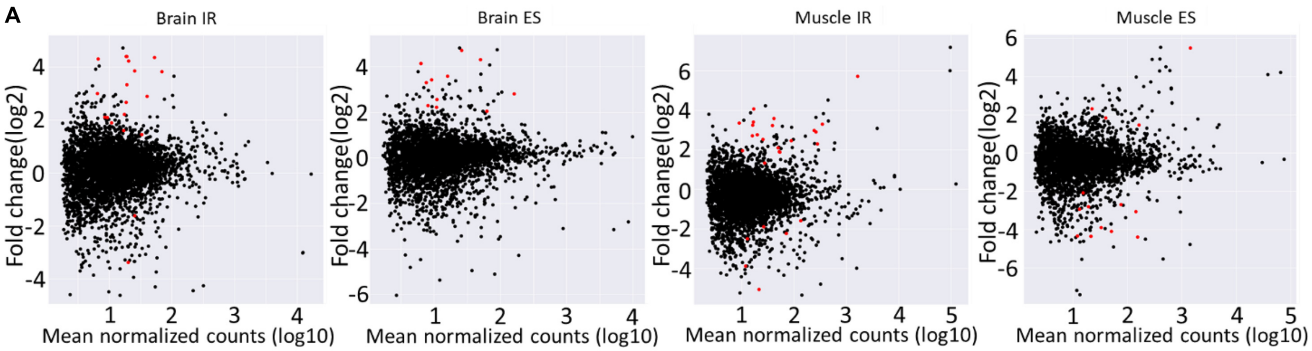


FIGURE 4 | (A) Most of the identified splice junctions were from the known canonical category (90.3%). **(B)** Novel splicing junctions tend to be far from the RefSeq transcription start site (TSS) compared with known junctions. **(C)** Types of AS. **(D)** Intron retention was the most frequent AS event in the corrected PacBio transcriptome and RNA-Seq data. In contrast, exon skipping (ES) predominated the RefSeq.

To validate the PacBio findings, the frequency of six types of AS (alternative 3′ splice sites; alternative 5′ splice sites; ES; multiple exon skips, ME; mutually exclusive exons, MX; and IR) was evaluated by RNA-Seq dataset generated from 13 tissues (Accession # PRJNA389609). In agreement with the PacBio data, IR was the most frequent event (36.3%), suggesting the reliability of the findings obtained from PacBio. IR and ES were reported as major AS forms in eukaryotes, with ES higher in animals and IR frequent in all eukaryotes, including plants (Grau-Bove et al., 2018). Our findings improved the transcriptome catalog for rainbow trout.

Furthermore, the availability of a PacBio-improved genome annotation facilitated the identification of differentially regulated AS patterns among tissues. Short-read datasets from nine tissues, collected from two Swanson fish (Berthelot et al., 2014; Salem et al., 2015), were mapped to the Swanson reference genome. A total of 156 differentially regulated events were identified (**Supplementary Tables 7, 8**). Of them, 33.3% were IR, whereas 21.8% were ES. The splicing event was considered as tissue specific when the event counted in a tissue was at least eight-fold higher than the other tissues (log2 FC $>$ 3; adj. p-value $<$ 0.05). A total of 66 splicing events in 44 unique genes were identified as tissue specific (**Supplementary Tables 7, 8**). Of them, 39.4% were IR, whereas 21.2% were ES. Brain and white muscle had 89% of the tissue-specific splicing events (**Supplementary Table 7** and **Figure 5A**). Similar to our findings, Rodriguez et al. (2020) reported a high abundance of tissue-specific alternative forms in nervous and muscle tissues. A few tissue-specific splicing events were identified from liver, head kidney, and stomach. It is worth mentioning that no differentially regulated/tissue-specific events were identified in spleen, kidney, intestine, and gill when independently compared with other tissues. The top tissue-specific AS patterns in muscle were identified in genes encoding cold shock domain-containing protein E1 (*CSDE1*) and phosphate carrier protein, mitochondrial (*SLC25A3*) (**Figure 5B**). *CSDE1* is critical for the efficient formation of stress granules (Youn et al., 2018). *SLC25A3* transports inorganic phosphate (Pi) across the mitochondrial membranes, which is necessary for the final step of oxidative phosphorylation. Pathologic variants of the *SLC25A3* have been reported in association with skeletal myopathy phenotype in humans (Mayr et al., 2011; Bhoj et al., 2015). In comparison, the top tissue-specific AS forms in the brain were identified in genes coding for protein tweety homolog 1 (*Ttyh1*) and ras-related protein Rab-6A (*Rab6a*) (**Figure 5B**). In mammals, the expression of the *Ttyh1* gene is mainly restricted to nervous tissue, where it revealed a role in cell adhesion and as a transmembrane receptor (Matthews et al., 2007). *Rab6a*



B

event_id	gene	chrom	strand	p-value adj	MEC_A	MEC_B	log2FC_EC	condition A	condition B	annotation
alt 3prime	PB.11858	NC_035085.1	-	1.98E-03	8464.13	4.28	10.95	Muscle	Other tissues	cold shock domain-containing protein E1
alt 5prime	PB.11858	NC_035085.1	-	7.74E-03	7316.22	4.16	10.78	Muscle	Other tissues	cold shock domain-containing protein E1
IR	PB.2665	NC_035078.1	+	2.18E-04	3296.25	13.04	7.98	Muscle	Other tissues	phosphate carrier protein, mitochondrial
ES	PB.2665	NC_035078.1	+	6.54E-04	2869.28	13.30	7.75	Muscle	Other tissues	phosphate carrier protein, mitochondrial
alt 3prime	PB.1357	NC_035078.1	+	3.70E-05	241.49	1.99	6.92	Brain	Other tissues	protein tweety homolog 1
alt 5prime	PB.1357	NC_035078.1	+	5.43E-03	164.08	2.11	6.28	Brain	Other tissues	protein tweety homolog 1
IR	PB.28488	NC_035101.1	+	1.97E-02	101.74	1.50	6.08	Brain	Other tissues	ras-related protein Rab-6A
ES	PB.28488	NC_035101.1	+	2.27E-02	96.24	1.47	6.03	Brain	Other tissues	ras-related protein Rab-6A
IR	PB.28488	NC_035101.1	+	5.49E-03	132.99	2.82	5.56	Brain	Other tissues	ras-related protein Rab-6A

MEC_A: mean event count_A; MEC_B: mean event count_B; log2FC_EC: log2FC_event count

FIGURE 5 | (A) MA plots showing major regulated AS forms (IR and ES) in brain and white muscle compared with eight other tissues. The red dots represent the differentially regulated AS forms at adjusted $p \leq 0.05$. **(B)** Top tissue-specific AS patterns in brain and white muscle.

knockdown led to defects of the cytoskeletal structures in mice (Ma et al., 2016). Tissue-specific alternative forms were previously identified in genes related to cytoskeleton, cell-cell adherens junction, focal adhesion, and structure of muscle fibers (Rodriguez et al., 2020). Further investigation is needed to study the role of tissue-specific AS forms in muscle and brain development.

PacBio sequencing generates FL transcripts containing poly(A) tails, which help to detect APA sites accurately. We searched for possible motifs within 50 nt upstream of the polyadenylation sites. We detected 14 poly(A) signals located within ∼18 nt upstream of the polyadenylation cleavage site (**Supplementary Figure 6**). The AATAAA (60.6%) and ATTAAA (19.8%) were the most frequent motifs in the PacBio transcriptome (**Supplementary Figure 6**), suggesting that these motifs are essential for polyadenylation. AATAAA is a well-known conserved poly(A) signal in plants (Feng S. et al., 2019) and animals (Proudfoot and Brownlee, 1976).

Reconstruction of Coding Regions From the Unmapped/Poorly Mapped Reads

There were 103,193 reads that were unmapped or poorly mapped to the genome and filtered out due to low alignment identity and coverage. The COding GENome reconstruction Tool (Cogent) was used to reconstruct coding regions from the unmapped and poorly mapped reads, generating 10,057 gene families and 8,636 unassigned sequences (Tseng, 2020). All coding bases in isoforms transcribed from a single locus were combined, yielding a reconstructed contig representing each gene family. All reconstructed sequences ($n = 30,445$) (**Supplementary Files 3, 4**) were employed as a reference to realign the unmapped/poorly mapped reads and make them suitable to be processed through the ToFU pipeline, which filters out reads exhibiting identity less than 0.95 and alignment coverage below 0.99 to each gene family locus. Afterward, redundant isoforms were successfully collapsed into 60,926 FL isoforms (avg. length = 2.9 kb), harboring 41,414 ORFs (**Supplementary File 5** and **Supplementary Figure 7**). Collapsed isoforms were annotated as shown in **Supplementary Tables 9–11**. Remarkably, when all collapsed reconstructed transcripts were mapped to the Swanson genome sequence, only 388 (0.64%) transcripts were mappable at identity ≥ 0.95 and coverage ≥ 0.99. In contrast, when the 60,926 reads were mapped to the newly released genome sequence of rainbow trout Arlee strain [USDA OmykA_1.1 assembly (GCF_013265735.2)], 35,218 (∼58%) transcripts were mappable, suggesting the reliability of the Cogent reconstructing the coding sequences and perhaps a necessity to improve the current version of the Swanson strain genome reference of this study. It is worth mentioning that the contiguity of the Arlee genome assembly has recently been improved using long reads. Furthermore, the Bionano optical mapping and Hi-C proximity ligation sequence data were used to join the Arlee contigs into scaffolds, which were then anchored to and ordered on chromosomes using genetic linkage information. The Swanson genome assembly has 139,799 unplaced scaffolds compared with 939 scaffolds in the Arlee assembly (Gao et al., 2021).

rnaQUAST 1.2.0[1] was used to further assess the PacBio transcriptome quality compared with Swanson RefSeq (Bushmanova et al., 2016). The collapsed isoforms were mapped to the Swanson trout reference genome using GMAP and BLAT to match the alignments to the reference coordinates (**Supplementary Tables 12, 13**). Based on the common alignment output, a total of 1,479 collapsed transcripts showed no significant alignment with the Swanson trout genome; of these, 346 transcripts (23%) had significant hits with the Arlee strain (identity ≥ 0.95 and coverage ≥ 0.99), suggesting those transcripts are missing in the Swanson RefSeq annotation; 8,348

[1]https://de.cyverse.org/

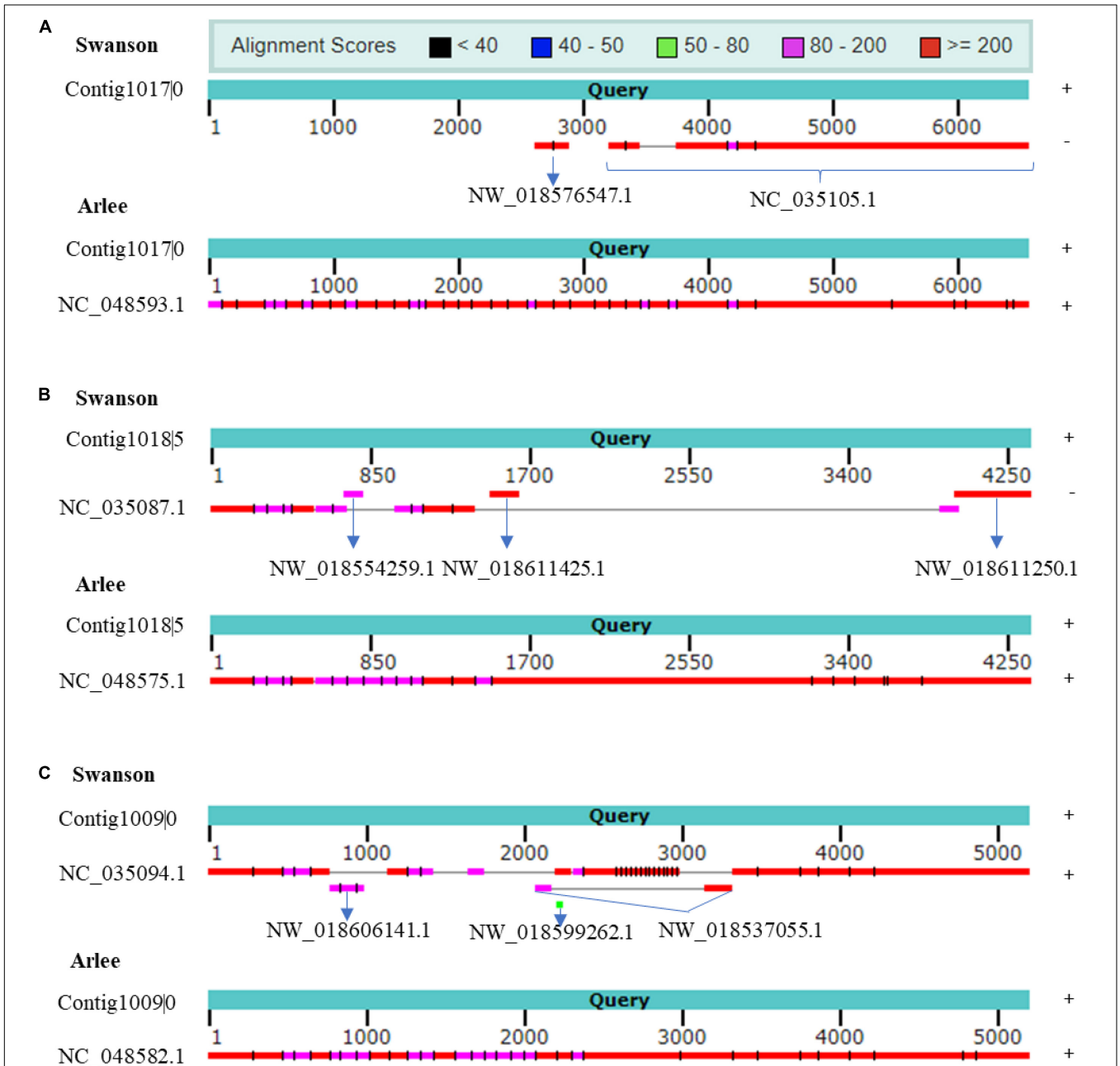


FIGURE 6 | (A) Cogent contig shows chromosome NC_035105.1 (Omy29) missing ∼3.2 kb. Cogent contig shows that the genomic scaffold NW_018576547.1 is placed on Omy29. The reconstructed contig aligns to chromosome Y on Arlee assembly, which provides evidence for the accuracy of Cogent reconstruction. **(B)** Cogent contig shows chromosome NC_035087.1 (Omy11) missing ∼2.5 kb. Cogent reconstructed one contig to which three scaffolds were mapped in this order, namely, "NW_018554259.1," "NW_018611425.1," and "NW_018611250.1" and validated by Arlee assembly. **(C)** The reconstruction reveals several discrete misassemblies on Omy18 (NC_035094.1) and anchors three scaffolds to the chromosome. Arlee assembly was used to validate the reconstruction. Strand orientations are provided on the right side of the figure. Cogent contigs and Arlee loci have the same strand orientations.

unannotated transcripts did not match any reference transcripts. The mapping revealed 15,120 misassembled transcripts (mapped to a different chromosome, strand, reverse order, etc.). To prove that the misassembled transcripts are not due to a high error rate in the PacBio sequencing, we mapped the ∼15K misassembled transcripts to the Arlee and Atlantic salmon genome sequences. A total of 9,935 (∼66%) and 1,209 (∼8%) transcripts matched the Arlee and Atlantic salmon genomes at identity ≥ 0.95 and coverage ≥ 0.99. For instance, Iso-Seq identified seven isoforms; Cogent resolved it to one contig. Mapping the contig back to NC_035105.1 (Omy29) showed a misassembly where the strand orientation is opposite, and Omy29 is missing the first ∼3.2 kb of the contig (**Figure 6A**). The Arlee assembly provided evidence for the presence of the 3.2 kb on the Y chromosome (**Figure 6A**) and, in turn, the accuracy of our reconstructed contig. Similarly, the reconstruction yielded a contig mapped to Omy11, which lacks ∼2.5 kb (**Figure 6B**). The contig also maps to three unplaced genomic scaffolds: "NW_018554259.1," "NW_018611425.1," and "NW_018611250.1" (**Figure 6B**). Also, Iso-Seq identified six transcripts that Cogent reconstructs into a single contig. The reconstructed Cogent contig, mapped to Omy18 and three scaffolds, showed that scaffold order should be "NW_018606141.1" followed by "NW_018537055.1" and "NW_018599262.1" (**Figure 6C**). Overall, Arlee assembly provided evidence for the accuracy of contig reconstruction, suggesting the necessity to refine the gene models in the current Swanson genome assembly.

The completeness of the PacBio transcriptome was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) (Seppey et al., 2019). BUSCO v5.1.2 checked for single and duplicate orthologs for members of the Actinopterygii lineage. A total of 3,640 BUSCO groups were searched to assess the transcriptome completion. Overall, 89% annotation completeness was achieved. BUSCO alignments revealed 8,679 well-mapped FL transcripts and 2,564 FL unmapped/poorly mapped collapsed transcripts that have hits to 2,662 and 1,185 orthologs, respectively. Remarkably, the unmapped and poorly mapped transcripts had hits to 564 orthologs with no matches in the well-mapped transcripts. Our results showed that the characterization of the rainbow trout transcriptome is close to complete and that sequencing more tissues from different biological conditions may help identify more FL transcripts to complete the genome annotation.

