frontiers | Research Topics

Forensic investigative genetic genealogy and fine-scale structure of human populations, 2nd edition

Edited by
Guanglin He, Mengge Wang and Ryan Lan-Hai Wei

Published in
Frontiers in Genetics
Frontiers in Ecology and Evolution

FRONTIERS EBOOK COPYRIGHT STATEMENT

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

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

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

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

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

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

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

ISSN 1664-8714
ISBN 978-2-8325-7432-4
DOI 10.3389/978-2-8325-7432-4

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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

Forensic investigative genetic genealogy and fine-scale structure of human populations, 2nd edition

Topic editors
Guanglin He — Sichuan University, China
Mengge Wang — Sun Yat-sen University, China
Ryan Lan-Hai Wei — Inner Mongolia Normal University, China

Citation
He, G., Wang, M., Wei, R. L.-H., eds. (2026). *Forensic investigative genetic genealogy and fine-scale structure of human populations, 2nd edition*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-7432-4

Publisher's note: This is a 2nd edition due to an article retraction.

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 05 January 2023
DOI 10.3389/fgene.2022.1067865

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Ronny Decorte,
Faculty of Medicine, KU Leuven, Belgium

*CORRESPONDENCE
He Guanglin,
Guanglinhescu@163.com
Wang Mengge,
Menggewang2021@163.com

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 12 October 2022
ACCEPTED 22 December 2022
PUBLISHED 05 January 2023

CITATION
Guanglin H, Lan-Hai W and Mengge W (2023), Editorial: Forensic investigative genetic genealogy and fine-scale structure of human populations.
Front. Genet. 13:1067865.
doi: 10.3389/fgene.2022.1067865

COPYRIGHT
© 2023 Guanglin, Lan-Hai and Mengge. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Forensic investigative genetic genealogy and fine-scale structure of human populations

He Guanglin[1]*, Wei Lan-Hai[2] and Wang Mengge[3]*

[1]Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu, China, [2]School of Ethnology and Anthropology, Inner Mongolia Normal University, Hohhot, China, [3]Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China

KEYWORDS

population structure, admixture history, genome-wide genetic markers, forensic investigative genetic genealogy, genetic diversity

Editorial on the Research Topic
Forensic investigative genetic genealogy and fine-scale structure of human populations

Introduction

Anatomically modern humans originated in Africa and separated from their most likely recent common ancestor hundreds and thousands of years ago (Bergstrom et al., 2020; Wang et al., 2021a). They followingly migrated out of Africa around 50 thousand years ago and evolved in concert with the complicated interplay of gene flow and adaptive selection during the peopling of Eurasia, Oceania, and America (Bergstrom, et al., 2020; Wang et al., 2021a). Genomic studies have demonstrated the pervasiveness of population differentiation and genetic admixture between long-isolated ethnic groups (Bergstrom et al., 2020; Pan et al., 2022). Extensive population bottleneck, adaptive evolution in changing environments, and introgression from archaic hominins further shaped the complicated patterns of human genetic heritage. In general, complex population divergence, migration, and admixture events extensively shaped the patterns of genetic diversity of ethnolinguistically diverse populations.

There is increasing evidence to suggest that the differences in the susceptibility of many common and rare diseases are primarily attributed to human populations' diverse cultural, environmental, demographic, and genetic histories. The comprehensive understanding of fine-scale population evolutionary history will gradually change our understanding of the genetic architecture of diseases (Timpson et al., 2018; Benton et al., 2021; Pan et al., 2022). Thus, there is an urgent need to expand genetic research to populations with different ancestries. In addition, population genetic studies based on multiple genome-wide genetic markers (short tandem repeat, STR; single nucleotide polymorphism, SNP; Insertion/Deletion, InDel; copy number variation, CNV, and so on) could provide new insights into the detailed process of population admixture and evolutionary history of ethnolinguistically and geographically diverse populations. Understanding the fine-scale population structure also helps the better study design in medical genomics and the comprehensive practice applications in population genetics and forensic science.

With the rapid development of genotyping technologies, sequencing platforms, and computational methods, previous studies have provided the basal framework of the genetic

landscape of worldwide populations from different perspectives (Li et al., 2008; Lippold et al., 2014). However, most genetic studies focused on the relationships and fine-scale population structures were conducted *via* low-density genetic markers. It is also now possible to capture and sequence ancient DNA from ancient samples, which could provide pivotal insights into the formation of spatiotemporally diverse populations with unprecedented resolution. Although huge nationwide biobanks for characterizing the genotypes and phenotypes of millions of people have been established (Barton et al., 2021; Zhang et al., 2021; Chiu et al., 2022), more geographically, linguistically, and culturally diverse populations (especially non-metropolitan populations) are needed to be studied systematically at different spatio-temporal scales.

A clear understanding of genetic background and diversity of ethnolinguistically diverse populations and decoding their demographic history can provide new medical and forensic application opportunities. Genetic studies have illuminated the population-specific reference database, effective algorithm and the developed panel for the targeted forensic applications were the fundamentals of forensic intelligence inference of external visual appearance, biogeographical ancestry inference and forensic investigative genetic genealogy (FIGG) (He et al., 2018). FIGG, one new and rapidly growing field of forensic genetics since 2018, has attracted the attention of geneticists focused on complex familial search (Phillips 2018). Currently, many projects aim to develop and validate new FIGG panels, construct and complement forensic FIGG databases, and develop new statistical models to promote the practice of FIGG (Kling et al., 2021; Tillmar et al., 2021).

To summarize the new advances in FIGG and fine-scale population structure and illuminate the importance of full-scale genetic structure and diversity as the basis for the FIGG, we organized this Research Topic of the "*Forensic investigative genetic genealogy and fine-scale structure of Human Populations*". In detail, this Research Topic aimed to characterize the genetic background and demographic history of ethnolinguistically and geographically diverse populations based on different densities of genetic marker. It would be helpful in exploring the long-range familial searches and fine-scale genetic localization within subgroups in other continents. This Research Topic attracted research focused on the basic knowledge exploration of the genetic background of one targeted population and included applied research focused on developing and validating forensic amplification systems.

Exploration of theoretical knowledge—Fine-scale genetic structure and demographic history reconstruction

Recent studies have demonstrated that population history reconstruction leveraging high-density genetic markers could uncover previously unrecognized population structures at a fine scale compared with forensically relevant loci (Li, et al., 2008; Bergstrom, et al., 2020). Zhou et al. generated and analyzed genome-wide data of Liaoning Han people and found that genetic differences existed in geographically different Sinitic-speaking Han populations, which might result from other migration and admixture events of Hans during the period of "Chuang Guandong". Hou et al. investigated the population history of Liaoning Mongolians based on ~700,000 SNPs and provided new insights into the admixture history of Mongolic-speaking Mongolians according to shared allele-based analyses. He et al. explored the demographic history of Qiang people based on Eurasian modern and ancient reference populations. This study revealed that the Tibeto-Burman-speaking Qiang people derived their primary ancestry from Tibetan-related ancestral populations in North China. Wang et al. performed a genome-wide association study on 26,806 Chinese individuals. They identified 21 SNPs associated with widow's peak, unibrow, double eyelid, earlobe attachment, and freckles. This study may facilitate a better understanding of the genetic basis of facial development in Chinese populations and provide new markers for forensic phenotype predictions. These studied Han, Mongolian and Qiang people are widely distributed in North China, and other populations from southern China and surrounding regions need to be further explored based on the array or sequencing data, such as Austronesian, Austroasiatic, Hmong-Mien and Tai-Kadai people. Generally, population genetic studies showed a strong correlation between Chinese cultural language and geographic patterns and population structure.