Alternative Splicing, Polyadenylation, and TSS Associated With Economically Important Phenotypes

AS and APA are interesting complexity aspects of the eukaryote transcriptome. The mechanism of AS and APA generates more transcripts per gene locus and, thus, expands the proteome diversity. Previous studies showed that the posttranscriptional mechanisms play important roles in immune responses (Martinez and Lynch, 2013), muscular atrophy (Lorson et al., 1999), cancer (Hube et al., 2006), and neurological disorders (Zhou et al., 2008). Therefore, we used DEXSeq to profile differential exon usage (DEU) in rainbow trout across different biological conditions using publicly available data (see "Materials and Methods" section) to identify AS and APA associated with the studied phenotypes. Change in relative exon usage could be due to (1) a change in the rate of exon splicing (i.e., AS), (2) a change in usage of alternative TSS, or (3) a change in usage of APA sites.

Fish Growth and Muscle Accretion

To identify AS and APA events contributing to fish growth and muscle accretion, RNA-Seq data previously generated from fish families exhibiting extreme whole-body weight (WBW) and muscle yield phenotypes (Ali et al., 2018) were mapped to the rainbow trout genome using TopHat2. DEU analysis revealed two exons differentially spliced in fish families showing divergent WBW phenotypes (**Supplementary Table 14**). The spliced transcripts are coding for the negative elongation factor C/D (*NELFCD*) and titin genes. The differentially used exon (DUE) in *NELFCD* (NC_035093.1 :41691708-41692396) was upregulated in fish families with low WBW (**Supplementary Figure 8**). Knockdown of *NELFCD* suppressed cancer cell proliferation *in vitro* (Song et al., 2018). Conversely, the DUE in titin (exon9) was upregulated in fish families with high WBW. Titin guides the assembly of myofibrils from premyofibrils. In zebrafish, knockout of titin from two titin homologs developed exon-dependent phenotypes of variable severity, including susceptibility to biomechanical stresses and degeneration during development explained by the exon usage hypothesis (Shih et al., 2016). Additionally, a single exon (NC_035087.1:57851656-57851747) was significantly DU and upregulated in fish families showing high muscle yield (**Supplementary Figure 8** and **Supplementary Table 14**). This exon is in a novel isoform coding for THO complex subunit 5 homolog (*THOC5*). *THOC5* is an essential element for normal proliferation and differentiation processes (reviewed in Tran et al., 2014). Depletion of *THOC5* in the embryonic fibroblasts inhibited cell growth (Guria et al., 2011). It is noteworthy that all identified DUEs were in the perfectly mapped transcripts.

To identify AS and APA events involved in muscle atrophy associated with sexual maturation, RNA-Seq data previously generated from gravid and sterile rainbow trout were used (Paneru et al., 2018). A total of 747 DUEs (adj. p-value < 0.05) were identified (**Supplementary Table 14**). The eukaryotic translation initiation factor 4E binding protein 2 (*EIF4EBP2*) had the most significant DUE (exon3; log2 FC = 4.4; adj. p-value = 4.44E-47) in the sterile fish relative to gravid fish. *EIF4EBP2* is known to inhibit protein synthesis, and the mTOR signaling pathway inactivates it to stimulate cell growth and metabolic process (Ding et al., 2018). Since this exon is highly used in sterile fish, we speculate that this exon is likely inactivating the *EIF4EBP2*. Conversely, mucosa-associated lymphoid tissue lymphoma translocation protein 1-like (*MALT1*) and four other genes had exons totally absent in the sterile fish. *MALT1* is a signaling component with protease functions (Coornaert et al., 2008). A total of 258 exons in the reconstructed poorly mapped/unmapped transcripts were DU (**Supplementary Table 15**). Of them, MHC class I heavy chain (PB.5976) (**Figure 7A**) and protein-tyrosine kinase 2-beta (PB.17301) were at the top of the list. Gene enrichment analysis revealed that

the isoforms harboring DUE are significantly enriched in the ribosome KEGG pathway and have GO terms belonging to translation (**Supplementary Table 14**).

Disease Resistance

Flavobacterium psychrophilum, the causative agent of BCWD, causes worldwide economic losses to the aquaculture industry (Nematollahi et al., 2003). Resistance to BCWD was demonstrated to be a moderately heritable trait that responds to selection (Silverstein et al., 2009; Leeds et al., 2010). Selective breeding programs have the potential to improve heritable phenotypes through existing genetic variation among individual animals or families (Ali et al., 2020). To gain insights into the molecular mechanism associated with resistance BCWD, RNA-Seq datasets previously generated from two genetic lines exhibiting contrasting resistance to BCWD (Marancik et al., 2014) were used for exon usage analysis (**Supplementary Tables 14, 15**). On day 1 post-infection, 77 exons were DU in the resistant and susceptible genetic lines. Of them, the first exon in a gene encoding interferon-induced very large GTPase 1 (*GVIN1*) was upregulated in the resistant line (log2 FC = 18.0). *GVIN1* was differentially expressed among survivors of three carp clones following herpesvirus (*Ca*HV) challenge (Lu et al., 2019). DEU analysis of all transcriptomic datasets from resistant and susceptible genetic lines showed that, regardless of the infectious status and days of infection, the *GVIN1* exon (**Figure 7B**) was completely absent in the susceptible line (log2 FC = 20.9; **Supplementary Table 14**). To further validate these data, the *GVIN1* exon was amplified by qPCR; the exon expression level in the resistant line exceeded that from the susceptible line by about 25-fold (**Supplementary Figure 8**). Transcript abundance analysis revealed that one of the three transcripts harboring the *GVIN1* exon was the most upregulated transcript in the resistant line (log2 FC = 2.5) compared with the susceptible line (**Supplementary Table 17**). The *GVIN1* DUE encodes 670 amino acids representing 36% of the whole ORF (1,836 amino acids long). These results suggest a role for *GVIN1* in disease resistance to BCWD.

Among all datasets from resistant and susceptible genetic lines (eight RNA-Seq dataset/genetic line), 238 DUEs were identified in the reconstructed unmapped/poorly mapped contigs (**Supplementary Table 15**). For instance, exon 3 in three isoforms coding for dystrophin was completely absent in resistant fish and showed log2 FC = −13.6 when compared with fish from the susceptible line (**Supplementary Figure 8**). In addition, an exon in a non-coding transcript (**Figure 7C**) and an exon in a transcript encoding microfibril-associated glycoprotein 4 were alternatively polyadenylated in fish showing divergent resistance to BCWD. In *Oreochromis niloticus*, a microfibril-associated glycoprotein 4 has demonstrated agglutination and opsonization capability to bacteria (Wu et al., 2020). A small/partial IR event was detected in a transcript coding for pentraxin-related protein PTX3 (**Figure 8**). The latter is a mediator of innate resistance to bacterial pathogens (Doni et al., 2019).

Our results indicate substantial genetic variation among fish from the resistant and susceptible lines explained by DUEs.

Fish Migration

To identify the molecular mechanism driving fish migration, we sought to compare the brain transcriptome at the time of smoltification (i.e., the physiological transition into seagoing forms) to an early time point during the second year of development in both male and female anadromous fish. For this purpose, an RNA-Seq dataset was obtained from Hale et al. (2016). A total of 533 and 349 DUEs belonging to the well-mapped and reconstructed reads were identified in anadromous female fish during smoltification (24 months) compared with 20-month-old presmolt fish (**Supplementary Tables 14, 15**). The DUEs were not identified in the resident rainbow trout population simultaneously (**Supplementary Table 14**), suggesting a potential role in fish migration. The list of DUE included downregulation of the first exon of an isoform encoding glycogen phosphorylase B (log2 FC = −16.9; **Supplementary Tables 14, 15**). During migration, muscle proteins in salmon act as a fuel and a carbon skeleton source to maintain the hepatic glycogen levels (Halver and Hardy, 2002). Hepatic glycogen content in Atlantic salmon after seawater entry became much lower (Plisetskaya et al., 1994). This may explain the complete absence of the glycogen phosphorylase exon at the 24-month-old fish. Also, the upregulation of two exons in a gene coding for E3 ubiquitin-protein ligase RNF115, which is involved in protein ubiquitination, may provide evidence of the role of proteins as a fuel source during migration. Our analysis revealed enrichment of the DUE-harboring isoforms in carbohydrate metabolism and ribosome KEGG pathways. Up on smoltification (24 months), the usage of two small exons belonging to the gamma-aminobutyric acid receptor-associated protein (*GABARAP*) and serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform (*PPP2CA*) dramatically decreased. *GABARAP* has a role in increasing the activity of the major inhibitory neurotransmitter (GABA), which is associated with behavioral traits in mice and Atlantic salmon (Thornqvist et al., 2015). Hypomethylated cytosines associated with *PPP2CA* were previously identified in 20-month-old fish relative to 8-month-old fish (Gavery et al., 2019). We noticed exceptional upregulation of two exons in novel isoforms expressed from a gene encoding Ig mu chain C region membrane-bound form (**Figure 7D**). Changes in immune response in migrating salmon were previously reported not to be due to infection but rather to the life history of salmon (Dolan et al., 2016). Upregulated exon was identified in a transcript encoding unconventional myosin-VIIa, which is required for sensory perception of the light stimulus (Ahmed et al., 2001) and sound (Weil et al., 1995). Migratory salmon rely on the sensory system (Farrell, 2011). Several other DUEs were identified in transcripts encoding proteins with a role in maintaining the nervous system such as Aladin (Tullio-Pelet et al., 2000); Na/K ATPase alpha subunit isoform 1b (Peng et al., 1997; Edwards et al., 2013); glycerophosphodiester phosphodiesterase 1 (Yanaka, 2007); protein-arginine deiminase type-2 (Asaga and Ishigami, 2007); lysyl oxidase homolog 2A (Du and Zhu, 2018); glutamate receptor ionotropic, kainate 2 (*GRIK2*) (Martin et al., 2007); and ST8 alpha-*N*-acetyl-neuraminide alpha-2,8-sialyltransferase 1 (Husain et al., 2008).

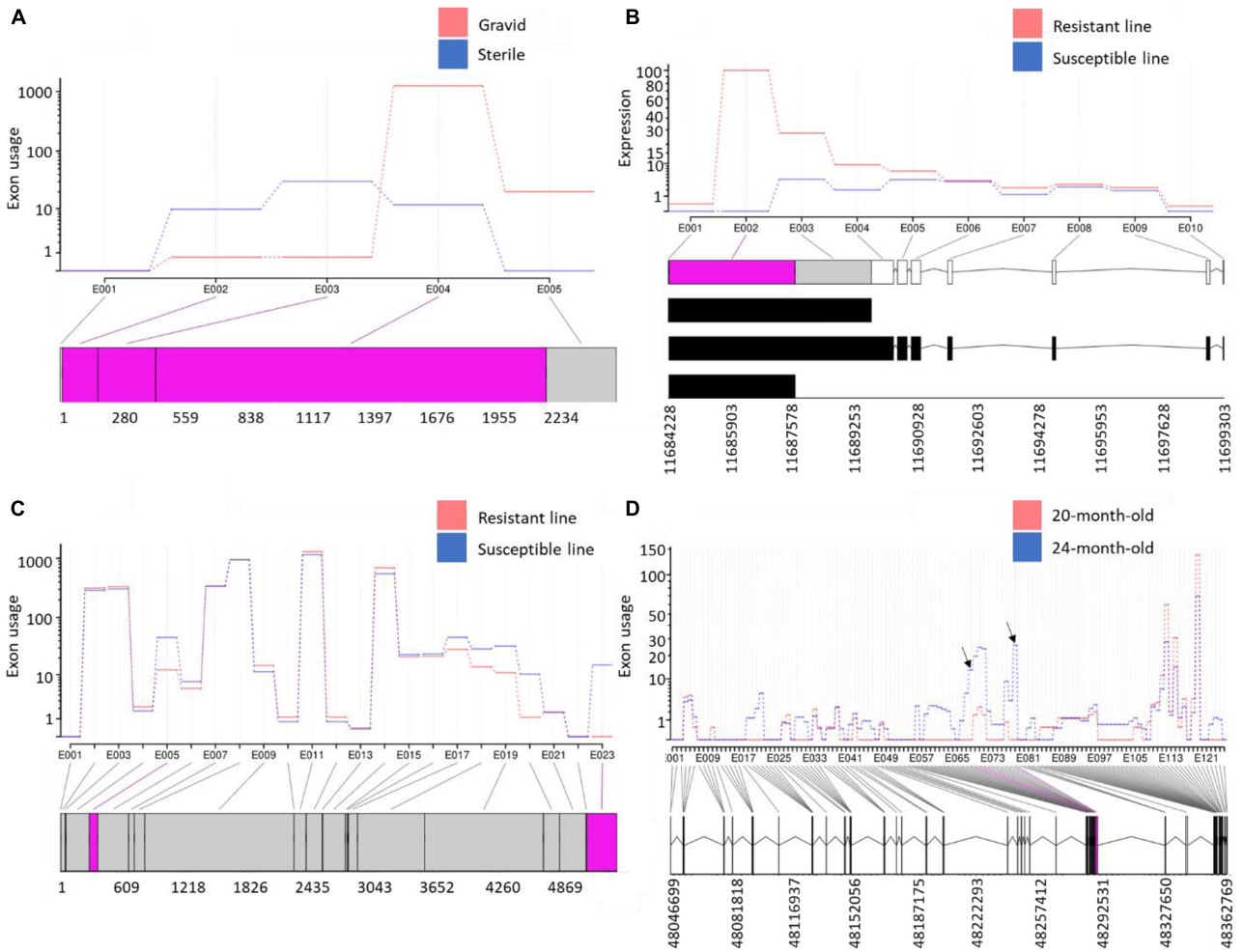


FIGURE 7 | (A) Variable exonic features of isoforms transcribed from a gene encoding MHC class I heavy chain showed opposite usage patterns in atrophying muscle during sexual maturation; **(B)** the first exon in a gene encoding *GVIN1* revealed exceptional downregulation in fish from the susceptible genetic line for BCWD; **(C)** a non-coding transcript was alternatively polyadenylated in fish showing divergent resistance to BCWD; and **(D)** exceptional upregulation of two exons in novel isoforms, encoding Ig mu chain C region membrane-bound form, in anadromous female smolts (24-month-old) compared with presmolts (20-month-old). Differentially used exonic features are shown in pink at the bottom of each panel.

A total of 1,163 and 164 DUEs in the well-mapped and reconstructed reads, respectively, were identified among 20– and 24-month-old anadromous males (**Supplementary Tables 14, 15**). Functional annotation analysis showed that the DUE-harboring transcripts are enriched in brain development, axon extension, response to activity, ATP binding, and tricarboxylic acid cycle. Remarkably, only 10 DUEs were common between male and female smolts. The list included protocadherin alpha-C2 (*PCDHAC2*) and *GRIK2*. *PCDHAC2* is involved in establishing and maintaining complex networks of neuronal connections in the brain (Wu and Maniatis, 1999), and *GRIK2* has GO terms related to the detection of cold stimulus involved in thermoception. Atlantic salmon smolts start their migration at a water temperature of 5°C and reach a peak of migration at water temperature > 8°C (Whalen et al., 1999). The current study showed sex-biased exon usage and suggests a role for AS in regulating the developmental plasticity in anadromous fish toward smoltification.

Response to Stress

Under intensive rearing conditions, fish experience diverse stressors, which negatively affect fish health, growth, and filet quality. Understanding the molecular mechanisms underlying stress responses will help to develop strategies that target improving animal welfare and aquaculture industry profitability. Therefore, we investigated DUEs in rainbow trout fish under five different stress conditions. For this purpose, RNA-Seq datasets were downloaded from the NCBI SRA (PRJNA312486). A total of 665, 37, 286, 554, and 124 DUEs were identified in fish exposed to high salinity, high temperature, low temperature, reused water, and crowding, respectively (**Supplementary Table 14**). Under all five stress conditions, there was a single common DUE (NC_035086.1:9191793-9192329) belonging to

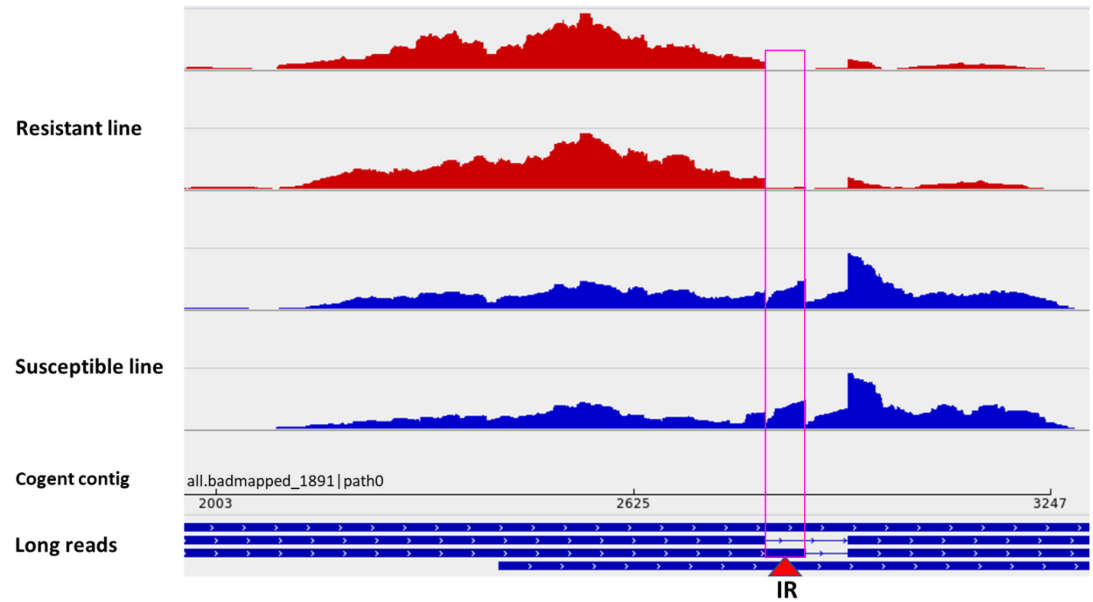


FIGURE 8 | A small/partial intron retention event in a transcript isoform coding for pentraxin-related protein PTX3 (red triangle). Two short-read datasets from a BCWD-resistant genetic line (red panels) showing no intron retention and two sets from a susceptible genetic line showing intron retention (blue panels).

transcript isoforms encoding prolyl 4-hydroxylase subunit alpha-2 (*P4HA2*) (**Figure 9**). Prolyl hydroxylation is a posttranslational modification to modulate protein folding and stability (Xiong et al., 2018). Prolyl 4-hydroxylase requires ascorbate to catalyze hydroxylation of proline residues in newly synthesized collagen chains to form 4-hydroxyproline. The hydroxylated residues stabilize the collagen triple helices under different physiological conditions (Pihlajaniemi et al., 1991). For instance, it was reported that stressed animals have a low concentration of ascorbic acid, which is not sufficient for collagen hydroxylation. This abnormal collagen affects the basement membrane structure of epithelial layers, causing skin lesions and blood vessel fragility (Pandey, 2007).

Three DUEs were identified in response to at least four stressors. Of them, delta-1-pyrroline-5-carboxylate synthase (*P5CSA*) and heat shock protein HSP 90-alpha (*HSP90α*) had an upregulated DUE in response to high salinity, low temperature, reused water, and crowding. In plants, *P5CSA* is induced in response to salt stress (Yoshiba et al., 1995) and water deprivation (Sharma et al., 2011). Pollutant-exposed fish hepatocytes induced *HSP90α*, which enabled the hepatocytes to become tolerant to oxidative stress (Padmini and Usha Rani, 2011). Conversely, two novel antisense transcripts had a downregulated DUE in response to high temperature, low temperature, reused water, and crowding.

A total of 24 DUEs were identified in response to at least three stressors. For instance, six exons were DU in response to high salinity, reused water, and crowding such as angiomotin, adenylate cyclase type 2, and lysine-specific histone demethylase 1A. Salt adaptation in teleost fish modulates the adenylate cyclase activity (Guibbolini and Lahlou, 1987). Acute stress in mouse models was regulated by the lysine-specific demethylase 1 (Longaretti et al., 2020). Fish differentially used three common exons when they were subjected to extreme temperatures. Two of them were belonging to transcripts encoding myozenin-2 and *P4HA2*, and a single DUE belonged to two non-coding antisense transcript isoforms. Transcripts undergoing posttranscriptional events in response to stress are enriched/involved in glycolysis and protein processing in the endoplasmic reticulum (**Supplementary Table 16**). Stress levels are often assessed according to plasma glucose and lactate levels (Arends et al., 1999; Acerete et al., 2004). Endoplasmic reticulum stress in the hepatopancreas of white shrimp was reported in response to low temperature (Fan et al., 2016). Our findings provide a basis for further investigation of molecular response to stress in rainbow trout, leading to better breeding practices to improve aquaculture production efficiency.

CONCLUSION

Iso-Seq data were used to construct a high-confidence FL transcriptome for rainbow trout. The study identified ∼76K FL transcripts that are well-mapped to the current Swanson reference genome and contain ∼65K ORFs longer than 100 amino acids. We identified 1,068 (4.1%) novel gene loci not previously annotated in the RefSeq reference. Additionally, ∼60K FL isoforms that were either poorly mapped or unmapped (∼1.4K transcripts) to the current genome were reconstructed into 30,445 Cogent contigs. Unlike the RefSeq annotation, PacBio and RNA-Seq data revealed that IR is the most frequent AS event in the rainbow trout. The PacBio-improved transcriptome was used to identify AS and isoform expression associated with fish

growth and muscle accretion, disease resistance, migration, and stress response. The improved transcriptome provides an avenue for future genetics and genomics studies to enhance aquaculture production efficiency.

MATERIALS AND METHODS

Production of Doubled Haploid Fish

Fish from the Swanson clonal line were obtained from the Washington State University (WSU) trout hatchery. Fish were produced as previously described (Scheerer et al., 1986; Young et al., 1996; Robison et al., 1999). Androgenesis was used to produce first-generation homozygous fish where eggs were gamma-irradiated before fertilization (Young et al., 1996; Robison et al., 1999). Sperms were collected from sexually matured homozygous males to perform another cycle of androgenesis producing homozygous clones (Scheerer et al., 1986). Three fish were dissected to collect tissues for Iso-Seq. Tissues included white muscle, red muscle, kidney, head kidney, spleen, stomach, gill, testis, heart, bone, skin, brain, liver, and intestine. Also, fertilized eggs at different developmental stages were collected.

Library Preparation and Sequencing

RNA was isolated from frozen tissues using TRIzol reagent (Life Technologies, Carlsbad, CA, United States) according to the guidelines of the manufacturer. RNA integrity was checked using Bioanalyzer (Agilent, Santa Clara, CA, United States[2]). RNA samples with RIN > 9 were used for Iso-Seq library preparation. The Clontech SMARTer PCR cDNA Synthesis Kit was used for first-strand cDNA synthesis according to PacBio instructions.

[2]https://www.agilent.com/

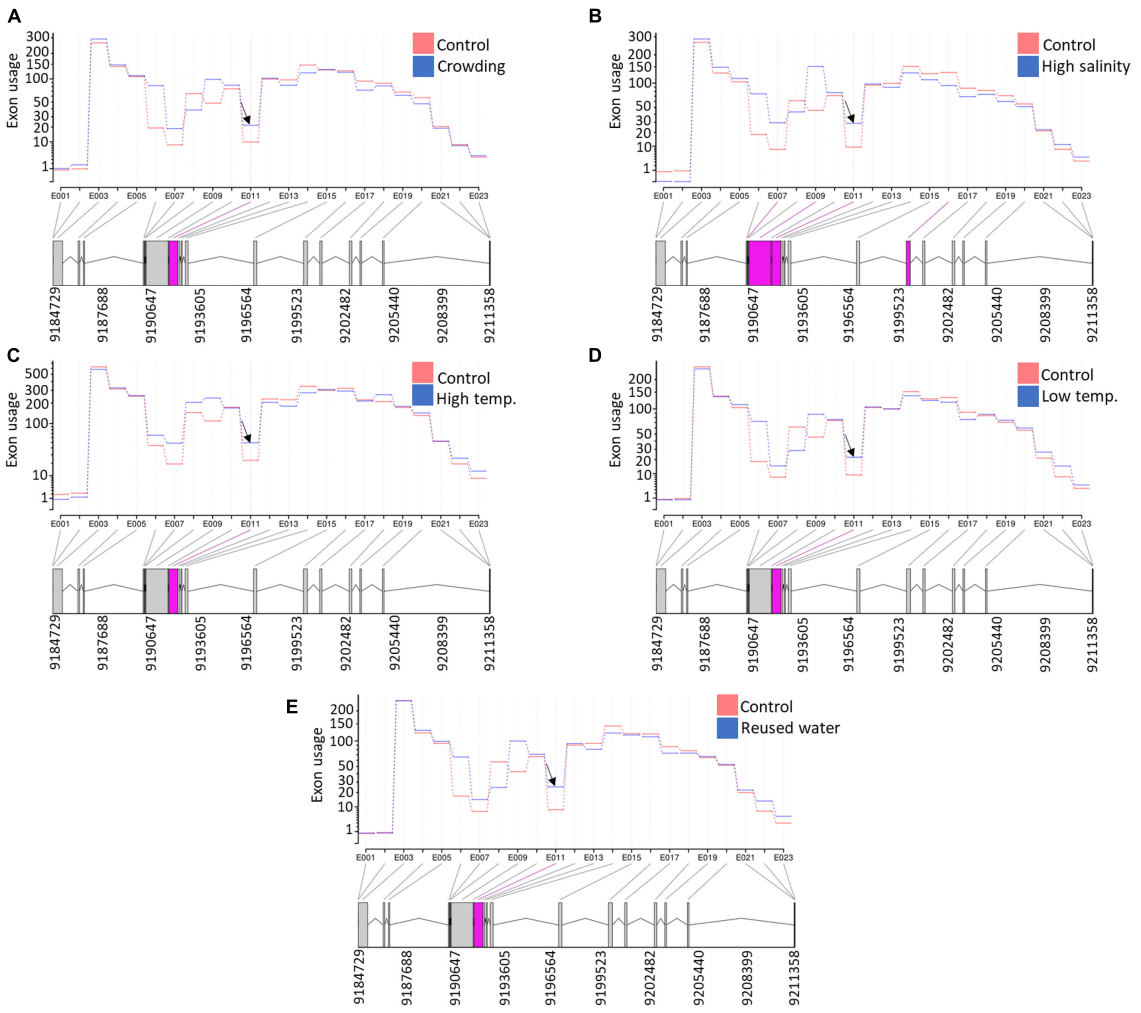


FIGURE 9 | A single exon (NC_035086.1:9191793-9192329) in an isoform encoding prolyl 4-hydroxylase subunit alpha-2 (*P4HA2*) was differentially used under all five studied stress conditions [**(A)** crowding, **(B)** high salinity, **(C)** high temperature, **(D)** low temperature, and **(E)** reused water]. Differentially used exonic features are shown in pink at the bottom of each panel.

Following PCR cycle optimization, a large-scale PCR was performed to generate double-stranded cDNA for SMRTbell library construction. AMPure PB Bead Purification of large-scale PCR products was performed and exonuclease was used to remove failed ligation products. AMPure PB beads were used to purify SMRTbell templates twice. The sequencing libraries were prepared by annealing a sequencing primer and binding a polymerase to SMRTbell templates. In total, 19 SMRT cells were sequenced.

Iso-Seq Analysis Pipeline

The pipeline included three initial steps: generation of CCS subreads, classification of FL reads, and clustering of FLnc reads. Polished CCS subreads were generated, using *CCS* v4.0.0, from the subreads bam files with a minimum quality of 0.9 (–min-rq 0.9). The default minimum number of FL subreads (n = 3) required to generate CCS for a zero-mode waveguide (ZMW) was used. FL transcripts were determined when the sequences had the poly(A) and the 5′ and 3′ cDNA primers. *Lima* v1.10.0 and *isoseq3 refine* v3.2.2 were used to remove the primers and poly(A) tails, respectively. The clustering algorithm ICE was used to obtain high-quality FL consensus sequences. The consensus transcripts were mapped to the Swanson rainbow trout reference genome (Pearse et al., 2020) using minimap2-2.17 (r941) (-ax splice -uf –secondary = no –C5 –O6,24 –B4) (Li, 2018). SAM files were sorted and used to collapse redundant isoforms using Cupcake v9.1.1[3]. Unmapped and poorly mapped isoforms were used as input to Cogent v6.0.0[4] to reconstruct the coding genome. The reconstructed contigs were used as a fake genome to process and collapse the unmapped and poorly mapped reads through the ToFU pipeline.

Transcriptome Characterization

SQANTI2 v6.0.0 was used to characterize and curate the long-read transcriptome (Tardaguila et al., 2018). The Swanson rainbow trout reference genome sequence [Omyk_1.0 (GCF_002163495.1)], annotation file (GTF), and quantification data were used as input to SQANTI2 to characterize/classify the collapsed isoforms and assess the quality of the sequencing data and the preprocessing ToFU pipeline (Gordon et al., 2015). A reference-guided error correction was implemented. Transcripts were classified into eight structural categories. Transcripts having splice junctions in a complete match with the reference transcripts were labeled as "FSM," whereas transcripts with partial consecutive matches with the annotated transcripts were labeled as "ISM." Novel isoforms of known genes were classified into Novel in Catalog "existing in the RefSeq annotation" (NIC) if containing a combination of annotated donor/acceptor sites or into Novel Not in Catalog (NNC) if at least containing one unannotated donor or acceptor site. In addition, "Genic Genomic" isoforms partially overlap with exons/introns of an annotated gene, whereas "Fusion" transcripts span two annotated loci. Transcripts in novel genes, compared with the RefSeq annotation, were classified as "intergenic" if existing outside the body of known genes, "Genic Intron" if completely contained within a known intron, and "Antisense" if overlapping the complementary strand of a known transcript. Potential artifacts were removed using SQANTI machine learning classifier. Transcripts flagged as intrapriming and RT-switching candidates were filtered out. The GeneMarkS-T (GMST) algorithm was implemented to predict ORFs from the corrected transcripts (Tang et al., 2015). Predicted ORFs were mapped to the Pfam protein domain database, Swiss-Prot, and TrEMBL database. The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 was used for gene enrichment analysis (FDR < 0.05) (Huang da et al., 2009).

AStalavista[5] was used with the raw annotation file generated from the Iso-Seq data to identify and classify AS events. SplAdder (Kahles et al., 2016) was used to identify AS events in rainbow trout tissues using bam files generated from the RNA-Seq datasets (Accession # PRJNA389609 and PRJEB4450) mapped to the trout genome. The frequencies of the six AS events (IR, ES, MX, ME, alternative 3′ splice sites, and alternative 5′ splice sites) and significant quantitative differences among tissues were determined from the SplAdder output files.

Non-coding RNA

Transcripts lacking ORFs or harboring ORFs less than 100 amino acids long were considered as potential non-coding transcripts. Those transcripts were aligned to all miRNA stem-loop sequences in miRbase[6] (release 22.1) and trout miRNA precursors (Juanchich et al., 2016) to identify pre-miRNAs. Also, putative non-coding transcripts were aligned to Rfam database to identify miRNAs and other non-coding classes. The remaining transcripts that did not match miRbase, trout miRNA precursors, and Rfam and longer than 200 bp were assessed for coding potential using CPC (CPC score $\leq$ 1) (Kong et al., 2007) and CPC2 web servers (Kang et al., 2017). Transcripts that were evaluated as non-coding were considered as putative lncRNA transcripts. Finally, these transcripts were aligned to the previous lncRNA assembly from rainbow trout (Al-Tobasei et al., 2016) using BLASTn (*E*-value 1e-5).

Differential Exon Usage Analysis

FastQC v0.11.9 was used to check the quality of the RNA-Seq datasets generated from rainbow trout fish under different biological conditions. Low-quality sequences were trimmed/removed using Trimmomatic v0.36 (Bolger et al., 2014). High-quality reads were mapped to the reference genome sequence by TopHat2 (Kim et al., 2013) with the default parameters.

DEXSeq package (v1.34.1) (Anders et al., 2012) was used to infer the DEU in the RNA-Seq datasets (Liu et al., 2014; Marancik et al., 2014; Hale et al., 2016; Ali et al., 2018; Paneru et al., 2018). DEXSeq counts the number of reads mapped to each exon (or part of an exon) in all samples. To infer changes in the relative

[3]https://github.com/Magdoll/cDNA_Cupcake

[4]https://github.com/Magdoll/Cogent

[5]http://astalavista.sammeth.net/

[6]http://www.mirbase.org/

exon usage, DEXSeq considers the change in the ratio of the number of reads mapped to an exon to read counts mapped to other exons of the same gene across conditions. DEXSeq uses two python scripts to prepare the GFF file and count the mapped reads. The first script, *dexseq_prepare_annotation.py*, was used to convert the GTF file with gene models into a gff file with collapsed exon counting bins. The second python script, *dexseq_count.py*, uses the sorted BAM/SAM alignment files to count the number of overlapping reads with each exon counting bin defined in the prepared GFF file. Default parameters were implemented in the DEXSeq analyses.

RT-PCR Validation of PacBio Isoforms and DEU

Reverse transcription (RT)-PCR was carried out to validate the long-read isoforms and to quantify exon usage as previously described (Ali et al., 2018). Primers used for RT-PCR analysis were designed using Primer3. First-strand cDNAs were synthesized using a Verso cDNA Synthesis Kit (Thermo Scientific, Hudson, NH, United States) following the instructions of the manufacturer. Each qPCR reaction contained a template (100 ng/μl), forward and reverse primers (10 μM working solution), and SYBR Green master mix (Bio-Rad, Hercules, CA, United States). Nuclease-free water was added to each reaction to achieve a final reaction volume of 10 μl. Quantification was performed in triplicates. β-*Actin* gene was used as an internal standard for normalization of expression. The PCR for all reactions started with 95°C for 30 s followed by 40 cycles. Each cycle lasted 15 s at 95°C, 30 s at the appropriate annealing temperature for each primer, and 30 s at 60°C. The expression was quantified using the delta delta Ct ($\Delta\Delta$Ct) method.