The estimated patterns of genetic diversity in China can help accurately biogeographic ancestry inference and FIGG. Similar population stratifications were also identified in Siberian populations. Currently, available tools could effectively distinguish populations with different continental origins, but most of these are not efficient for differentiating people within the same continent. Gorin et al. selected 5,229 AISNPs and tested various mathematical models for biogeographic ancestry inference. The results showed that the accuracy of the prediction of this panel on one of 29 studied ethnic groups reached 71% and the proposed method could be employed to predict ancestries from Russian and neighboring populations. Huang et al. developed a machine learning approach for estimating the relationships with high error SNP profiles and found that this approach was more accurate and robust than the individual measures.

Forensic potential applications—Development and validation of forensic systems focused on personal identification and family research

STR genotyping has been applied in forensic investigations for nearly 30 years (Hagelberg et al., 1991; Kayser and de Knijff 2011). Nowadays, several commercial STR kits have been developed based on the expanded CODIS loci (Oostdik et al., 2014; Wang et al., 2018; Qu et al., 2019; Batham et al., 2020; Green et al., 2021). However, with the rapid increase in the number of STR genotypes in forensic databases, more novel non-CODIS STRs with high genetic polymorphisms are required to minimize the incidence of adventitious matches. Huang et al. validated the forensic performance of a novel multiplex autosomal STR panel (including six CODIS STRs and 20 non-CODIS STRs). They found that this novel kit could be applied as a promising tool for forensic human identification and complex paternity analysis. Than et al. genotyped seven Lao Isan and three Laotian populations using Verifiler plus PCR Amplification kit. The allelic frequency results provided the genetic background of Austroasiatic and Tai-Kadai people from Laos and Thailand.

InDel loci, the second most abundant polymorphism across the human genome, possess low mutation rates and small amplicon lengths compared with STRs (Weber et al., 2002; Mills et al.,

2006), which have been proven to be of value in forensic investigations (Pereira et al., 2009; Zhang et al., 2018). And InDels showing significant allele frequency differences among geographically and linguistically diverse populations can be adopted as ancestry-informative markers (AIMs) (Sun et al., 2016; Inacio et al., 2017). Moreover, previous studies showed that multi-InDel loci behaved well in parentage tests and could be used for forensic applications (Fan et al., 2016; Sun, et al., 2016; Qu et al., 2020). Liu et al. developed and validated a six-color fluorescence multiplex panel including 59 autosomal InDels. Subsequently, Tibetan groups from China have been genotyped using the newly-developed 59-plex InDel panel. The comprehensive population genetic analyses showed that this homemade panel could be used as a powerful tool for individual forensic identification and paternity testing in Chinese Tibetan groups. Jin et al. developed a Next-Generation Sequencing (NGS) InDel panel, including 17 multi-InDels on the X chromosome. They found that the newly-developed panel could be adopted as an effective tool for individual forensic identification, paternity testing, and biogeographical ancestry inference.

Genetic surveys based on uniparentally inherited markers have identified many paternal and maternal founding lineages in regional populations and their corresponding expansion events (Poznik et al., 2016; Li et al., 2019). Jia et al. sequenced complete mitochondrial genomes of 146 Daur individuals in China. The results showed that the Daur ethnic group has high maternal genetic diversity and may have experienced recent population expansion. He et al. developed and validated the AGCU-Y30 Y-STR panel and conducted a Y-STR-based study to explore the paternal history of the Qiang people. The validated results showed that the novel Y-STR kit was sensitive and robust enough for forensic applications. Population genetic analyses revealed that the Qiang people are closely related to lowland Tibetan-Yi corridor populations.

Prospects and challenges

Human genetics needed the fundamental foundation supporting large-scale population genomic projects to characterize the full landscape of human genetic diversity and population structure, such as NHLBI TOPMed (Taliun et al., 2021), gnomAD (Collins et al., 2020) and UK10K (Wang et al., 2021). These projects made significant advances in European human genetics. Other projects, including the Chinese 10K_CPGDP (Chinese population genomic diversity project), GSRD-100K^{WCH} and ChinaMap, were recently launched to explore the genetic features of under-represented populations. Characterizing the genetic architecture of ethnolinguistically and geographically diverse populations will promote our understanding of population origin, separation, admixture, adaptation, and gene flow from archaic individuals.

Knowledge of the fine-scale genetic structure and population-specific genomic database is the base for FIGG. Synthesizing our growing knowledge of evolutionary history with forensic investigations will help to achieve the promise of long-range familial searches and fine-scale genetic localization. For human population genetics and FIGG, we also had some challenges that needed to be overcome in the next step:

Firstly, European bias in worldwide human genetic studies and Han Chinese bias in Chinese cohort research hinder in heath equality of genetic studies and forensic genetics. More population genomic studies from under-reported populations need to be conducted to characterize their uncharacterized genetic polymorphisms and the genetic spectrum of different genetic markers.

Second, a population-specific genomic database should be constructed to provide a comprehensive landscape of different genetic markers and even include structural variations and mobile elements via the PacBio sequencing plateau.

Third, cooperation of national genomic studies should be formed to promote data sharing in different institutions.

Forth, population genomic projects should be conducted among included subjects with deep phenotypes, which would provide values for exploring the genetic basis for physical traits and medical phenotypes.

Fifth, regional population-specific genomic datasets, FIGG panels, algorithms, and forensic databases should be developed and validated. FIGG has promoted successful inspection of criminal suspects among American or European populations, as there are publicly available European/American genomic databases and DNA. Land and GEDmatch severs. Similar databases from other countries or regions should be constructed in the future.

Seventh, IBD-based algorithms need the high-density phased SNPs or whole genome sequencing data. Many forensic samples were highly degraded or had minute amounts of genetic materials. Thus, the primary focus is to develop more FIGG panels based on low-density SNP markers and non-linkage algorithms (allele-sharing status and allele frequency spectrum).

In short, more ethnolinguistically and geographically diverse populations are needed to be studied based on different genetic markers and algorithm models, and population-specific panels based on different genetic variations and corresponding forensic databases need to be developed to achieve long-range familial searches and fine-scale genetic structure reconstruction.

Author contributions

MW and GH drafted it, and MW, L-HW, and GH proofread it. Before submitting it, all authors reviewed it.

Funding

This work was funded by the National Natural Science Foundation of China (82202078).

Acknowledgments

Thanks to all the brilliant Frontiers team whose invaluable assistance at each step made the Research Topic successful. We also thank Ronny Decorte from the Faculty of Medicine, KU Leuven, who gave the essential revised advice for our work.

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.