DATA AVAILABILITY STATEMENT

All the raw sequence data for this study were submitted to the NCBI BioProject at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA389609.

ETHICS STATEMENT

The Washington State University Institutional Animal Care and Use Committee reviewed and approved the animal study under protocol #02456.

AUTHOR CONTRIBUTIONS

MS conceived and designed the experiments. AA, GT, and MS performed the experiments. AA and MS analyzed the data and wrote the manuscript. GT contributed reagents, materials, and analysis tools. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by competitive grants (Nos. 2020-67015-30770 and 2021-67015-33388) from the United States Department of Agriculture, National Institute of Food and Agriculture (MS).

ACKNOWLEDGMENTS

We thank Paul Wheeler for providing tissues from the Swanson doubled haploid trout.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://osf.io/gwks6/?view_only=cbcffed4fda948ac974808c246517a4d

Supplementary Figure 1 | RT-switching and intra-priming.

Supplementary Figure 2 | Isoform distribution across structural categories before and after filtration.

Supplementary Figure 3 | Comparison between novel and annotated genes & transcript length across structural categories.

Supplementary Figure 4 | Distribution and characteristics of splice junctions.

Supplementary Figure 5 | Characterization of splice junctions in the corrected transcriptome generated from PacBio dataset.

Supplementary Figure 6 | Location, distribution, and frequency of polyadenylation cleavage site.

Supplementary Figure 7 | Characterization of unmapped/poorly-mapped collapsed transcripts.

Supplementary Figure 8 | Validation of a subset of differentially used exons by qPCR.

Supplementary Table 1 | Classification and annotation of all well-mapped PacBio transcripts.

Supplementary Table 2 | Hits of PacBio ORF (predicted from well-mapped transcripts) with three protein databases (Pfam, TrEMBL, and Swiss-Prot).

Supplementary Table 3 | PacBio well-mapped transcripts exhibiting homology with rainbow trout pre-miRNAs, miRNA stem-loop sequences available in the miRbase, and Rfam.

Supplementary Table 4 | Long non-coding PacBio well-mapped transcripts and their matches with our previous lncRNA assembly.

Supplementary Table 5 | Conservation of long non-coding PacBio well-mapped transcripts.

Supplementary Table 6 | genomic position and classification of splice junctions in PacBio well-mapped transcripts.

Supplementary Table 7 | Differentially regulated AS forms among rainbow trout tissues.

Supplementary Table 8 | Genomic coordinates of differentially regulated AS events among tissues (gff-version 3).

Supplementary Table 9 | Hits of PacBio ORF (predicted from unmapped and poorly mapped transcripts) with three protein databases (Pfam, TrEMBL, and Swiss-Prot).

Supplementary Table 10 | Coding potential of lncRNAs identified from unmapped/poorly-mapped transcripts.

Supplementary Table 11 | Unmapped/poorly mapped non-coding transcripts exhibiting homology with miRNA precursors.

Supplementary Table 12 | Statistics of rnaQUAST to assess the PacBio transcriptome quality compared to Swanson RefSeq.

Supplementary Table 13 | Unaligned, misassembled, and unannotated PacBio transcripts in the Swanson RefSeq genome.

Supplementary Table 14 | Differentially used exonic feature of well-mapped transcripts under different biological conditions.

Supplementary Table 15 | Differentially used exonic feature of unmapped/poorly-mapped transcripts under different biological conditions.

Supplementary Table 16 | Gene enrichment analysis of transcripts harboring differentially used exons.

Supplementary Table 17 | Differentially expressed transcripts between fish from resistant and susceptible genetic lines for BCWD.

Supplementary File 1 | Genomic coordinates of all well-mapped PacBio transcripts (gtf file).

Supplementary File 2 | ∼65K predicted ORFs from well-mapped PacBio transcripts (FASTA format).

Supplementary File 3 | ∼30K Cogent reconstructed contigs (FASTA format).

Supplementary File 4 | Annotation file for all unmapped/poorly mapped transcripts (gff file).

Supplementary File 5 | 41K predicted ORFs from unmapped/poorly mapped transcripts (FASTA format).

REFERENCES

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., et al. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7:11706.

Acerete, L., Balasch, J. C., Espinosa, E., Josa, A., and Tort, L. (2004). Physiological responses in Eurasian perch (*Perca fluviatilis*, L.) subjected to stress by transport and handling. *Aquaculture* 237, 167–178. doi: 10.1016/j.aquaculture.2004.03.018

Ahmed, Z. M., Riazuddin, S., Bernstein, S. L., Ahmed, Z., Khan, S., Griffith, A. J., et al. (2001). Mutations of the protocadherin gene PCDH15 cause Usher syndrome type 1F. *Am. J. Hum. Genet.* 69, 25–34. doi: 10.1086/321277

Ali, A., Al-Tobasei, R., Kenney, B., Leeds, T. D., and Salem, M. (2018). Integrated analysis of lncRNA and mRNA expression in rainbow trout families showing variation in muscle growth and fillet quality traits. *Sci. Rep.* 8:12111.

Ali, A., Al-Tobasei, R., Lourenco, D., Leeds, T., Kenney, B., and Salem, M. (2020). Genome-wide identification of loci associated with growth in rainbow trout. *BMC Genomics* 21:209.

Ali, A., Rexroad, C. E., Thorgaard, G. H., Yao, J., and Salem, M. (2014). Characterization of the rainbow trout spleen transcriptome and identification of immune-related genes. *Front. Genet.* 5:348.

Al-Tobasei, R., Paneru, B., and Salem, M. (2016). Genome-wide discovery of long non-coding RNAS in rainbow trout. *PLoS One* 11:e0148940. doi: 10.1371/journal.pone.0148940

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Res.* 22, 2008–2017. doi: 10.1101/gr.133744.111

Arends, R. J., Mancera, J. M., Munoz, J. L., Wendelaar Bonga, S. E., and Flik, G. (1999). The stress response of the gilthead sea bream (*Sparus aurata* L.) to air exposure and confinement. *J. Endocrinol.* 163, 149–157. doi: 10.1677/joe.0.1630149

Asaga, H., and Ishigami, A. (2007). Microglial expression of peptidylarginine deiminase 2 in the prenatal rat brain. *Cell Mol. Biol. Lett.* 12, 536–544.

Au, K. F., Sebastiano, V., Afshar, P. T., Durruthy, J. D., Lee, L., Williams, B. A., et al. (2013). Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 110, E4821–E4830.

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., et al. (2010). Deciphering the splicing code. *Nature* 465, 53–59. doi: 10.1038/nature09000

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noel, B., et al. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.* 5:3657.

Bhoj, E. J., Li, M., Ahrens-Nicklas, R., Pyle, L. C., Wang, J., Zhang, V. W., et al. (2015). Pathologic variants of the mitochondrial phosphate carrier SLC25A3: two new patients and expansion of the cardiomyopathy/skeletal myopathy phenotype with and without lactic acidosis. *JIMD Rep.* 19, 59–66. doi: 10.1007/8904_2014_364

Boley, N., Stoiber, M. H., Booth, B. W., Wan, K. H., Hoskins, R. A., Bickel, P. J., et al. (2014). Genome-guided transcript assembly by integrative analysis of RNA sequence data. *Nat. Biotechnol.* 32, 341–346. doi: 10.1038/nbt.2850

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bushmanova, E., Antipov, D., Lapidus, A., Suvorov, V., and Prjibelski, A. D. (2016). rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. *Bioinformatics* 32, 2210–2212. doi: 10.1093/bioinformatics/btw218

Cocquet, J., Chong, A., Zhang, G., and Veitia, R. A. (2006). Reverse transcriptase template switching and false alternative transcripts. *Genomics* 88, 127–131. doi: 10.1016/j.ygeno.2005.12.013

Coornaert, B., Baens, M., Heyninck, K., Bekaert, T., Haegman, M., Staal, J., et al. (2008). T cell antigen receptor stimulation induces MALT1 paracaspase-mediated cleavage of the NF-kappaB inhibitor A20. *Nat. Immunol.* 9, 263–271. doi: 10.1038/ni1561

De Gobbi, M., Viprakasit, V., Hughes, J. R., Fisher, C., Buckle, V. J., Ayyub, H., et al. (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 312, 1215–1217. doi: 10.1126/science.1126431

Denoeud, F., Aury, J. M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., et al. (2008). Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* 9:R175.

Ding, M., Van der Kwast, T. H., Vellanki, R. N., Foltz, W. D., McKee, T. D., Sonenberg, N., et al. (2018). The mTOR Targets 4E-BP1/2 restrain tumor growth and promote hypoxia tolerance in PTEN-driven prostate cancer. *Mol. Cancer Res.* 16, 682–695. doi: 10.1158/1541-7786.mcr-17-0696

Dolan, B. P., Fisher, K. M., Colvin, M. E., Benda, S. E., Peterson, J. T., Kent, M. L., et al. (2016). Innate and adaptive immune responses in migrating spring-run adult chinook salmon, *Oncorhynchus tshawytscha*. *Fish. Shellfish Immunol.* 48, 136–144. doi: 10.1016/j.fsi.2015.11.015

Doni, A., Stravalaci, M., Inforzato, A., Magrini, E., Mantovani, A., Garlanda, C., et al. (2019). The long pentraxin PTX3 as a link between innate immunity. *Tissue Remodel. Cancer Front. Immunol.* 10:712.

Du, X. G., and Zhu, M. J. (2018). Clinical relevance of lysyl oxidase-like 2 and functional mechanisms in glioma. *Onco. Targets Ther.* 11, 2699–2708. doi: 10.2147/ott.s164056

Edwards, I. J., Bruce, G., Lawrenson, C., Howe, L., Clapcote, S. J., Deuchars, S. A., et al. (2013). Na+/K+ ATPase alpha1 and alpha3 isoforms are differentially expressed in alpha- and gamma-motoneurons. *J. Neurosci.* 33, 9913–9919. doi: 10.1523/jneurosci.5584-12.2013

Fan, L. F., Wang, A. L., Miao, Y. T., Liao, S. A., Ye, C. X., and Lin, Q. C. (2016). Comparative proteomic identification of the hepatopancreas response to cold stress in white shrimp. *Litopenaeus Vannamei. Aquacult.* 454, 27–34. doi: 10.1016/j.aquaculture.2015.10.016

Farrell, A. P. (2011). *Encyclopedia of Fish Physiology: From Genome to Environment*. Amsterdam: Elsevier Science.

Feng, S., Xu, M., Liu, F., Cui, C., and Zhou, B. (2019a). Reconstruction of the full-length transcriptome atlas using PacBio Iso-Seq provides insight into the alternative splicing in *Gossypium australe*. *BMC Plant Biol.* 19:365.

Feng, X., Jia, Y., Zhu, R., Chen, K., and Chen, Y. (2019b). Characterization and analysis of the transcriptome in *Gymnocypris selincuoensis* on the Qinghai-Tibetan Plateau using single-molecule long-read sequencing and RNA-seq. *DNA Res.* 26, 353–363. doi: 10.1093/dnares/dsz014

Fox, S. E., Christie, M. R., Marine, M., Priest, H. D., Mockler, T. C., and Blouin, M. S. (2014). Sequencing and characterization of the anadromous steelhead (*Oncorhynchus mykiss*) transcriptome. *Mar. Genomics* 15, 13–15. doi: 10.1016/j.margen.2013.12.001

Gao, G., Magadan, S., Waldbieser, G. C., Youngblood, R. C., Wheeler, P. A., Scheffler, B. E., et al. (2021). A long reads-based de-novo assembly of the genome of the *Arlee homozygous* line reveals chromosomal rearrangements in rainbow trout. *G3 (Bethesda)* 11:jkab052.

Gavery, M. R., Nichols, K. M., Berejikian, B. A., Tatara, C. P., Goetz, G. W., Dickey, J. T., et al. (2019). Temporal dynamics of DNA methylation patterns in response to rearing juvenile steelhead (*Oncorhynchus mykiss*) in a hatchery versus simulated stream environment. *Genes (Basel)* 10:356. doi: 10.3390/genes10050356

Gordon, S. P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., et al. (2015). Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One* 10:e0132628. doi: 10.1371/journal.pone.0132628

Grau-Bove, X., Ruiz-Trillo, I., and Irimia, M. (2018). Origin of exon skipping-rich transcriptomes in animals driven by evolution of gene architecture. *Genome Biol.* 19:135.

Grelet, S., Link, L. A., Howley, B., Obellianne, C., Palanisamy, V., Gangaraju, V. K., et al. (2017). Addendum: a regulated. *Nat. Cell Biol.* 19:1443.

Guibbolini, M. E., and Lahlou, B. (1987). Adenylate cyclase activity in fish gills in relation to salt adaptation. *Life Sci.* 41, 71–78. doi: 10.1016/0024-3205(87)90558-3

Guria, A., Tran, D. D., Ramachandran, S., Koch, A., El Bounkari, O., Dutta, P., et al. (2011). Identification of mRNAs that are spliced but not exported to the cytoplasm in the absence of THOC5 in mouse embryo fibroblasts. *RNA* 17, 1048–1056. doi: 10.1261/rna.2607011

Hale, M. C., McKinney, G. J., Thrower, F. P., and Nichols, K. M. (2016). RNA-seq reveals differential gene expression in the brains of juvenile resident and migratory smolt rainbow trout (*Oncorhynchus mykiss*. *Comp. Biochem. Physiol. Part D Genom. Proteom.* 20, 136–150. doi: 10.1016/j.cbd.2016.07.006

Halver, J. E., and Hardy, R. W. (2002). "Fish nutrition," in *The Lipids*, 3rd Edn, eds J. R. Sargent, D. R. Tocher, and G. Bell (Cambridge MA: Academic Press), 182–246.

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Hube, F., Guo, J., Chooniedass-Kothari, S., Cooper, C., Hamedani, M. K., Dibrov, A. A., et al. (2006). Alternative splicing of the first intron of the steroid receptor RNA activator (SRA) participates in the generation of coding and noncoding RNA isoforms in breast cancer cell lines. *DNA Cell Biol.* 25, 418–428. doi: 10.1089/dna.2006.25.418

Husain, S., Yildirim-Toruner, C., Rubio, J. P., Field, J., Southern, M. S. G. C., Schwalb, M., et al. (2008). Variants of ST8SIA1 are associated with risk of developing multiple sclerosis. *PLoS One* 3:e2653. doi: 10.1371/journal.pone.0002653

Jeffries, A. R., Leung, S. K., Castanho, I., Moore, K., Davies, J. P., Dempster, E. L., et al. (2020). Full-length transcript sequencing of human and mouse identifies widespread isoform diversity and alternative splicing in the cerebral cortex. *bioRxiv*[Preprint] bioRxiv 2020.10.14.339200,

Juanchich, A., Bardou, P., Rue, O., Gabillard, J. C., Gaspin, C., Bobe, J., et al. (2016). Characterization of an extensive rainbow trout miRNA transcriptome by next generation sequencing. *BMC Genomics* 17:164.

Kahles, A., Ong, C. S., Zhong, Y., and Ratsch, G. (2016). SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* 32, 1840–1847. doi: 10.1093/bioinformatics/btw076

Kalsotra, A., and Cooper, T. A. (2011). Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* 12, 715–729. doi: 10.1038/nrg3052

Kang, Y. J., Yang, D. C., Kong, L., Hou, M., Meng, Y. Q., Wei, L., et al. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* 45, W12–W16.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., et al. (2007). , CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349.

Kornblihtt, A. R. I, Schor, E., Allo, M., Dujardin, G., Petrillo, E., and Munoz, M. J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* 14, 153–165. doi: 10.1038/nrm3525

Leeds, T. D., Silverstein, J. T., Weber, G. M., Vallejo, R. L., Palti, Y., Rexroad, C. E. III, et al. (2010). Response to selection for bacterial cold water disease resistance in rainbow trout. *J. Anim. Sci.* 88, 1936–1946. doi: 10.2527/jas.2009-2538

Li, C. Y., Chu, J. Y., Yu, J. K., Huang, X. Q., Liu, X. J., Shi, L., et al. (2004). Regulation of alternative splicing of Bcl-x by IL-6, GM-CSF and TPA. *Cell Res.* 14, 473–479. doi: 10.1038/sj.cr.7290250

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Liu, S., Gao, G., Palti, Y., Cleveland, B. M., Weber, G. M., and Rexroad, C. E. III (2014). RNA-seq analysis of early hepatic response to handling and confinement stress in rainbow trout. *PLoS One* 9:e88492. doi: 10.1371/journal.pone.0088492

Longaretti, A., Forastieri, C., Gabaglio, M., Rubino, T., Battaglioli, E., and Rusconi, F. (2020). Termination of acute stress response by the endocannabinoid system is regulated through lysine-specific demethylase 1-mediated transcriptional repression of 2-AG hydrolases ABHD6 and MAGL. *J. Neurochem.* 155, 98–110. doi: 10.1111/jnc.15000

Lorson, C. L., Hahnen, E., Androphy, E. J., and Wirth, B. (1999). A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc. Natl. Acad. Sci. U.S.A.* 96, 6307–6311. doi: 10.1073/pnas.96.11.6307

Lu, W. J., Gao, F. X., Wang, Y., Zhang, Q. Y., Li, Z., Zhang, X. J., et al. (2019). Differential expression of innate and adaptive immune genes in the survivors of three gibel carp gynogenetic clones after herpesvirus challenge. *BMC Genomics* 20:432.