References

Barton, A. R., Sherman, M. A., Mukamel, R. E., and Loh, P. R. (2021). Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet. Aug* 53, 1260–1269. Epub 2021/07/07. doi:10.1038/s41588-021-00892-1

Batham, M. S., Kushwaha, K. P. S., Chauhan, T., Kumawat, R. K., and Shrivastava, P. (2020). Autosomal STR allele frequencies in Kahars of Uttar Pradesh, India, drawn with PowerPlex® 21 multiplex system. *Int. J. Leg. Med. Mar.* 134, 517–519. Epub 2019/03/29. doi:10.1007/s00414-019-02046-9

Benton, M. L., Abraham, A., LaBella, A. L., Abbot, P., Rokas, A., and Capra, J. A. (2021). The influence of evolutionary history on human health and disease. *Nat. Rev. Genet. May* 22, 269–283. Epub 20210106. doi:10.1038/s41576-020-00305-9

Bergstrom, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Sci. Mar* 20, eaay5012. Epub 2020/03/21. doi:10.1126/science.aay5012

Chiu, A. M., Molloy, E. K., Tan, Z., Talwalkar, A., and Sankararaman, S. (2022). Inferring population structure in biobank-scale genomic data. *Am. J. Hum. Genet. Apr* 7 (109), 727–737. Epub 2022/03/18. doi:10.1016/j.ajhg.2022.02.015

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alfoldi, J., Francioli, L. C., et al. (2020). A structural variation reference for medical and population genetics. *Nat. May* 581, 444–451. Epub 20200527. doi:10.1038/s41586-020-2287-8

Fan, G. Y., Ye, Y., and Hou, Y. P. (2016). Detecting a hierarchical genetic population structure via Multi-InDel markers on the X chromosome. *Sci. Rep.* 6 (6), 32178. Epub 2016/08/19. doi:10.1038/srep32178

Green, R., Elliott, J. L., Norona, W., Go, F., Nguyen, V. T., Ge, J., et al. (2021). Developmental validation of VeriFiler™ plus PCR amplification kit: A 6-dye multiplex assay designed for casework samples. *Forensic Sci. Int. Genet. Jul* 53, 102494. Epub 2021/03/20. doi:10.1016/j.fsigen.2021.102494

Hagelberg, E., Gray, I. C., and Jeffreys, A. J. (1991). Identification of the skeletal remains of a murder victim by DNA analysis. *Nature* 352, 427–429. doi:10.1038/352427a0

He, G., Wang, Z., Wang, M., Luo, T., Liu, J., Zhou, Y., et al. (2018). Forensic ancestry analysis in two Chinese minority populations using massively parallel sequencing of 165 ancestry-informative SNPs. *Electrophor. Nov.* 39, 2732–2742. Epub 20180628. doi:10.1002/elps.201800019

Inacio, A., Costa, H. A., da Silva, C. V., Ribeiro, T., Porto, M. J., Santos, J. C., et al. (2017). Study of InDel genetic markers with forensic and ancestry informative interest in PALOP's immigrant populations in Lisboa. *Int. J. Leg. Med. May* 131, 657–660. Epub 2016/11/01. doi:10.1007/s00414-016-1484-3

Kayser, M., and de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nat. Rev. Genet. Mar.* 12, 179–192. Epub 2011/02/19. doi:10.1038/nrg2952

Kling, D., Phillips, C., Kennett, D., and Tillmar, A. (2021). Investigative genetic genealogy: Current methods, knowledge and practice. *Forensic Sci. Int. Genet. May* 52, 113–124. Epub 2021/02/17. doi:10.1016/j.fsigen.2019.06.019

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Sci. Feb* 22, 3191100–3191104. Epub 2008/02/23. doi:10.1126/science.1153717

Li, Y. C., Ye, W. J., Jiang, C. G., Zeng, Z., Tian, J. Y., Yang, L. Q., et al. (2019). River valleys shaped the maternal genetic landscape of han Chinese. *Mol. Biol. Evol.* 36, 1643–1652. Epub 2019/05/22. doi:10.1093/molbev/msz072

Lippold, S., Xu, H., Ko, A., Li, M., Renaud, G., Butthof, A., et al. (2014). Human paternal and maternal demographic histories: Insights from high-resolution Y chromosome and mtDNA sequences. *Investig. Genet.* 5, 13. Epub 20140924. doi:10.1186/2041-2223-5-13

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., et al. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res. Sep.* 16, 1182–1190. Epub 20060810. doi:10.1101/gr.4565806

Oostdik, K., Lenz, K., Nye, J., Schelling, K., Yet, D., et al. (2014). Developmental validation of the PowerPlex(®) fusion system for analysis of casework and reference samples: A 24-locus multiplex for new database standards. *Forensic Sci. Int. Genet. Sep.* 12, 69–76. Epub 2014/06/07. doi:10.1016/j.fsigen.2014.04.013

Pan, Y., Zhang, C., Lu, Y., Ning, Z., Lu, D., Gao, Y., et al. (2022). Genomic diversity and post-admixture adaptation in the Uyghurs. *Natl. Sci. Rev. Mar.* 9, nwab124. nwab124. Epub 20210911. doi:10.1093/nsr/nwab124

Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, A., and Gusmao, L. (2009). A new multiplex for human identification using insertion/deletion polymorphisms. *Electrophor. Nov.* 30, 3682–3690. Epub 2009/10/29. doi:10.1002/elps.200900274

Phillips, C. (2018). The Golden State Killer investigation and the nascent field of forensic genealogy. *Forensic Sci. Int. Genet. Sep.* 36, 186–188. Epub 2018/07/25. doi:10.1016/j.fsigen.2018.07.010

Poznik, G. D., Xue, Y., Mendez, F. L., Willems, T. F., Massaia, A., Wilson Sayres, M. A., et al. (2016). Punctuated bursts in human male demography inferred from 1, 244 worldwide Y-chromosome sequences. *Nat. Genet. Jun* 48, 593–599. Epub 2016/04/26. doi:10.1038/ng.3559

Qu, S., Li, H., Li, Y., Lv, M., Yang, F., Zhu, J., et al. (2019). Developmental validation of the Microreader™ 20A ID system. *Electrophoresis* 40, 3099–3107. Epub 2019/10/10. doi:10.1002/elps.201900221

Qu, S., Lv, M., Xue, J., Zhu, J., Wang, L., Jian, H., et al. (2020). Multi-indel: A microhaplotype marker can Be typed using capillary electrophoresis platforms. *Front. Genet.* 11, 567082. Epub 2020/11/17. doi:10.3389/fgene.2020.567082

Sun, K., Ye, Y., Luo, T., and Hou, Y. (2016). Multi-InDel analysis for ancestry inference of sub-populations in China. *Sci. Rep.* 6 (6), 39797. Epub 2016/12/23. doi:10.1038/srep39797

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., et al. (2021). Sequencing of 53, 831 diverse genomes from the NHLBI TOPMed Program. *Nat. Feb* 590, 290–299. Epub 20210210. doi:10.1038/s41586-021-03205-y

Tillmar, A., Sturk-Andreaggi, K., Daniels-Higginbotham, J., Thomas, J. T., and Marshall, C. (2021). The FORCE panel: An all-in-one SNP marker set for confirming investigative genetic genealogy leads and for general forensic applications. *Genes. (Basel)* 12, 1968. Epub 20211210. doi:10.3390/genes12121968

Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J., and Richards, J. B. (2018). Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet. Feb* 19, 110–124. Epub 20171211. doi:10.1038/nrg.2017.101

Wang, C. C., Yeh, H. Y., Popov, A. N., Zhang, H. Q., Matsumura, H., Sirak, K., et al. (2021a). Genomic insights into the formation of human populations in East Asia. *Nat. Mar.* 591, 413–419. Epub 2021/02/23. doi:10.1038/s41586-021-03336-2

Wang, M., Wang, Z., He, G., Jia, Z., Liu, J., and Hou, Y. (2018). Genetic characteristics and phylogenetic analysis of three Chinese ethnic groups using the Huaxia Platinum System. *Sci. Rep. Feb* 5 (8), 2429. Epub 2018/02/07. doi:10.1038/s41598-018-20871-7

Wang, Q., Dhindsa, R. S., Carss, K., Harper, A. R., Nag, A., Tachmazidou, I., et al. (2021b). Rare variant contribution to human disease in 281, 104 UK Biobank exomes. *Nat. Sep.* 597, 527–532. Epub 20210810. doi:10.1038/s41586-021-03855-y

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., and Marth, G. (2002). Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet. Oct.* 71, 854–862. Epub 20020904. doi:10.1086/342727

Zhang, P., Luo, H., Li, Y., Wang, Y., Wang, J., Zheng, Y., et al. (2021). NyuWa genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell. Rep. Nov.* 16, 37110017. Epub 2021/11/18. doi:10.1016/j.celrep.2021.110017

Zhang, S., Zhu, Q., Chen, X., Zhao, Y., Zhao, X., Yang, Y., et al. (2018). Forensic applicability of multi-allelic InDels with mononucleotide homopolymer structures. *Electrophoresis* 39, 2136–2143. Epub 20180710. doi:10.1002/elps.201700468

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 16 May 2022
doi: 10.3389/fgene.2022.902309

Determining the Area of Ancestral Origin for Individuals From North Eurasia Based on 5,229 SNP Markers

Igor Gorin[1,2,3†], Oleg Balanovsky[1,3,4†], Oleg Kozlov[1], Sergey Koshel[3,5], Elena Kostryukova[6], Maxat Zhabagin[7], Anastasiya Agdzhoyan[1,3], Vladimir Pylev[3,4] and Elena Balanovska[1,3,4]*

[1]Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia, [2]Moscow Institute of Physics and Technology, Dolgoprudny, Russia, [3]Research Centre for Medical Genetics, Moscow, Russia, [4]Biobank of North Eurasia, Moscow, Russia, [5]Faculty of Geography, Lomonosov Moscow State University, Moscow, Russia, [6]Federal Research and Clinical Center of Physical-Chemical Medicine, Moscow, Russia, [7]National Center for Biotechnology, Nur-Sultan, Kazakhstan

OPEN ACCESS

Edited by:
Mengge Wang,
Sichuan University, China

Reviewed by:
Ma Pengcheng,
Jilin University, China
Jiang Huang,
Guizhou Medical University, China

***Correspondence:**
Igor Gorin
gorin.io@phystech.edu

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

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 23 March 2022
Accepted: 26 April 2022
Published: 16 May 2022

Citation:
Gorin I, Balanovsky O, Kozlov O, Koshel S, Kostryukova E, Zhabagin M, Agdzhoyan A, Pylev V and Balanovska E (2022) Determining the Area of Ancestral Origin for Individuals From North Eurasia Based on 5,229 SNP Markers. Front. Genet. 13:902309. doi: 10.3389/fgene.2022.902309

Currently available genetic tools effectively distinguish between different continental origins. However, North Eurasia, which constitutes one-third of the world's largest continent, remains severely underrepresented. The dataset used in this study represents 266 populations from 12 North Eurasian countries, including most of the ethnic diversity across Russia's vast territory. A total of 1,883 samples were genotyped using the Illumina Infinium Omni5Exome-4 v1.3 BeadChip. Three principal components were computed for the entire dataset using three iterations for outlier removal. It allowed the merging of 266 populations into larger groups while maintaining intragroup homogeneity, so 29 ethnic geographic groups were formed that were genetically distinguishable enough to trace individual ancestry. Several feature selection methods, including the random forest algorithm, were tested to estimate the number of genetic markers needed to differentiate between the groups; 5,229 ancestry-informative SNPs were selected. We tested various classifiers supporting multiple classes and output values for each class that could be interpreted as probabilities. The logistic regression was chosen as the best mathematical model for predicting ancestral populations. The machine learning algorithm for inferring an ancestral ethnic geographic group was implemented in the original software "Homeland" fitted with the interface module, the prediction module, and the cartographic module. Examples of geographic maps showing the likelihood of geographic ancestry for individuals from different regions of North Eurasia are provided. Validating methods show that the highest number of ethnic geographic group predictions with almost absolute accuracy and sensitivity was observed for South and Central Siberia, Far East, and Kamchatka. The total accuracy of prediction of one of 29 ethnic geographic groups reached 71%. The proposed method can be employed to predict ancestries from the populations of Russia and its neighbor states. It can be used for the needs of forensic science and genetic genealogy.

Keywords: gene geography, ancestry prediction, human population genetics, ancestral origin, machine learning

INTRODUCTION

Now and then, criminal investigators are faced with the need to infer the ancestral geographical origin of an individual from their genotype. Advances in genome analysis technologies and customization of genotyping arrays have shaped the diversity of currently available platforms for biogeographical ancestry prediction from individual DNA samples. Some of them rely on only dozens or hundreds of SNPs and can predict the continent of a person's origin (or a large region at best) rather than a specific population (Mehta et al., 2017; Lan et al., 2019; Pakstis et al., 2019). Such platforms are in high demand in countries where individuals of different continental or subcontinental origins constitute the population majority. They are designed to account for human genetic variation at the global rather than local level, even at the cost of sacrificing the number of informative ancestry markers (Phillips et al., 2019). Other arrays can generate more specific predictions, but the markers they use are geographically limited to large regions or subcontinents, like East or South Asia, Oceania, North Africa, Middle East, and Europe (Al-Asfi et al., 2018; Pereira et al., 2019; Lan et al., 2020; Xavier et al., 2020). One of such panels featuring 48 SNPs has proved to be powerful enough to successfully differentiate between three Chinese populations with very different ancestries: Mongol, Uighur, and Han (Jin et al., 2019). However, it is unclear whether the same set of markers can accurately predict the 40 remaining East Asian Chinese populations.

Commercial arrays for genealogy tracing comprise hundreds of thousands of SNPs and produce accurate results, but high costs preclude their use in routine forensic practice, which is limited to dozens or hundreds of SNPs.