Ma, R., Zhang, J., Liu, X., Li, L., Liu, H., Rui, R., et al. (2016). Involvement of Rab6a in organelle rearrangement and cytoskeletal organization during mouse oocyte maturation. *Sci. Rep.* 6:23560.

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

Marancik, D., Gao, G., Paneru, B., Ma, H., Hernandez, A. G., Salem, M., et al. (2014). Whole-body transcriptome of selectively bred, resistant-, control-, and susceptible-line rainbow trout following experimental challenge with *Flavobacterium psychrophilum*. *Front. Genet.* 5:453.

Martin, S., Nishimune, A., Mellor, J. R., and Henley, J. M. (2007). SUMOylation regulates kainate-receptor-mediated synaptic transmission. *Nature* 447, 321–325. doi: 10.1038/nature05736

Martinez, N. M., and Lynch, K. W. (2013). Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. *Immunol. Rev.* 253, 216–236. doi: 10.1111/imr.12047

Matthews, C. A., Shaw, J. E., Hooper, J. A., Young, I. G., Crouch, M. F., and Campbell, H. D. (2007). Expression and evolution of the mammalian brain gene Ttyh1. *J. Neurochem.* 100, 693–707. doi: 10.1111/j.1471-4159.2006.04237.x

Mayr, J. A., Zimmermann, F. A., Horvath, R., Schneider, H. C., Schoser, B., Holinski-Feder, E., et al. (2011). Deficiency of the mitochondrial phosphate carrier presenting as myopathy and cardiomyopathy in a family with three affected children. *Neuromuscul. Disord* 21, 803–808. doi: 10.1016/j.nmd.2011.06.005

Miura, P., Sanfilippo, P., Shenker, S., and Lai, E. C. (2014). Alternative polyadenylation in the nervous system: to what lengths will 3′ UTR extensions take us? *Bioessays* 36, 766–777. doi: 10.1002/bies.201300174

Nematollahi, A., Decostere, A., Pasmans, F., and Haesebrouck, F. (2003). Flavobacterium psychrophilum infections in salmonid fish. *J. Fish. Dis.* 26, 563–574. doi: 10.1046/j.1365-2761.2003.00488.x

Nudelman, G., Frasca, A., Kent, B., Sadler, K. C., Sealfon, S. C., Walsh, M. J., et al. (2018). High resolution annotation of zebrafish transcriptome using long-read sequencing. *Genome Res.* 28, 1415–1425. doi: 10.1101/gr.223586.117

Padmini, E., and Usha Rani, M. (2011). Heat-shock protein 90 alpha (HSP90alpha) modulates signaling pathways towards tolerance of oxidative stress and

enhanced survival of hepatocytes of *Mugil cephalus*. *Cell Stress Chaperones* 16, 411–425. doi: 10.1007/s12192-011-0255-9

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259

Pandey, B. N. (2007). *Fisheries And Fish Toxicology*. Daryaganj: APH Publishing Corporation.

Paneru, B., Ali, A., Al-Tobasei, R., Kenney, B., and Salem, M. (2018). Crosstalk among lncRNAs, microRNAs and mRNAs in the muscle 'degradome' of rainbow trout. *Sci. Rep.* 8:8416.

Parada, G. E., Munita, R., Cerda, C. A., and Gysling, K. (2014). A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Res.* 42, 10564–10578. doi: 10.1093/nar/gku744

Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadia-Cardoso, A., et al. (2020). Publisher Correction: Sex-dependent dominance maintains migration supergene in rainbow trout. *Nat. Ecol. Evol.* 4:170. doi: 10.1038/s41559-019-1076-y

Peng, L., Martin-Vasallo, P., and Sweadner, K. J. (1997). Isoforms of Na,K-ATPase alpha and beta subunits in the rat cerebellum and in granule cell cultures. *J. Neurosci.* 17, 3488–3502. doi: 10.1523/jneurosci.17-10-03488.1997

Pihlajaniemi, T., Myllyla, R., and Kivirikko, K. I. (1991). Prolyl 4-hydroxylase and its role in collagen synthesis. *J. Hepatol.* 3 13 Suppl, S2–S7.

Plisetskaya, E. M., Moon, T. W., Larsen, D. A., Foster, G. D., and Dickhoff, W. W. (1994). Liver glycogen, enzyme activities, and pancreatic hormones in juvenile atlantic salmon (Salmo salar) during their first summer in seawater. *Can. J. Fish. Aquat. Sci.* 51, 567–576. doi: 10.1139/f94-059

Proudfoot, N. J., and Brownlee, G. G. (1976). 3′ non-coding region sequences in eukaryotic messenger RNA. *Nature* 263, 211–214. doi: 10.1038/263211a0

Ramberg, S., Hoyheim, B., Ostbye, T. K., and Andreassen, R. (2021). A de novo Full-Length mRNA transcriptome generated from hybrid-Corrected pacbio long-reads improves the transcript annotation and identifies thousands of novel splice variants in atlantic salmon. *Front. Genet.* 12:656334.

Robison, B. D., Wheeler, P. A., and Thorgaard, G. H. (1999). Variation in development rate among clonal lines of rainbow trout (*Oncorhynchus mykiss*). *Aquaculture* 173, 131–141. doi: 10.1016/s0044-8486(98)00481-5

Rodriguez, J. M., Pozo, F., di Domenico, T., Vazquez, J., and Tress, M. L. (2020). An analysis of tissue-specific alternative splicing at the protein level. *PLoS Comput. Biol.* 16:e1008287. doi: 10.1371/journal.pcbi.1008287

Salem, M., Paneru, B., Al-Tobasei, R., Abdouni, F., Thorgaard, G. H., Rexroad, C. E., et al. (2015). Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout. *PLoS One* 10:e0121778. doi: 10.1371/journal.pone.0121778

Salem, M., Rexroad, C. E. III, Wang, J., Thorgaard, G. H., and Yao, J. (2010). Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches. *BMC Genomics* 11:564.

Scheerer, P. D., Thorgaard, G. H., Allendorf, F. W., and Knudsen, K. L. (1986). Androgenetic rainbow-trout produced from inbred and outbred sperm sources show similar survival. *Aquaculture* 57, 289–298. doi: 10.1016/0044-8486(86)90207-3

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., et al. (2000). *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 101, 671–684. doi: 10.1016/s0092-8674(00)80878-8

Seo, P. J., Park, M. J., and Park, C. M. (2013). Alternative splicing of transcription factors in plant responses to low temperature stress: mechanisms and functions. *Planta* 237, 1415–1424. doi: 10.1007/s00425-013-1882-4

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol* 1962, 227–245. doi: 10.1007/978-1-4939-9173-0_14

Sharma, S., Villamor, J. G., and Verslues, P. E. (2011). Essential role of tissue-specific proline synthesis and catabolism in growth and redox balance at low water potential. *Plant Physiol.* 157, 292–304. doi: 10.1104/pp.111.183210

Sherstnev, A., Duc, C., Cole, C., Zacharaki, V., Hornyik, C., Ozsolak, F., et al. (2012). Direct sequencing of *Arabidopsis thaliana* RNA reveals patterns of cleavage and polyadenylation. *Nat. Struct. Mol. Biol.* 19, 845–852. doi: 10.1038/nsmb.2345

Shih, Y. H., Dvornikov, A. V., Zhu, P., Ma, X., Kim, M., Ding, Y., et al. (2016). Exon- and contraction-dependent functions of titin in sarcomere assembly. *Development* 143, 4713–4722.

Silverstein, J. T., Vallejo, R. L., Palti, Y., Leeds, T. D., Rexroad, C. E. III, Welch, T. J., et al. (2009). Rainbow trout resistance to bacterial cold-water disease is moderately heritable and is not adversely correlated with growth. *J. Anim. Sci.* 87, 860–867. doi: 10.2527/jas.2008-1157

Song, S., Li, D., Yang, C., Yan, P., Bai, Y., Zhang, Y., et al. (2018). Overexpression of NELFCD promotes colorectal cancer cells proliferation, migration, and invasion. *Onco. Targets Ther.* 11, 8741–8750. doi: 10.2147/ott.s186266

Steijger, T., Abril, J. F., Engstrom, P. G., Kokocinski, F., Consortium, R., Hubbard, T. J., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184. doi: 10.1038/nmeth.2714

Tang, S., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* 43, e78. doi: 10.1093/nar/gkv227

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., et al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* 28, 396–411. doi: 10.1101/gr.222976.117

Thornqvist, P. O., Hoglund, E., and Winberg, S. (2015). Natural selection constrains personality and brain gene expression differences in Atlantic salmon (Salmo salar). *J. Exp. Biol.* 218, 1077–1083. doi: 10.1242/jeb.114314

Tian, Y., Wen, H., Qi, X., Zhang, X., Liu, S., Li, B., et al. (2019). Characterization of full-length transcriptome sequences and splice variants of *Lateolabrax maculatus* by single-molecule long-read sequencing and their involvement in salinity regulation. *Front. Genet.* 10:1126.

Tilgner, H., Grubert, F., Sharon, D., and Snyder, M. P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* 111, 9869–9874. doi: 10.1073/pnas.1400447111

Tran, D. D., Koch, A., and Tamura, T. (2014). THOC5, a member of the mRNA export complex: a novel link between mRNA export machinery and signal transduction pathways in cell proliferation and differentiation. *Cell Commun. Signal.* 12:3. doi: 10.1186/1478-811x-12-3

Treutlein, B., Gokce, O., Quake, S. R., and Sudhof, T. C. (2014). Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 111, E1291–E1299.

Tseng, E. (2020). *Cogent-8.0.0*. Available online at: https://github.com/Magdoll/Cogent (accessed February 14, 2021).

Tullio-Pelet, A., Salomon, R., Hadj-Rabia, S., Mugnier, C., de Laet, M. H., Chaouachi, B., et al. (2000). Mutant WD-repeat protein in triple-A syndrome. *Nat. Genet.* 26, 332–335. doi: 10.1038/81642

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., et al. (2017). The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.* 45, D626–D634.

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., et al. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7:11708.

Weil, D., Blanchard, S., Kaplan, J., Guilford, P., Gibson, F., Walsh, J., et al. (1995). Defective myosin VIIA gene responsible for Usher syndrome type 1B. *Nature* 374, 60–61. doi: 10.1038/374060a0

Whalen, K. G., Parrish, D. L., and McCormick, S. D. (1999). Migration timing of Atlantic salmon smolts relative to environmental and physiological factors. *Trans. Am. Fish. Soc.* 128, 289–301. doi: 10.1577/1548-8659(1999)128<0289:mtoass>2.0.co;2

Williamson, L., Saponaro, M., Boeing, S., East, P., Mitter, R., Kantidakis, T., et al. (2017). UV irradiation induces a non-coding RNA that functionally opposes the protein encoded by the same gene. *Cell* 84:e13.

Wu, H., Mu, L., Yin, X., Han, K., Yan, F., Zhou, E., et al. (2020). A microfibril-associated glycoprotein 4 (MFAP4) from *Nile tilapia* (*Oreochromis niloticus*) possesses agglutination and opsonization ability to bacterial pathogens. *Fish. Shellfish Immunol.* 104, 182–191. doi: 10.1016/j.fsi.2020.06.009

Wu, Q., and Maniatis, T. (1999). A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* 97, 779–790. doi: 10.1016/s0092-8674(00)80789-8

Wu, X., Liu, M., Downie, B., Liang, C., Ji, G., Li, Q. Q., et al. (2011). Genome-wide landscape of polyadenylation in *Arabidopsis* provides evidence for extensive

alternative polyadenylation. *Proc. Natl. Acad. Sci. U.S.A.* 108, 12533–12538. doi: 10.1073/pnas.1019732108

Xiong, G., Stewart, R. L., Chen, J., Gao, T., Scott, T. L., Samayoa, L. M., et al. (2018). Collagen prolyl 4-hydroxylase 1 is essential for HIF-1alpha stabilization and TNBC chemoresistance. *Nat. Commun.* 9:4456.

Yanaka, N. (2007). Mammalian glycerophosphodiester phosphodiesterases. *Biosci. Biotechnol. Biochem.* 71, 1811–1818. doi: 10.1271/bbb.70062

Yi, S., Zhou, X., Li, J., Zhang, M., and Luo, S. (2018). Full-length transcriptome of *Misgurnus anguillicaudatus* provides insights into evolution of genus *Misgurnus*. *Sci. Rep.* 8:11699.

Yoshiba, Y., Kiyosue, T., Katagiri, T., Ueda, H., Mizoguchi, T., Yamaguchi-Shinozaki, K., et al. (1995). Correlation between the induction of a gene for delta 1-pyrroline-5-carboxylate synthetase and the accumulation of proline in *Arabidopsis thaliana* under osmotic stress. *Plant J.* 7, 751–760. doi: 10.1046/j.1365-313x.1995.07050751.x

Youn, J. Y., Dunham, W. H., Hong, S. J., Knight, J. D. R., Bashkurov, M., Chen, G. I., et al. (2018). High-density proximity mapping reveals the subcellular organization of mRNA-associated granules and bodies. *Mol. Cell* 51:e11.

Young, W. P., Wheeler, P. A., Fields, R. D., and Thorgaard, G. H. (1996). DNA fingerprinting confirms isogenicity of androgenetically derived rainbow trout lines. *J. Hered* 87, 77–80. doi: 10.1093/oxfordjournals.jhered.a022960

Zhou, J., Yu, Q., and Zou, T. (2008). Alternative splicing of exon 10 in the tau gene as a target for treatment of tauopathies. *BMC Neurosci* 9 Suppl 2:S10.

Zhou, J., Zheng, X., and Shen, H. (2012). Targeting RNA-splicing for SMA treatment. *Mol. Cells* 33, 223–228. doi: 10.1007/s10059-012-0005-6

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.

Copyright © 2021 Ali, Thorgaard and Salem. 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: 30 August 2021
doi: 10.3389/fgene.2021.664974

PacBio Single-Molecule Long-Read Sequencing Provides New Light on the Complexity of Full-Length Transcripts in Cattle

Tianpeng Chang[1], Bingxing An[1], Mang Liang[1], Xinghai Duan[2], Lili Du[1], Wentao Cai[1], Bo Zhu[1], Xue Gao[1], Yan Chen[1], Lingyang Xu[1], Lupei Zhang[1], Huijiang Gao[1] and Junya Li[1]**

[1] Laboratory of Molecular Biology and Bovine Breeding, Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China, [2] College of Animal Science and Technology, Southwest University, Chongqing, China

OPEN ACCESS

Edited by:
Christa Kuehn,
Leibniz Institute for Farm Animal Biology (FBN), Germany

Reviewed by:
Emily Louise Clark,
University of Edinburgh, United Kingdom
Hamid Beiki,
Iowa State University, United States

***Correspondence:**
Huijiang Gao
gaohuijiang@caas.cn
Junya Li
lijunya@caas.cn

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

Received: 06 February 2021
Accepted: 06 August 2021
Published: 30 August 2021

Citation:
Chang T, An B, Liang M, Duan X, Du L, Cai W, Zhu B, Gao X, Chen Y, Xu L, Zhang L, Gao H and Li J (2021) PacBio Single-Molecule Long-Read Sequencing Provides New Light on the Complexity of Full-Length Transcripts in Cattle. Front. Genet. 12:664974. doi: 10.3389/fgene.2021.664974

Cattle (*Bos taurus*) is one of the most widely distributed livestock species in the world, and provides us with high-quality milk and meat which have a huge impact on the quality of human life. Therefore, accurate and complete transcriptome and genome annotation are of great value to the research of cattle breeding. In this study, we used error-corrected PacBio single-molecule real-time (SMRT) data to perform whole-transcriptome profiling in cattle. Then, 22.5 Gb of subreads was generated, including 381,423 circular consensus sequences (CCSs), among which 276,295 full-length non-chimeric (FLNC) sequences were identified. After correction by Illumina short reads, we obtained 22,353 error-corrected isoforms. A total of 305 alternative splicing (AS) events and 3,795 alternative polyadenylation (APA) sites were detected by transcriptome structural analysis. Furthermore, we identified 457 novel genes, 120 putative transcription factors (TFs), and 569 novel long non-coding RNAs (lncRNAs). Taken together, this research improves our understanding and provides new insights into the complexity of full-length transcripts in cattle.

Keywords: cattle, PacBio single-molecule long-read sequencing, full-length transcript, alternative splicing, alternative polyadenylation, long non-coding RNA

INTRODUCTION

Cattle (*Bos taurus*) are an agriculturally important species that provide human beings with large quantities of high-quality protein. As a typical ruminant animal, cattle still play a great role in sustainable agriculture since they can effectively utilize pastures, silage, and high-fiber crop residues. Nowadays, genomic information plays an important role in accelerating the molecular breeding process of cattle, so an accurate and complete reference genome and annotation are essential for genetic mechanism research, Quantitative trait locus (QTL) mapping, and genomic selection of important production traits for cattle breeding. The latest reference genome assembly (ARS-UCD1.2) was first reported in 2018, assembling 2.7 Gb of the genome (Rosen et al., 2020). The annotation of the ARS-UCD1.2 assembly (NCBI release 106), resulted in 21,039 protein-coding genes, 9,357 non-coding genes, and 4,569 pseudogenes (Rosen et al., 2020).

This assembly has a higher sequence continuity and accuracy than the previous reference (UMD3.1.1), and the protein models predicted in ARS-UCD1.2 assembly annotation are generally more complete than in UMD3.1.1 annotation (Zimin et al., 2009; Rosen et al., 2020). However, due to the diversity of cattle breeds, there are distinct genetic characteristics and allelic variations between breeds, so more genomic information is needed to explore to complete the annotation of the structural and functional of the current cattle reference genome (Crysnanto and Pausch, 2020).

The next-generation sequencing (NGS) technologies, such as the Illumina platform, has stimulated the construction of genome and transcriptome resources for many species (Lan et al., 2012; Li et al., 2012; Oono et al., 2013; Du et al., 2017; Carruthers et al., 2018). NGS is accurate, cost-effective, and supported by a wide range of analysis software and pipelines. However, natural nucleic acid polymers span eight orders of magnitude in length, and sequencing them in short amplified fragments complicates the task of reconstructing and counting the original molecules (Conesa et al., 2016; Amarasinghe et al., 2020). Therefore, it is difficult to accurately reconstruct expressed full-length transcripts, predicting splice isoforms and analyzing the transcriptome diversity based on NGS reads (Wang L. et al., 2019). The third-generation sequencing (TGS) technologies, which include Pacific Biosciences single-molecule real-time (SMRT) and Oxford Nanopore Technologies (ONT) nanopore sequencing technology, can avoid the disadvantages of NGS technology and obtain high-quality long-read transcripts due to their ability to sequence reads up to 50 kb (Jia et al., 2018; Zuo et al., 2018). Then, we can obtain more complete transcripts and analyze structural variations in the genome and transcriptome (Eid et al., 2009; Koren et al., 2012; Sharon et al., 2013; Tilgner et al., 2015; Amarasinghe et al., 2020). However, the long-read sequencing technologies still have some limitations, such as higher error rates and relatively low throughput (Wang et al., 2016; Wang X. et al., 2019). Several studies have indicated that the error rate for SMRT sequencing (15%) is higher than the Illumina platform (1%) (Weirather et al., 2017; Amarasinghe et al., 2020). Moreover, NGS and TGS technologies have different error models. The Illumina short reads mainly contain miscalled bases with increasing frequency toward read ends, while SMRT sequencing generates primarily insertion-deletion errors in a random pattern (Hackl et al., 2014; Beiki et al., 2019). Fortunately, some research has shown that the accurate and abundant NGS reads can be used to correct errors in TGS and improve the accuracy of long reads sequencing (Sharon et al., 2013; Xu et al., 2015).

Recently, the strategy of combining SMRT sequencing and Illumina RNA-seq data to detect structural variation, novel genes, or isoforms and reveal functional variety at the transcriptional level has become more prevalent (Amarasinghe et al., 2020). According to this strategy, 24,797 alternative splicing (AS) and 11,184 alternative polyadenylation (APA) events were detected in rabbit (Chen et al., 2017). In pig, researchers detected 28,127 novel isoforms from 26,881 novel genes based on the high-quality full-length isoforms. Meanwhile, they identified more than 92,000 novel AS events and found intron retention (RI) and exon skipping (ES) were the main AS events in AS model (Li et al., 2018). Another pig study observed many unique transcripts and extended more than 6,000 known gene borders, and the extensions were verified by independent ChIP-seq, 3′-RNA-seq experiment, and human CAGE data (Beiki et al., 2019). In maize, many unique isoforms and higher isoform densities were detected with SMRT sequencing, and 867 novel lncRNAs were identified which had a much longer mean length than those identified by Illumina short-read sequencing (Wang et al., 2016). These findings have provided important information for improving genome annotation and gene models for different species.

To improve the transcriptional information and explore the complexity of cattle transcriptome, we generated high-quality FLNC reads in the present study by PacBio SMRT sequencing. We first used Illumina short reads to correct the relatively high error rates of SMRT long reads. Then, AS and APA events were detected to explore the structural complexity of transcripts. We predicted the novel genes and annotated them using seven databases, and the transcript factor (TF) and lncRNAs were investigated. Accordingly, our research contributed to the exploration of the splice isoforms and transcriptome diversity of cattle, increased our understanding of the structure of the transcript, and facilitates the further study of the genetics of cattle.

MATERIALS AND METHODS

Collection of Samples and RNA Preparation

The sample collection experiment was conducted on the farm of Hanjiang Beef Cattle Co., Ltd. (Hubei, China). The use of animals and private land in this study was approved by their respective legal owners. The cattle were raised in the same feeding strategies and conditions. Three unrelated male Simmental beef calves were collected at 0 days of age. The calves were stunned by electrical shock and killed while unconscious. Then, six tissues, consisting of cerebrum, rumen, liver, spleen, renal cortical, and longissimus muscle, were sampled for each cattle, snap frozen, and stored in liquid nitrogen until use. Subsequently, all 18 samples were subjected to RNA extraction using TRIzol reagent (Takara, Dalian, China) according to the manufacturer's instructions. The RNA concentration was measured using Qubit® RNA Assay Kit in Qubit® 2.0 Fluorometer (Life Technologies, Carlsbad, CA, United States). RNA integrity and purity were assessed using the Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, United States) and RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, Palo Alto, CA, United States), respectively. Qualified

RNA samples were then used for further cDNA library construction and sequencing.

SMRT Library Preparation and PacBio Sequencing

One microgram for each RNA sample was equally pooled together and prepared for PacBio SMRT library construction. Full-length cDNA was synthesized by use of the Clontech SMARTer PCR cDNA Synthesis Kit (TaKaRa, Dalian, China), and cDNA fraction and length selection (<4 kb and >4 kb) was performed using the BluePippin^{TM} Size Selection System (Sage Science, Beverly, MA, United States). Then, one SMRT bell library was generated using the Pacific Biosciences DNA Template Prep Kit 2.0 (Pacific Biosciences, CA, United States) according to the standard method. All the samples were pooled onto one SMRT cell library, so SMRT data of different tissues were not available, and tissue-specific analysis of isoforms was not possible. Finally, SMRT sequencing was performed on the Pacific Bioscience Sequel platform.

Illumina cDNA Library Construction and NGS Analysis

All the 18 RNA samples were prepared for Illumina unstranded cDNA library construction. Briefly, polyadenylated RNA was isolated and fragmented into ∼200 bp fragments. The first and second cDNA strands were synthesized successively. The repaired and purified double-stranded cDNA fragments were selected by size. Then, the qualified and amplified mRNA libraries were finally sequenced on an Illumina NovaSeq 6000 platform, and 150 bp paired-end raw short reads were generated. The raw short reads were subjected to quality filtering using NGS QC Toolkit (v2.3.3) (Patel and Jain, 2012), for which we trimmed the first five bases from the 5′ end of the reads and removed reads consisting of the low-quality bases (QA $\leq$ 30) >20% or ambiguous bases >1%. To produce corrected PacBio long reads, Illumina clean reads are used for independently assembling transcripts using Hisat2 (v2.1.0) and Stringtie (v2.1.1) (Kim et al., 2015; Pertea et al., 2015).

Quality Filtering and Error Correction for PacBio Long Reads

The PacBio raw data were processed using the SMRTlink (v7.0) software with parameters: minReadScore = 0.75, minLength = 200. Circular consensus sequences (CCSs) were generated from subread BAM files (parameters: min_length 200, max_drop_fraction 0.8, no_polish TRUE, min_zscore –9999, min_passes 2, min_predicted_accuracy 0.8, max_length 18,000) and then a BAM file of CCS was generated. By searching for the 5′ and 3′ adapters and the poly(A) tail, the CCS was classified into full length and non-full length (NFL) reads. Full-length reads with all the three elements and any additional copies of the adapter sequence within the DNA fragment were classified as FLNC. We then used ICE (Iterative Clustering for Error Correction) to identify the consensus isoforms that formed FLNC and polished the consensus isoforms with NFL reads to obtain high-quality isoforms with post-correction accuracy above 99% using Quiver (parameters: hq_quiver_min_accuracy 0.99, bin_by_primer false, bin_size_kb 1, qv_trim_5p 100, qv_trim_3p 30). Next, the Illumina clean data generated above was used to correct nucleotide indels and mismatches in consensus reads with the LoRDEC software (v0.7) (Salmela and Rivals, 2014). LoRDEC uses a hybrid error correction strategy that builds a succinct de Bruijn graph representing the Illumina short reads, and seeks a corrective sequence for each erroneous region in the PacBio long reads by traversing chosen paths in the graph (Fu et al., 2019). Then, a high-quality PacBio corrected consensus reads dataset without redundant isoforms was constructed. Finally, we used the method proposed by Salmela and Rivals (2014) to evaluate the PacBio data error rate before and after error correction.

Mapping to the Reference Genome and Structural Analysis of Genes

Corrected isoforms were aligned to the cattle reference genome (ARS-UCD1.2) with the Genome Mapping and Alignment Program (GMAP, version: 2017-06-20) using the following parameters: –no-chimeras, –cross-species, –expand-offsets 1 -B 5 -K 50000 -f samse -n 1 (Wu and Watanabe, 2005). The genome annotation file (NCBI release 106) was used for gene and transcript determination. Genome-guided construction of the full transcriptome was successful. Transcripts structure analysis was performed using the TAPIS pipeline (Version 1.2.1) (Abdel-Ghany et al., 2016). AS events including IR, ES, alternative 3′ splice site (Alt.3′), alternative 5′ splice site (Alt.5′), mutually exclusive exon (MEE), alternative first exons (AF), and alternative last exons (AL) were identified and classified using SUPPA (v2.3) (Alamancos et al., 2015). Among them, IR is defined as when one intron is retained within a longer exon and flanked by two shorter exons simultaneously. ES is defined as when an exon is absent in some transcripts but present in others. If an intron is excised at more than one site and linked to its 5′ or 3′ exons with different boundaries, they are considered as the Alt. 5′ and Alt. 3′ (Chen et al., 2017). Alternative terminal exon regulations including AF or AL are types of AS, which couples with alternative transcription start sites and APA, respectively (Lian et al., 2020). APA events were analyzed by TAPIS described previously. The transcription factors (TFs) were predicted using the animalTFDB 2.0 database (Zhang et al., 2015).

Due to the limitation of library construction, we can only obtain lncRNA containing polyA tails. The following four tools were combined: Coding Potential Calculator (CPC) (Kong et al., 2007), Coding–Non-Coding Index (CNCI) (Sun et al., 2013), PLEK (Li et al., 2014), and Pfam database (Finn et al., 2016). They were used to sort non-protein-coding RNA candidates from putative protein-coding RNAs in the transcripts. Putative protein-coding RNAs were filtered out using minimum length and exon number thresholds. The transcripts longer than 200 bp with more than two exons were selected as lncRNAs candidates and then screened using CPC/CNCI/PLEK/Pfam, as these tools can distinguish protein-coding from the non-protein-coding

genes. Only the transcripts identified in the four databases were regarded as lncRNAs.

Novel Gene Prediction and Functional Annotation

Here we defined a novel (compared to NCBI gene-build) gene as a gene putatively encoding a detected transcript that does not match any annotated gene in the cattle reference genome (ARS-UCD1.2). To obtain comprehensive annotation information, functional annotations of the novel genes were conducted using the following seven databases: NR (NCBI non-redundant protein sequences); NT (NCBI non-redundant nucleotide sequences); Pfam (Protein family); KOG (EuKaryotic Ortholog Groups of proteins) (Tatusov et al., 2000); Swiss-Prot (a manually annotated and reviewed protein sequence database); KO (KEGG Ortholog database) (Kanehisa et al., 2004); and GO (Gene Ontology). We used the software of BLAST and set the *e*-value "1e-10" in the NT database analysis. We used the software of Diamond BLASTX and set the *e*-value "1e-10" in the NR, KOG, Swiss-Prot, and KEGG database analyses. We used the software of Hmmscan in the Pfam database analysis. For each transcript searched in the four databases, functional information for the best-matched sequence was assigned to the query transcript.

RESULTS

General Properties of PacBio Sequencing

To reveal the complexity of the transcriptome in cattle, six tissues (cerebrum, rumen, liver, spleen, renal cortical, and longissimus muscle) were collected and a pooled RNA sample of them was sequenced with the Pacific Bioscience Sequel platform to accurately capture full-length sequences and uncover full-length splice variants. With SMRT sequencing, 23.6 Gb of raw data consisting of 441,444 raw polymerase reads was generated. Then, a total of 14,750,730 subreads (22.5 Gb) were obtained, with an average read length of 1,526 bp and N50 of 2,367 bp. To provide more accurate sequence information, CCS was generated from subreads that pass at least 2 times through the insert, and a total of 381,423 CCSs were obtained. In all, 286,688 CCSs were identified as full-length reads, and 276,295 were identified as FLNC reads with low artificial concatemers. The mean length of FLNC reads was 2,241 bp. The length distribution of the subreads, CCSs, FLNC reads is shown in **Figure 1** and **Table 1**.

Error Correction of PacBio Long Reads Using Illumina Reads

The FLNC reads with similar sequences were clustered together using the ICE (Iterative isoform-clustering) algorithm, and each cluster was considered as a consistent sequence. Combined with NFL sequences, the Quiver program was used to polish the consistent sequences in each cluster. To further correct the relatively high error rates of PacBio long reads, we generated ∼981.4 million clean reads of NGS sequencing clean data. Then, the Illumina short reads were used for correcting the consensus isoform sequences of PacBio long reads. The LoRDEC software was used to correct polished consensus sequences, resulting in 22,353 corrected sequences, with an N50 length of 2,921 bp and a mean read length of 2,379 bp (**Table 1** and **Figure 1D**). To calculate the error rate, the raw and corrected PacBio long reads were aligned to the cattle reference genome (ARS-UCD1.2) with BLASR (Chaisson and Tesler, 2012). The error rate is defined as the sum of the numbers of bases of insertions, deletions, and substitutions in the alignment divided by the length of aligned regions for each read. After calculation, the error rates of PacBio long reads before and after error correction were 9.72 and 2.84%, respectively.

Genome Mapping

We compared all the corrected sequences against the cattle reference genome using GMAP software. A total of 21,863 reads (97.81%) were mapped to the reference genome. According to mapping results, these reads could be divided into four groups: unmapped, multiple mapped, mapped to "+" and mapped to "−" (**Table 2** and **Figure 2A**). The unmapped group consisted of 490 reads (2.19%) with no significant mapping to the reference genome. The multiple mapped group consisted of 1,583 reads (7.08%) showing multiple alignments. The group of results mapped to "+" consisted of 10,778 reads (48.22%) that were mapped to the positive strand of the reference genome, and the group of results mapped to "−" consisted of 9,502 reads (42.51%) which were mapped to the opposite strand of the reference genome. The curve of the corrected isoform numbers reached a saturation level (**Figure 2B**), and high-quality reads (i.e., with coverage and identity values over 98%) accounted for over 90% (**Figure 2C**).