Although the arsenal of tools for ethnic geographic ancestry prediction is continuously expanding and more regions are getting covered, one-third of the world's largest continent remains severely underrepresented. The population of North Eurasia, which spans, among other states, post-Soviet countries, and Mongolia, is incredibly culturally diverse (200 peoples and ten language families) and highly genetically heterogeneous. The immenseness of its genome-wide variation was clearly visible on principal component plots for worldwide population datasets in the early days of SNP-based biogeographic ancestry studies (Li et al., 2008). Using a Humans Origins array featuring 600,000 autosomal markers, Jeong et al. (2019) demonstrated that the composition of the North Eurasian gene pool had been shaped by three major genetic components geographically linked to three ecoregions: forest-tundra, forest-steppe, and steppe. Notably, patterns revealed more than 50 years ago by research studies that relied on classic markers are reproduced today with genome-wide SNP arrays (Balanovskaia and Rychkov, 1990a, Balanovskaia and Rychkov, 1990b; Rychkov and Balanovska, 1992). According to the cited studies, genetic markers characterizing the North Eurasian gene pool occur at different frequencies across populations of North Eurasia. The populations of neighboring regions may not necessarily share them. So, commercial arrays for indigenous ancestry prediction based on dozens of SNPs will provide only rough estimates of European and Asian genetic components for Russian individuals, which is not enough for practical work.

There were attempts to describe the populations of Russia using autosomal STRs and to create an STR-based database for forensic needs. However, the array turned out to have only limited ability to predict ethnic geographic ancestry. The largest dataset representing this region was published in (Stepanov et al., 2011). It consisted of 1,156 samples from 17 populations genotyped for 15 autosomal STRs (Promega PowerPlex16 kit). The dataset represented six Russian cities, nine ethnic groups from Russia, and populations from two other North Eurasian countries (Ukraine and Belarus). The urban populations were shown to be virtually indistinguishable genetically, whereas many ethnic populations differed significantly from each other.

Russia is a vast country with a highly heterogeneous population. At present, there are no SNP arrays to match its diversity. Even the Humans Origins array turned out to be insufficient for the correct differentiation of the population of Northern Eurasia, since it is focused on the world gene pool as a whole. This study was an endeavor to improve the accuracy of biogeographic ancestry predictions for the populations of Inner Eurasia. To that end, the population of this region was divided into 29 ethnic geographic groups that fairly adequately represented its diversity. We determined the range of the most informative autosomal markers that effectively characterize North Eurasian populations and developed a model and software for ancestry inference based on these markers. For the sake of the end user's convenience, we supplied the software with a cartographic module that shows the most probable area of a person's ancestral origin on the geographic map.

MATERIALS AND METHODS

Samples

Genotype data was generated from samples representing North Eurasian populations using genome-wide SNP arrays. Most of the analysis was conducted on the data generated by an Infinium Omni5Exome-4 v1.3 BeadChip Kit (Illumina; United States) featuring 4.5 M SNPs. The dataset consisted of 1,883 samples from 266 populations of Russia and its neighbor states. The samples represented 92 ethnic groups from 12 North Eurasian countries: Armenia, Azerbaijan, Georgia, Kazakhstan, Kyrgyzstan, Lithuania, Moldova, Mongolia, Russia, Turkey, Ukraine, and Uzbekistan. The samples were provided by the Biobank of North Eurasia (Balanovska et al., 2016). To avoid terminological confusion when using the words "population", "people", "sub-ethnic group", "geographic group", "region", etc., we propose the term "ethnic geographic groups" (EGG) to denote groups of populations that in their totality represent an entire geographic region in such a way that each EGG is relatively genetically

homogeneous, but at the same time, its gene pool differs from that of other EGGs.

The study was approved by the Ethics Committee of the Research Centre for Medical Genetics, Moscow, Russia. All procedures performed in studies involving human participants were in accordance with the ethical standards and with the Helsinki declaration (1964).

The written informed consent was obtained from all individual participants included in the study.

Datasets

Quality control was performed with PLINK 1.9 (Chang et al., 2015). The following filters were applied to create datasets for PCA plots: --geno 0.05 (filters out SNPs with a missing rate over 5%), --maf 0.01 (filters out SNPs with a minor allele frequency below 0.01), --mind 0.1 (excludes individuals with over 10% missing genotype data), and --indep-pairwise 1500 150 0.2 (removes SNPs that are in high linkage disequilibrium with each other). The same filters were applied to create a dataset for SNP selection. The output data were converted to vcf and then to a csv file in which 0 denoted the 0/0 genotype, 1 denoted the 1/0 genotype, and 2 denoted the 1/1 genotype. Finally, missing genotypes were imputed. Imputation is needed because of the inability of a lot of machine learning algorithms to work with missing data. While we have a lot of markers in the initial dataset, we develop the software for the prediction that uses only a limited number of markers. Although haplotype imputation is more accurate, five thousand markers are not enough for this kind of approach. We decided that using a single method for all the data would be more appropriate, so that the training and the test datasets, as well as any newly generated data in the future, would undergo the same preprocessing. Thus, missing genotypes were imputed by replacing a missing value with the most frequent genotype for a given SNP across all 1883 profiles.

After raw data filtering, 51 samples were excluded. Additionally, we removed 19 related samples using KING 2.2.4 software (Manichaikul et al., 2010); all settings were set to default, relatedness was estimated using the --related option. The final dataset consisted of 1813 samples.

PCA and FST

PCA and FST were conducted using the smartpca tool from the EIGENSOFT software package (Price et al., 2006). Default parameters were used except for the number of iterations for outlier removal in PCA set to 3. The filtered dataset after quality control and pruning described in the previous section was used as input data.

Machine Learning Algorithms

All machine learning algorithms were used as implemented in the Python 3 Scikit-learn module (Pedregosa et al., 2011). The metrics used are also those implemented in Scikit-learn. All parameters were set to default values if not said otherwise in the Results section. The random seed was fixed for all of the methods to ensure the reproducibility of the study.

RESULTS

Workflow Overview

The dataset included 266 populations from 12 North Eurasian countries: Armenia, Azerbaijan, Georgia, Kazakhstan, Kyrgyzstan, Lithuania, Moldova, Mongolia, Russia, Turkey, Ukraine, and Uzbekistan. The studied populations represent most of the ethnic diversity across this vast territory (**Figure 1**, **Supplementary Table S1**). A platform for biogeographic ancestry identification was developed in 5 steps. We started by identifying "ancestry groups", or "ethnic geographic groups", i.e., groups of populations that are genetically distinguishable enough to trace individual ancestry. Then, we estimated how many SNPs were needed to differentiate between these groups and chose 5,000 most informative SNPs from the Illumina array of 4.5 M markers. In the third step, we developed a machine learning algorithm for inferring an ancestral EGG. After that, we implemented this algorithm in the original software supplied with a cartographic module for constructing geographic maps of ancestry probabilities. Finally, we validated the proposed method and evaluated its precision.

Subdividing North Eurasia Into 29 Ethnic Geographic Groups

We aimed to achieve the highest possible geographic resolution of ancestry estimates relying on the limited number of SNPs. There were 266 populations in our dataset (**Figure 1**), and obviously, it was impossible to genetically distinguish between closely related geographically neighboring populations. This raised the need for clustering the studied populations into groups that would be genetically distinguishable yet relatively internally homogenous. However, the populations were grouped by their genetic characteristics. Below, the groups will be referred to as ethnic geographic because most of them comprised ethnically and linguistically related populations that occupy contiguous territories. In addition to relative genetic homogeneity within a group and apparent differences between the groups, each group had to be represented by at least 25 samples.