Gene Structure Analysis

Gene structure analysis was performed using the TAPIS pipeline. The GMAP output file and genome annotation (NCBI release 106) file were used for gene and transcript isoforms determination. Reads that were mapped to different exons in known gene regions were considered new isoforms, and isoforms spanning two or more genes are removed from downstream splice isoform analysis. Then, 4,862 isoforms were identified, and they can be divided into three types: (1) 2,104 isoforms of known genes; (2) 2,250 novel isoforms from known genes; and (3) 508 novel isoforms from novel genes (**Figure 2D** and **Supplementary Table 1**). Also, 457 novel genes (**Supplementary Table 2**) were identified and they were annotated using the Nt, Nr, Swissprot, GO, KOG, Pfam, and KEGG databases. A total of 63 genes had hits on all 7 databases, and 457 had hits on at least 1 database (**Figure 3A** and **Supplementary Table 3**). We analyzed homologous species by comparing the novel genes to the NR database, and the results showed that the largest five number of the novel genes were distributed in *Bos taurus* (64), *Bos mutus* (28), *Bos indicus* (20), *Ovis aries* (12), and *Macaca fascicularis* (10) (**Figure 3B**). By using all annotated genes in the cattle genome as background, GO analysis showed that "Cell," "binding," and "cellular process" were ranked as the most enriched items in the "cellular components," "molecular functions," and "biological process" categories, respectively (**Figure 3C**). KOG

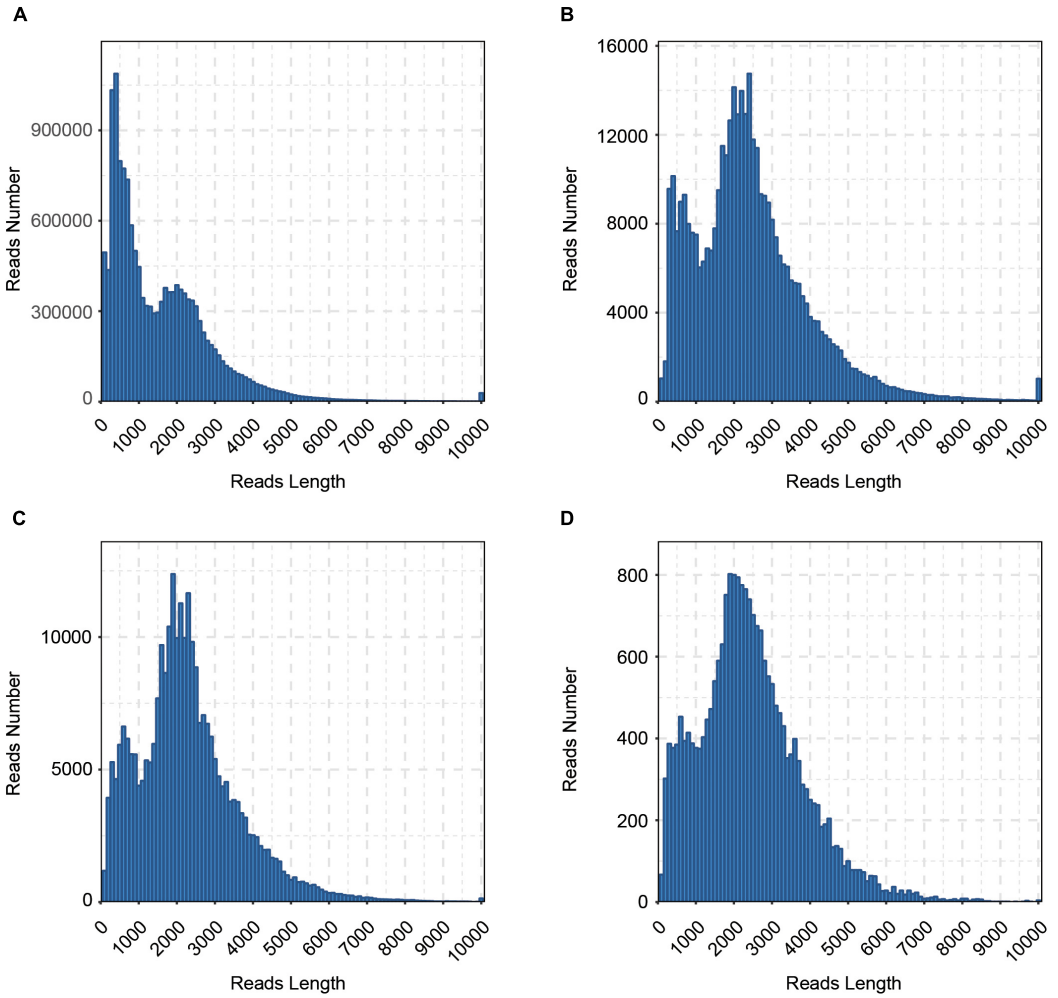


FIGURE 1 | Length distributions of PacBio SMRT sequencing. **(A)** Number and length distributions of 14,750,730 Subreads sequences. **(B)** Number and length distributions of 381,423 CCS sequences. **(C)** Number and length distributions of 276,295 FLNC sequences. **(D)** Number and length distributions of 22,353 corrected sequences.

TABLE 1 | Summary of reads from PacBio SMRT sequencing.

	Polymerase read	Subreads	CCS	FLNC	Corrected consensus
Number	441,444	14,750,730	381,423	276,295	22,353
Mean length	53,457	1,526	2,396	2,241	2,379
N50	109,539	2,367	2,963	2,725	2,921

TABLE 2 | Comparisons of PacBio and Illumina sequenced data for read mapping.

Terms	PacBio sequenced data		Illumina sequenced data	
	Number of reads	Percentage (%)	Number of reads	Percentage (%)
Total mapped	21,863	97.81	954,571,462	97.27
Multiple mapped	1,583	7.08	23,405,640	2.38
Uniquely mapped	20,280	90.73	931,165,822	94.88
Reads map to "+" strands	10,778	48.22	465,119,568	47.39
Reads map to "−" strands	9,502	42.51	466,046,254	47.49

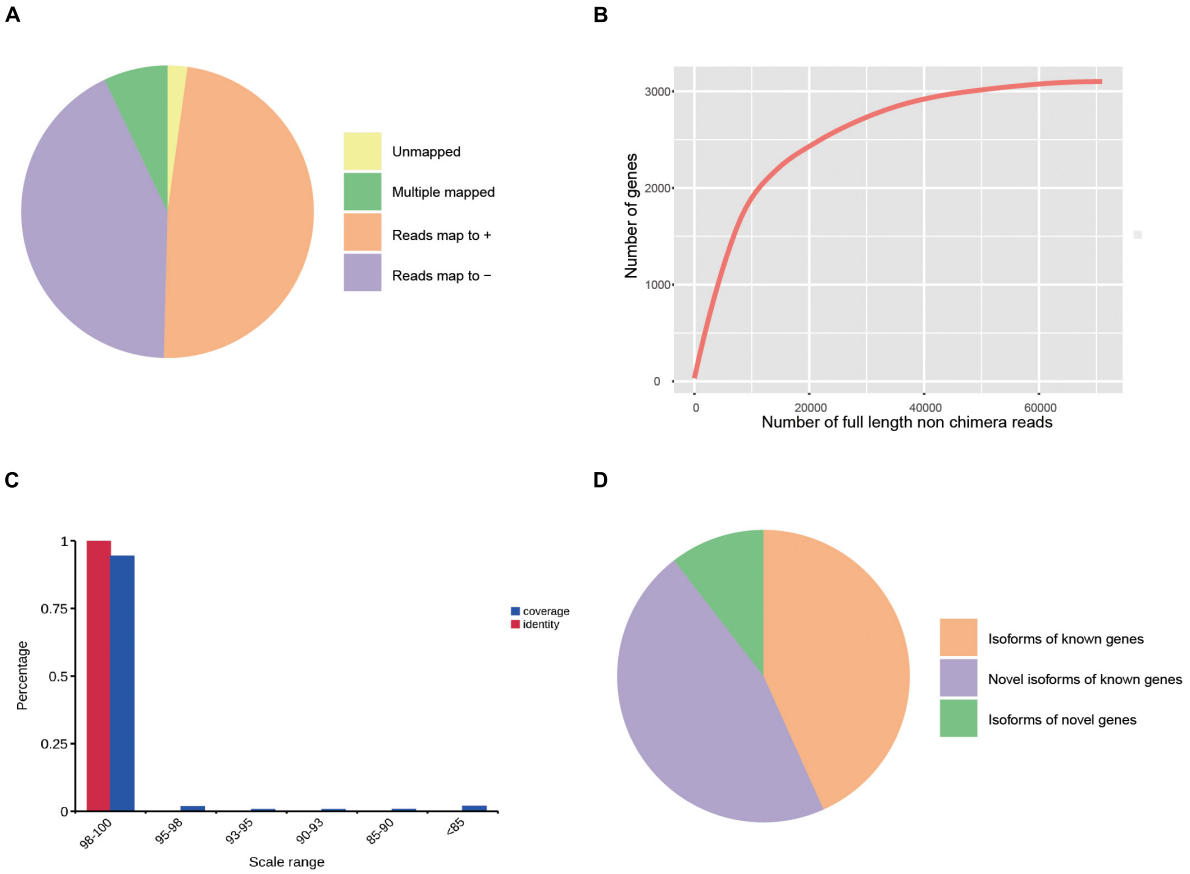


FIGURE 2 | Genome Mapping and Alignment Program (GMAP) analysis of SMRT sequencing. **(A)** GMAP mapping statistics of the corrected sequences. **(B)** Saturation curve of corrected sequences, *x*-axis represents numbers of full-length non-chimera (FLNC) reads, *y*-axis represents numbers of genes. **(C)** Range of mapping coverage and identity, *x*-axis represents the scale ranges, *y*-axis represents the percentages. **(D)** Classification of transcript isoforms identified.

analysis showed the novel genes were assigned to 17 functional clusters, and the "general function prediction only," "translation, ribosomal structure and biogenesis," and "cytoskeleton" ranked as the top three largest categories (**Figure 3D**). The KEGG results demonstrated that the novel genes were mapped to 90 KEGG pathways (**Figure 3E**).

Analysis of AS and APA Events

One of the most important advantages of PacBio sequencing is its ability to identify AS events by directly comparing different isoforms of the same gene. Here, AS events were analyzed with SUPPA software. Seven AS events (IR, SE, Alt.3′, Alt.5′, MX, AF, and AL) were identified. A total of 305 AS events were found based on the PacBio SMRT reads (**Supplementary Table 4**). Two kinds of events, skipped exons (95) and retained introns (83), were much more common than other AS events (**Figure 4A**). PacBio sequencing also enables the investigation of the APA sites. In our study, 3,795 poly(A) sites were identified among the 2,643 genes in the cattle reference genome, 1,929 genes showed 1 poly(A) site, and 21 genes contained at least 5 poly(A) sites (**Figure 4B** and **Supplementary Table 5**). The average number of poly(A) sites per gene was 1.43.

Identification of TF and lncRNA

Transcription factors (TFs) play important regulatory roles in animal growth and development. In this study, they were identified and classified with the animalTFDB 2.0 database. A total of 120 putative TFs from 24 families were identified, of which 11 TFs were identified as novel. The numbers of TFs enriched were as follows: zf-C2H2 (45), ZBTB (14), TF_bZIP (9), bHLH (8), and MYB (6) (**Figure 5** and **Supplementary Table 6**). Based on the prediction of CPC, CNCI, PLEK, and Pfam databases, 2086 transcripts were considered as putative non-coding RNAs. Finally, 588 transcripts found in all 4 prediction results were considered as lncRNAs and 569 (96.8%) of them were novel lncRNAs (**Figure 6A** and **Supplementary Table 7**). Length distribution analysis of the lncRNAs revealed that their lengths ranged from 0.2 to 7.65 kb and the mean length was 1.51 kb (**Figure 6B**). The lncRNAs predicted have fewer exons when compared to the mRNAs and 541 (92%) of the lncRNAs were single exons (**Figure 6C**). Additionally, the identified lncRNAs were further classified into four types, including 205 antisense lncRNAs (34.86%), 171 sense intronic lncRNAs (29.08%), 106 sense overlapping lncRNAs (18.03%), and 106 lincRNAs (18.03%) (**Figure 6D**).

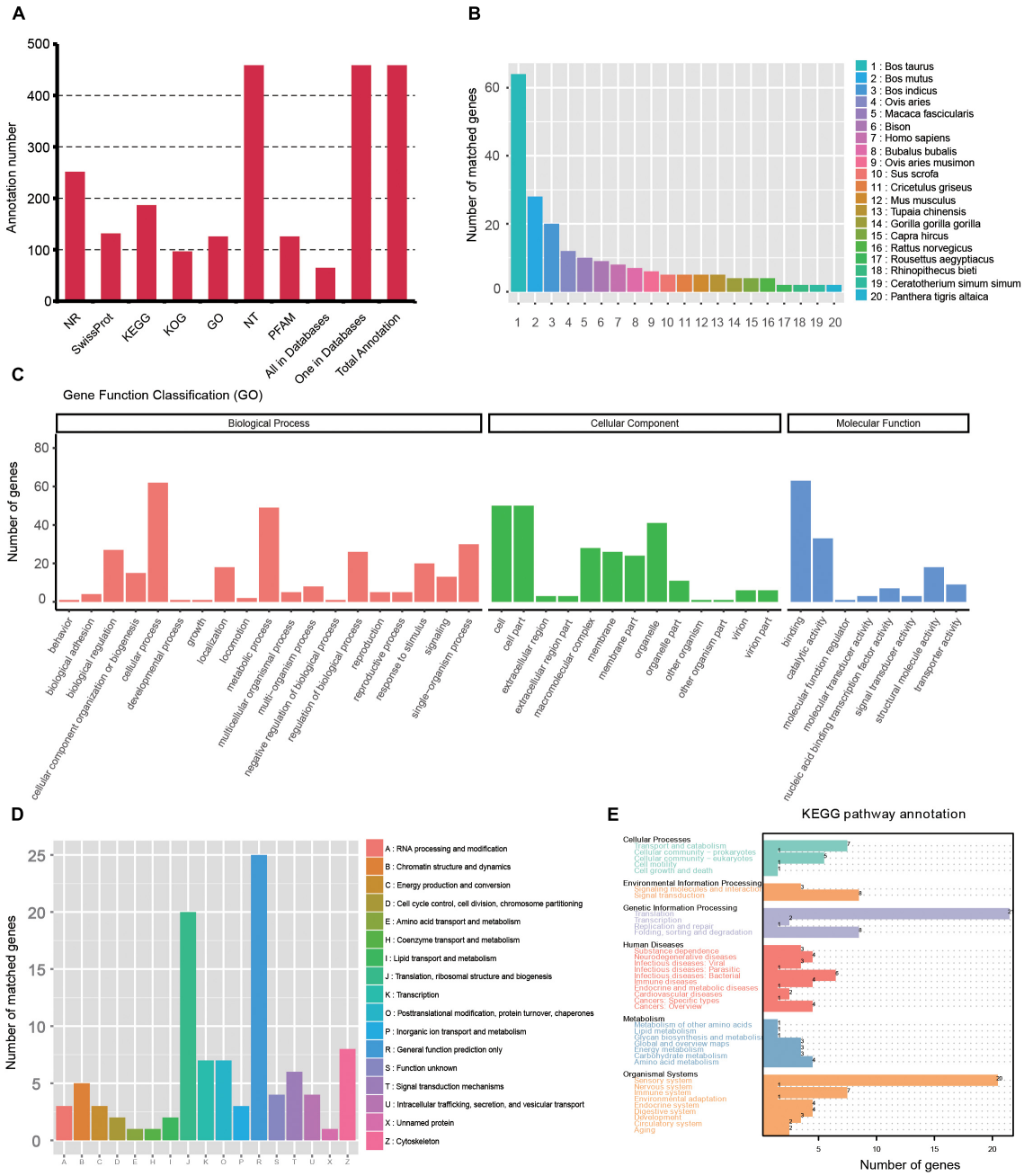


FIGURE 3 | Function annotation of novel genes. **(A)** Function annotation of novel genes in all databases (NR, NT, Pfam, KOG, Swiss-Prot, KEGG, and GO). **(B)** Nr Homologous species distribution diagram of novel genes. **(C)** Distribution of GO terms for all annotated transcripts in biological process, cellular component, and molecular function. **(D)** KOG enrichment of novel genes. **(E)** KEGG pathways enrichment of novel genes.

DISCUSSION

RNA sequencing (RNA-seq) has become a ubiquitous tool for transcriptome-wide analysis of differential gene expression and transcript structure. However, the major limitation of short-read is the difficulty in accurately reconstructing expressed full-length transcripts from the assembly of reads, which are useful for functional studies of important genes (Conesa et al., 2016). Developments in sequencing technology have produced long-read sequencing technology or TGS, which offer many advantages over short-read sequencing and can effectively solve the above limitation. As a representative of long-read sequencing, PacBio SMRT sequencing can capture full-length transcripts without the need for further assembly, which makes it an effective method to analyze full-length sequence, AS events, APA sites, lncRNAs, and gene structure at the transcriptome

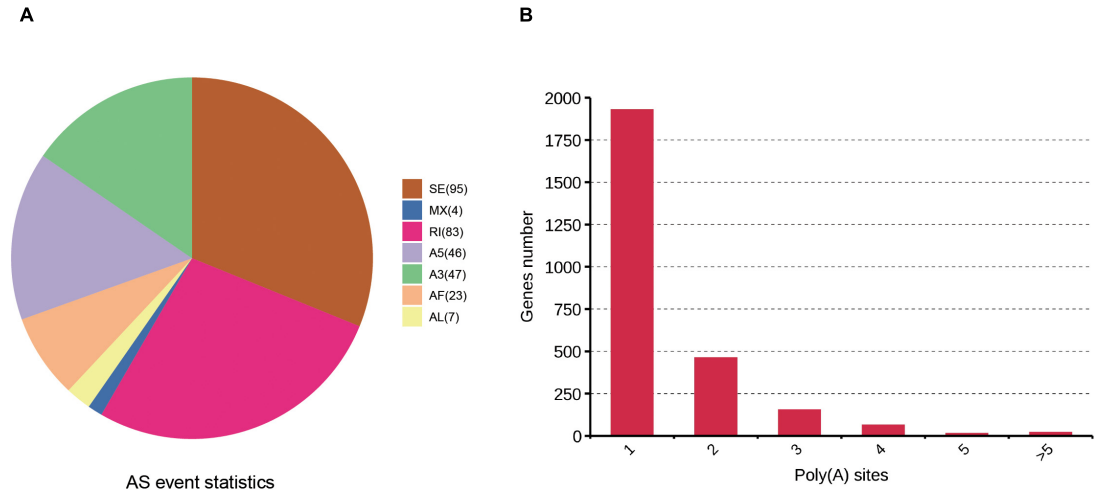


FIGURE 4 | Identification of AS and APA events based on the SMRT sequencing. **(A)** Number and categories of the AS events identified. **(B)** Distribution of the number of poly(A) sites per gene.