The grouping procedure was previously detailed in (Gorin et al., 2020). Briefly, three principal components were computed for the entire dataset of 1,813 samples (after raw data filtering) using three iterations for outlier removal. For each population, the mean value of each principal component was calculated, and the K-means algorithm was applied to these mean values to partition them into clusters. To obtain clusters with a desired average size, K was set to 30. The method produced 30 imbalanced EGGs (4 samples in the smallest group and 294 samples in the largest). To reduce the imbalance, some of the EGGs were merged while others were broken down into smaller groups so that their size was neither too small (<25 samples) nor too large (>150 samples). The validity of these changes was tested using additional PCA plots for the merged/divided populations and by calculating

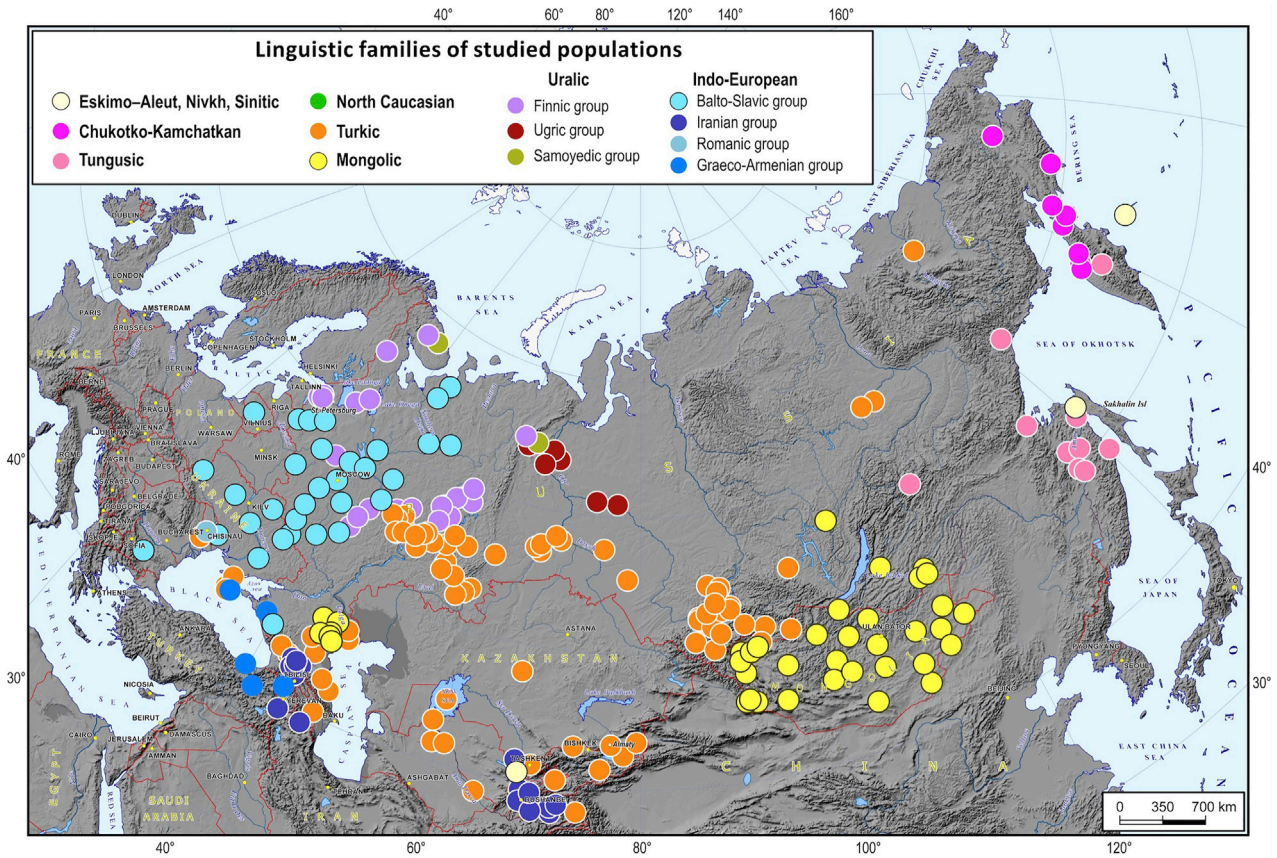


FIGURE 1 | A map of the 266 populations of North Eurasia used for the analysis. Notes. Dots of different colors on the map are languages spoken by the representatives of the studied populations (the color legend is provided at the top of the map).

FST for all pairs of populations. We were not able to merge some of the smaller populations due to their size and genetic difference from other populations, so we removed them from the dataset (40 samples in total). We ended up with 29 groups of populations (EGGs) identified from a set of 1,773 samples (**Table 1**). **Figure 2** shows the area on the map occupied by these groups. **Figure 3** and **Supplementary Figure S1** show PCA plots for the entire dataset, i.e., 4.5 M SNPs; the color of each sample coincides with the color of the ethnic geographic group it represents.

SNP Selection

Various methods of SNP selection were tested. The results were compared using an F1-score metric, which is a harmonic mean of precision and recall and therefore ensures a balanced evaluation of predicting power of the model. There were over 817,120 candidate SNPs after raw data filtering, which, considering the small number of samples (1,773), is overwhelming for most feature selection algorithms. At first, we tried the lasso method without univariate feature selection. The resulting F1 score was only 0.42 on average. So, univariate feature selection was performed as a preprocessing step. The chi-square test was applied to each SNP within each class, i.e., EGG. For each class, SNPs with the highest chi-squared values were selected for further analysis. Besides, we experimented with various numbers of SNPs to represent each class and finally settled on 2,000 SNPs. This approach allowed us to reduce the number of candidate SNPs to 50,000–60,000, which is high enough to prevent significant SNPs from being left out and low enough for feature selection algorithms to process the dataset.

For further feature selection, various models were tested. To choose the best feature selection model, we trained a few logistic regression models with identical parameters on the samples of the selected SNPs. The F1 scores obtained by the models were averaged between EGGs and compared to each other. The first tested model was the lasso method without univariate feature selection, which produced an average F1-score of 0.42. By applying the chi-square test, we were able to increase the score to 0.62. Further improvements were achieved by adding size-appropriate weights to classes (EGGs) during model training; this produced an F1 score of 0.65. The procedure was severely affected by overfitting, so we tested the models with less tendency to overfit. The best result was demonstrated by the ExtraTrees classifier less affected by overfitting due to the randomness of the algorithm. Besides, ExtraTrees assigns a score value to each feature and thus can be used to select SNPs with the best score. By

TABLE 1 | Populations and sizes of EGGs.

No	EGG	Populations	Size
1	Amur_Nanais&Nivkhs&Orochi&Ulchi	Nanais, Nivkhs, Ulchi, Orochs	55
2	Bashkirs	Bashkirs	44
3	Buryats&Khamnegan&Yakuts	Buryats, Khamnegan, Yakuts	59
4	Chechens&Ingush	Chechens, Ingush	39
5	Chukchi&Koryaks&Itelmen	Koryaks, Itelmens, Kamchadals, Chukchi, Itelmens	75
6	Dagestan	Avars, Kubachins, Dargins, Tabasarans, Laks, Lezgins, Rutuls	74
7	Evenks&Evens	Evens, Evenks	49
8	Karelians&Veps	Karelians, Vepsa	38
9	Kazakh&Karakalpak&Uigur&Nogais	Karakalpaks, Nogais_Astrakhan, Nogais_Stavropol, Uyghurs, Kazakhs	33
10	Khakass&AltaiSouth	Khakass, Altaians	46
11	Khanty&Mansi&Nenets	Khanty, Nenets, Mansi	53
12	Komi&Udmurts	Komi Permyaks, Komi Zyrians, Udmurts, Besermyan	84
13	Kyrghyz	Kyrghyz	43
14	Mari&Chuvash	Chuvashes, Mari	53
15	Mongols&Kalmyks	Mongols, Kalmyks	127
16	Mordovians	Mordovians Moksha, Mordovians Erzya, Mordovians Shoksha	41
17	Ossets	Ossetians	36
18	Russians_North	Russians, Izhora, Vod	81
19	Russians_Southern	Russians, Belorussians	240
20	Russians_VeryNorth	Russians	35
21	Shors&AltaiNorth	Shors, Altaians	37
22	Siberian Tatars	Tatars Siberian	68
23	Tajiks&Pomiri&Yaghnobi	Pomiri, Tajiks, Yaghnobi	72
24	Tatars	Tatars Krayshen, Tatars Kazan, Tatar _Mishar, Tatars from Bashkortostan, Tatars Astrakhan	60
25	Transcaucasia&Crimea	Armenians, Azeri, Tatars_Crimean, Karaites, Turks, Kurds, Ezids, Georgians	113
26	Tuvinians&Tofalars	Tuvinians, Mongols, Tofalars	64
27	Ukrainians	Ukrainians	79
28	Uzbeks&Turkmens	Turkmens, Uzbeks	55
29	West_Caucasus	Adyghe, Kabardinians, Shapsug, Karachays, Abkhazians, Circassians, Abazins, Balkars	87

adjusting a score threshold, the number of SNPs that get into the final list can be changed. Using the random forest algorithm, we were able to achieve an F1-score of 0.75.

Stratified k-fold cross-validation was applied to further reduce the influence of overfitting. The dataset was split into five subsets (k folds), and the random forest algorithm was trained on these five subsets. An SNP was included in the final list of selected SNPs if its score was above the threshold value in all five models. This allowed us to increase the F1 score to 0.79.

The model performed well for most EGGs, but there were two EGG pairs and one triplet that were often confused by the algorithm: “Northern Russians” and “Southern Russians”; “Mordovians” and “Ukrainians”; “Kazakh&Karakalpak&Uigur&Nogai”, “Kyrghyz” and “Mongols&Kalmyks”. However, all these EGGs were clearly distinguishable on the PCA plots, so we decided to expand the list of optimal SNPs with extra 100 markers with the highest weight that distinguished EGGs in the pairs. To overcome the problem with the triplet, we added 100 SNPs that distinguished two EGG in the triplet from the third and 100 more SNPs that distinguished the two EGGs from each other. This improved the average F1 score to 0.81.

To determine the optimal number of SNPs to be included in the final list of markers, the described workflow was run several times with various numbers of SNPs. Then the logistic regression model was trained on each of the SNP sublists and the performance of the models was compared based on the F1 score. The F1 plot for different numbers of SNPs chosen for ancestry prediction is shown in **Supplementary Figure S2**.

As seen from **Supplementary Figure S2**, 4,000 SNPs should be enough to achieve a prediction close to the best possible prediction that can be generated by this model. However, to compensate for imperfect genotyping, we expanded the list to 5,000 SNPs.

After preliminary experiments, we ran the final SNP selection process. First, we selected 2,000 SNPs for each EGG using the chi-square test. Due to the overlap of these SNP sets, our list was narrowed down to 54,522 SNPs. Then, we used ExtraTreesClassifier with balanced class weights and determined the optimal number of estimators to use with cross-validation (CV) and trained one-*vs*-rest logistic regression. The best results were achieved with 320 estimators. After training the model and selecting SNPs with scores above 0.000027 in all CV splits, we ended up with 4,851 SNPs. Then, we added 400 SNPs from the principal components of problematic pairs and triplets to the dataset. The final list comprised 5,229 selected SNPs. The dataset with 1,883 samples and 5,229 SNPs is available in a PLINK format *via* correspondence; characteristics of the samples are provided in **Supplementary Table S1**. The flowchart of the final SNP selection process is shown in **Supplementary Figure S3**.

To check whether the selected SNPs adequately reflected the population structure, we constructed PCA plots based on 5,229 SNPs included in the final list (**Supplementary Figure S4** and

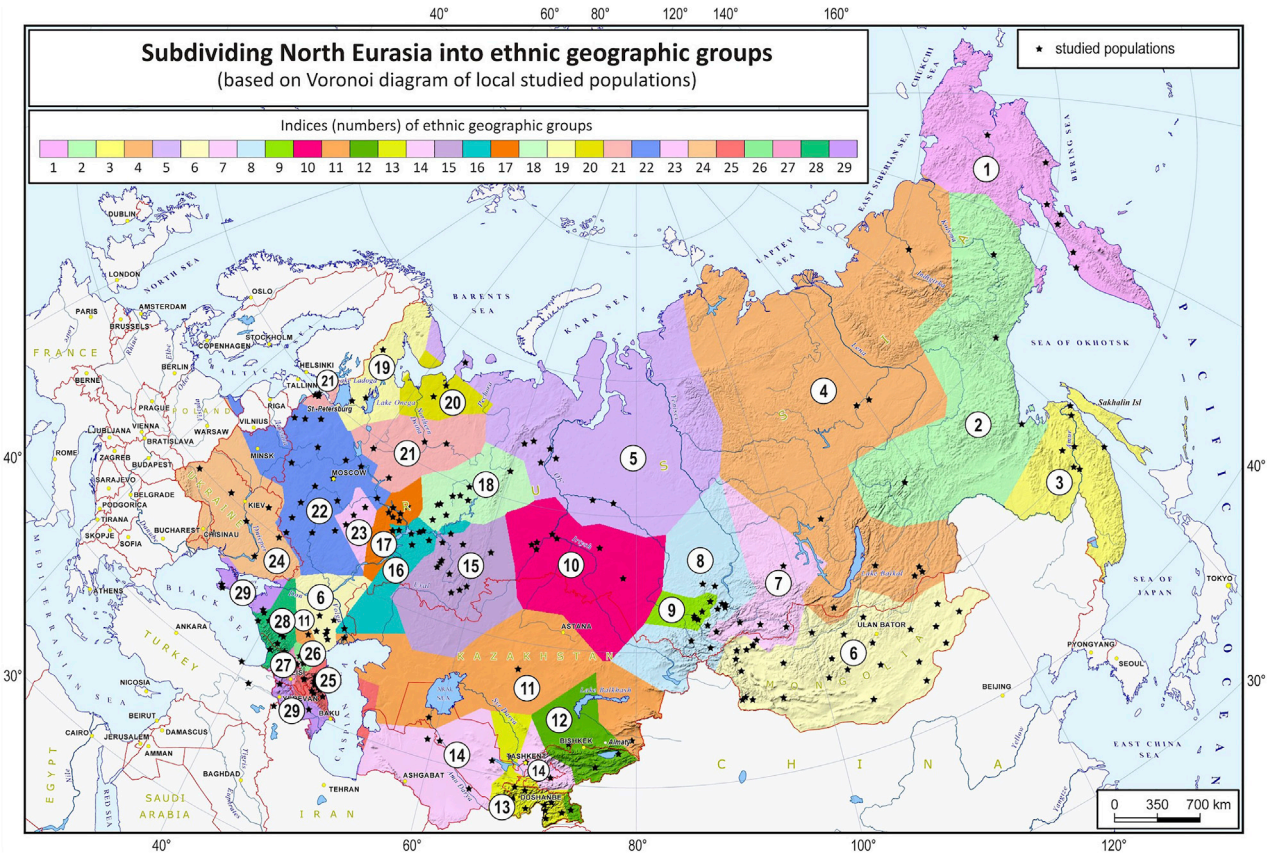


FIGURE 2 | North Eurasia divided into genetically distinguishable ethnic geographic groups. Notes. Colored zones on the map designate areas occupied by the identified ethnic geographic groups. Groups are numbered according to their geographic coordinates. Black stars represent local populations (coincide with the populations in **Figure 1**).

Supplementary Figure S5). Other PCA plots were constructed for the same population sample using the entire set of 4.5 M SNPs from the Illumina panel (**Figure 3** and **Supplementary Figure S1**). The two sets of plots were compared, revealing similar patterns. There was a greater overlap of some population clusters in the second pair of plots, and the distances between some clusters were shorter. However, a decrease in resolution is inevitable with fewer SNPs. By reducing the number of SNPs 1,000-fold, genotyping is made a lot simpler, while the general pattern of genetic similarities between populations remains the same, and the selected set of SNPs allows ancestries to be inferred.

Building the Prediction Model

After SNP selection was completed, the best mathematical model (classifier) for predicting ancestral populations was chosen and trained. We tested various classifiers supporting multiple classes and output values for each class that could be interpreted as probabilities, including logistic regression, multilayer perceptron (MLP), different variants of Support Vector Classifiers, Naive Bayesian classifiers, and some types of bagging and boosting random forest methods. Their performance was compared based on the average F1-score in all EGGs in 5 CV splits. The best results were demonstrated by MLP and logistic regression (the average F1-score was 0.81). We made an attempt to tune both models. Adjustment of MLP parameters did not improve the score. By tuning the logistic regression model, a slight improvement of the F1 score was achieved, but the score was lower than that obtained with MLP. However, logistic regression is more straightforward and trains much faster than MLP, so we opted for one-*vs*-rest logistic regression with the following parameters: L2 penalty, C equal to 1, and balanced class weight.

This model was trained on 1,773 population samples using previously selected 5,229 SNPs.

Developing Software for Ancestry Prediction and Mapping the Results

As a part of this study, we developed software for ancestry prediction. The software named Homeland (available *via* correspondence) consists of 3 modules: the interface module, the prediction module, and the cartographic module.

The interface module aggregates data from other modules and translates it into a user-friendly format. The module allows the user to submit genotype samples and returns the result of biogeographic ancestry estimation, which can be

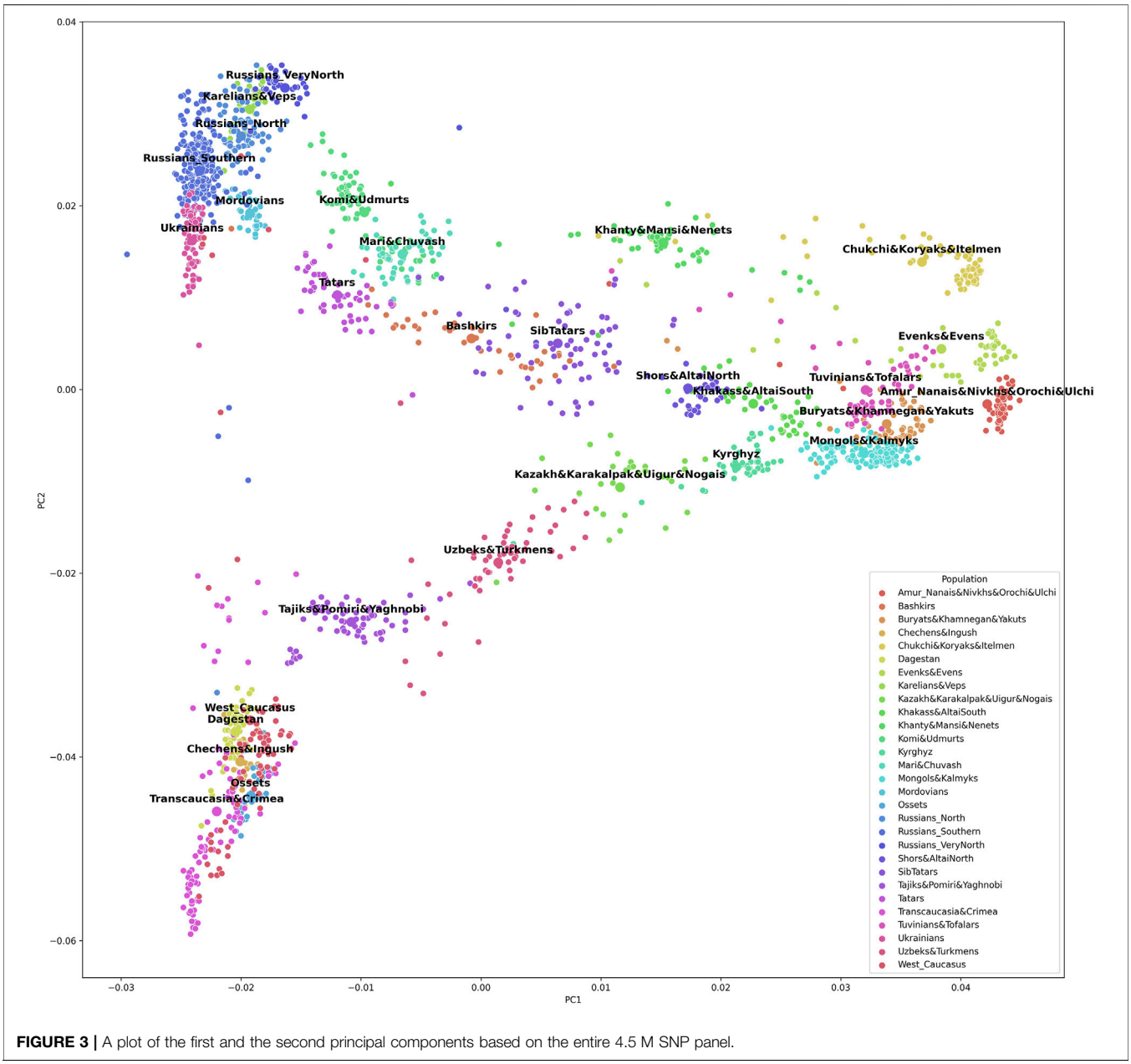


FIGURE 3 | A plot of the first and the second principal components based on the entire 4.5 M SNP panel.

subsequently printed out as a report or visualized on a