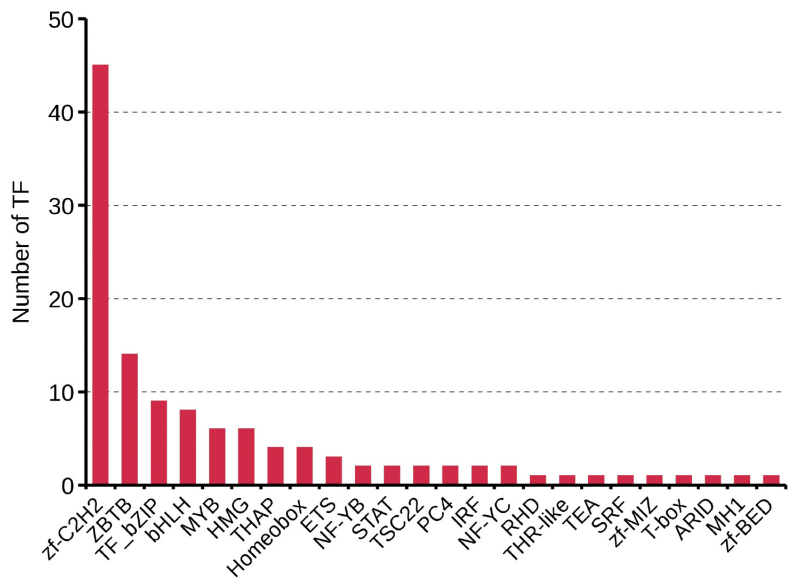


FIGURE 5 | Identification of transcription factors based on the SMRT sequencing.

level (Rhoads and Au, 2015; Feng et al., 2019). Furthermore, with continuing progress in accuracy, throughput, and cost reduction, long-read sequencing has become an option for a broad range of applications in genomics and transcriptomics for model and non-model organisms (Amarasinghe et al., 2020).

In this study, we used PacBio SMRT sequencing to assess the Simmental cattle transcriptome by pooling RNA samples from different tissues together. Then, a total of 22.5 Gb of subreads data were obtained and 381,423 CCSs were generated. By detecting the sequences, 276,295 were identified as full-length non-chimeric (FLNC) reads, which accounts for 72.44% of all CCSs. After removing redundant sequences, the consensus sequences were obtained. Meanwhile, we acquired paired-end reads on the Illumina platform, from which ∼981.4 million clean reads were retained after quality filtering. These short reads were subsequently used for correcting the consensus isoform sequences of SMRT sequencing data. Finally, combined SMRT with Illumina data, 22,353 corrected consensus reads were obtained in total. After mapping the consensus reads against the cattle reference genome, the mapping rate was 97.81%, which shows the high quality of the sequencing data.

We used the TAPIS pipeline to perform transcripts structure analysis. After further correction and clustering to eliminate redundancy, we finally got 4,862 high-quality isoforms, among which 2,104 (43.27%) isoforms were classified as known isoforms from known genes, 2,250 (46.28%) were classified as novel

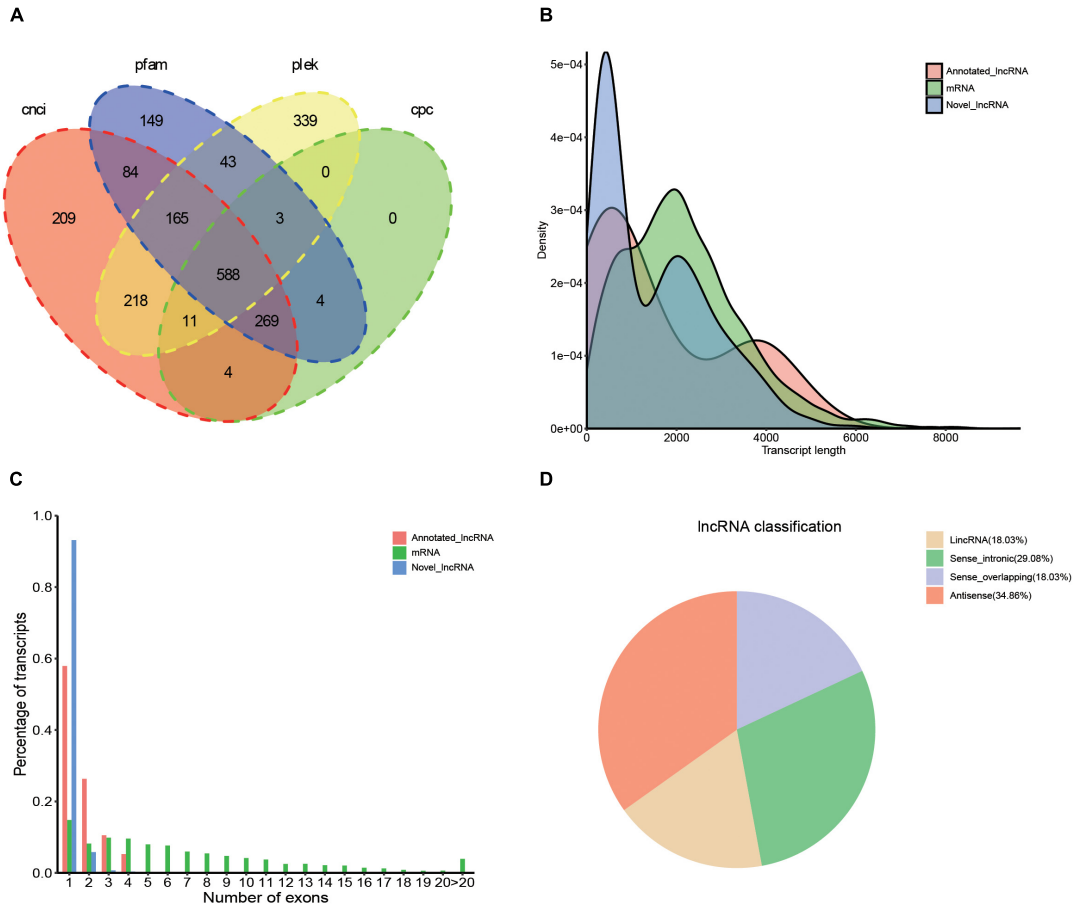


FIGURE 6 | Identification of lncRNA based on the SMRT sequencing. **(A)** Venn diagram of lncRNA predicted by CNCI, CPC, PLEK, and Pfam tools. **(B)** Length and density distribution of annotated lncRNA, novel lncRNA, and mRNA identified. **(C)** Comparison of exon number of annotated lncRNA, mRNA, and novel lncRNA. **(D)** Classification of the types of lncRNA.

isoforms from known genes, and 508 (10.45%) were novel isoforms from novel genes. AS is a crucial transcriptional regulation mechanism for increasing the structural and functional polymorphism of transcripts and proteins. Here, we found 305 AS events using PacBio sequences. The types of most AS were ES and IR. Previous studies have indicated that APA of RNA influenced gene function by changing transcriptome complexity and gene expression. Our study also provides a comprehensive genome-wide APA map draft consisting of 3,795 poly(A) sites from 2,643 genes. These results may underestimate the true number of APA genes because of the low expression of proximal poly(A) sites.

Long non-coding RNAs (LncRNAs) are important regulators of gene expression and are involved in a wide range of biological processes, such as cell proliferation differentiation and modification of chromatin. Several studies have been conducted to identify lncRNAs in cattle, but most of them were performed based on NGS data (Billerey et al., 2014; Koufariotis et al., 2015; Kern et al., 2018). In our study, 569 novel lncRNAs (mean length 1.51 kb) were identified based on PacBio sequencing data. These newly identified lncRNAs will provide additional valid candidates for future functional characterization. Besides, we used the animalTFDB 2.0 database to perform TFs prediction and classification, and then 120 putative TFs from 24 families were identified.

Taken together, our study generated a large number of gene models and alternative isoforms that have not been annotated yet and provide a general encyclopedia of gene transcriptions. These findings refined the annotation of the reference genome and are beneficial for characterizing full-length transcripts of cattle, which are useful for further genetically molecular breeding of cattle. Of course, this profiling of cattle transcriptome would not be exhaustive due to the limited number of sequencing samples.

CONCLUSION

Overall, we analyzed the full-length transcriptome of cattle with PacBio SMRT sequencing. Based on full-length transcripts, many AS events, APA sites, novel isoforms, novel lncRNAs, and TFs provide a more comprehensive foundation to explore cattle transcriptome diversity. Our results may provide valuable

information for improving cattle draft genome annotation, optimizing the genome structure, and fully characterizing the cattle transcriptome.

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: https://bigd.big.ac.cn/, PRJCA004285.

ETHICS STATEMENT

All animals used in the study were treated following the guidelines established by the Council of China Animal Welfare and were handled in strict accordance with good clinical practices to minimize suffering. Protocols of the experiments were approved by the Science Research Department of the Institute of Animal Sciences, Chinese Academy of Agricultural Sciences (approval number: IAS2020-33). Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

JL, HG, and TC designed the study and the interpretation of the findings. LZ, BA, ML, XD, and LD collected the experimental materials and prepared them for the sequencing. TC and BA performed the transcriptome and other computational analyses. WC, BZ, YC, and XG oversaw the project. TC wrote the manuscript. LX revised critically for content and grammar. All authors read and approved the final manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (31872975), the Chinese Academy of Agricultural Sciences of Technology Innovation Project (CAAS-XTCX2016010, CAAS-ZDXT2018006, and ASTIP-IAS03), and the Program of National Beef Cattle, Yak Industrial Technology System (CARS-37) and China Agriculture Research System of MOF and MARA.

ACKNOWLEDGMENTS

We acknowledge the Novogene Corporation (Beijing, China) for the facilities and expertise of the PacBio platform for libraries construction and sequencing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.664974/full#supplementary-material

Supplementary Table 1 | Prediction of novel isoforms from known genes.

Supplementary Table 2 | Prediction of the novel genes.

Supplementary Table 3 | Annotation of new genes based on seven databases.

Supplementary Table 4 | Results of AS identification.

Supplementary Table 5 | Results of poly(A) prediction.

Supplementary Table 6 | Prediction of the transcription factors.

Supplementary Table 7 | Prediction of the novel lncRNAs.

REFERENCES

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., et al. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7:11706. doi: 10.1038/ncomms11706

Alamancos, G. P., Pages, A., Trincado, J. L., Bellora, N., and Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* 21, 1521–1531. doi: 10.1261/rna.051557.115

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21:30. doi: 10.1186/s13059-020-1935-5

Beiki, H., Liu, H., Huang, J., Manchanda, N., Nonneman, D., Smith, T. P. L., et al. (2019). Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. *BMC Genom.* 20:344. doi: 10.1186/s12864-019-5709-y

Billerey, C., Boussaha, M., Esquerre, D., Rebours, E., Djari, A., Meersseman, C., et al. (2014). Identification of large intergenic non-coding RNAs in bovine muscle using next-generation transcriptomic sequencing. *BMC Genom.* 15:499. doi: 10.1186/1471-2164-15-499

Carruthers, M., Yurchenko, A. A., Augley, J. J., Adams, C. E., Herzyk, P., and Elmer, K. R. (2018). De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. *BMC Genom.* 19:32. doi: 10.1186/s12864-017-4379-x

Chaisson, M. J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinform.* 13:238. doi: 10.1186/1471-2105-13-238

Chen, S. Y., Deng, F., Jia, X., Li, C., and Lai, S. J. (2017). A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Sci. Rep.* 7:7648. doi: 10.1038/s41598-017-08138-z

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. doi: 10.1186/s13059-016-0881-8

Crysnanto, D., and Pausch, H. (2020). Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome Biol.* 21:184. doi: 10.1186/s13059-020-02105-0

Du, H., Yu, Y., Ma, Y., Gao, Q., Cao, Y., Chen, Z., et al. (2017). Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.* 8:15324. doi: 10.1038/ncomms15324

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-Time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi: 10.1126/science.1162986

Feng, X., Jia, Y. T., Zhu, R., Chen, K., and Chen, Y. F. (2019). Characterization and analysis of the transcriptome in *Gymnocypris selincuoensis* on the Qinghai-Tibetan Plateau using single-molecule long-read sequencing and RNA-seq. *DNA Res.* 26, 353–363. doi: 10.1093/dnares/dsz014

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi: 10.1093/nar/gkv1344

Fu, S., Wang, A., and Au, K. F. (2019). A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biol.* 20:26. doi: 10.1186/s13059-018-1605-z

Hackl, T., Hedrich, R., Schultz, J., and Forster, F. (2014). Proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* 30, 3004–3011. doi: 10.1093/bioinformatics/btu392

Jia, D., Wang, Y., Liu, Y., Hu, J., Guo, Y., Gao, L., et al. (2018). SMRT sequencing of full-length transcriptome of flea beetle *Agasicles hygrophila* (Selman and Vogt). *Sci. Rep.* 8:2197. doi: 10.1038/s41598-018-20181-y

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280. doi: 10.1093/nar/gkh063

Kern, C., Wang, Y., Chitwood, J., Korf, I., Delany, M., Cheng, H., et al. (2018). Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. *BMC Genom.* 19:684. doi: 10.1186/s12864-018-5037-7

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., et al. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349. doi: 10.1093/nar/gkm391

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., et al. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30, 693–700. doi: 10.1038/nbt.2280

Koufariotis, L. T., Chen, Y. P. P., Chamberlain, A., Jagt, C. V., and Hayes, B. J. (2015). A catalogue of novel bovine long noncoding RNA across 18 tissues. *PLoS One* 10:e0141225. doi: 10.1371/journal.pone.0141225

Lan, P., Li, W., and Schmidt, W. (2012). Complementary proteome and transcriptome profiling in phosphate-deficient *Arabidopsis* roots reveals multiple levels of gene regulation. *Mol. Cell Proteom.* 11, 1156–1166. doi: 10.1074/mcp.M112.020461

Li, A., Zhang, J., and Zhou, Z. (2014). PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinform.* 15:311. doi: 10.1186/1471-2105-15-311

Li, Y., Fang, C., Fu, Y., Hu, A., Li, C., Zou, C., et al. (2018). A survey of transcriptome complexity in *Sus scrofa* using single-molecule long-read sequencing. *DNA Res.* 25, 421–437. doi: 10.1093/dnares/dsy014

Li, Z., Xu, C., Li, K., Yan, S., Qu, X., and Zhang, J. (2012). Phosphate starvation of maize inhibits lateral root formation and alters gene expression in the lateral root primordium zone. *BMC Plant Biol.* 12:89. doi: 10.1186/1471-2229-12-89

Lian, H., Wang, A., Shen, Y., Wang, Q., Zhou, Z., Zhang, R., et al. (2020). Identification of novel alternative splicing isoform biomarkers and their association with overall survival in colorectal cancer. *BMC Gastroenterol.* 20:171. doi: 10.1186/s12876-020-01288-x

Oono, Y., Kawahara, Y., Yazawa, T., Kanamori, H., Kuramata, M., Yamagata, H., et al. (2013). Diversity in the complexity of phosphate starvation transcriptomes among rice cultivars based on RNA-Seq profiles. *Plant Mol. Biol.* 83, 523–537. doi: 10.1007/s11103-013-0106-4

Patel, R. K., and Jain, M. (2012). NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7:e30619. doi: 10.1371/journal.pone.0030619

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genom. Proteom. Bioinform.* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., et al. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* 9:giaa021. doi: 10.1093/gigascience/giaa021

Salmela, L., and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30, 3506–3514. doi: 10.1093/bioinformatics/btu538

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–1014. doi: 10.1038/nbt.2705

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., et al. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 41:e166. doi: 10.1093/nar/gkt646

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36. doi: 10.1093/nar/28.1.33

Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., et al. (2015). Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* 33, 736–742. doi: 10.1038/nbt.3242

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., et al. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7:11708. doi: 10.1038/ncomms11708

Wang, L., Jiang, X., Wang, L., Wang, W., Fu, C., Yan, X., et al. (2019). A survey of transcriptome complexity using PacBio single-molecule real-time analysis combined with Illumina RNA sequencing for a better understanding of ricinoleic acid biosynthesis in *Ricinus communis*. *BMC Genom.* 20:456. doi: 10.1186/s12864-019-5832-9

Wang, X., Chen, S., Shi, X., Liu, D., Zhao, P., Lu, Y., et al. (2019). Hybrid sequencing reveals insight into heat sensing and signaling of bread wheat. *Plant J.* 98, 1015–1032. doi: 10.1111/tpj.14299

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X. J., et al. (2017). Comprehensive comparison of Pacific biosciences and Oxford Nanopore technologies and their applications to transcriptome analysis. *F1000Research* 6:100. doi: 10.12688/f1000research.10571.2

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875. doi: 10.1093/bioinformatics/bti310

Xu, Z., Peters, R. J., Weirather, J., Luo, H., Liao, B., Zhang, X., et al. (2015). Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis. *Plant J.* 82, 951–961. doi: 10.1111/tpj.12865

Zhang, H. M., Liu, T., Liu, C. J., Song, S., Zhang, X., Liu, W., et al. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.* 43, D76–D81. doi: 10.1093/nar/gku887

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., et al. (2009). A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* 10:R42. doi: 10.1186/gb-2009-10-4-r42

Zuo, C., Blow, M., Sreedasyam, A., Kuo, R. C., Ramamoorthy, G. K., Torres-Jerez, I., et al. (2018). Revealing the transcriptomic complexity of switchgrass by PacBio long-read sequencing. *Biotechnol. Biofuels* 11:170. doi: 10.1186/s13068-018-1167-z

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 Chang, An, Liang, Duan, Du, Cai, Zhu, Gao, Chen, Xu, Zhang, Gao 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership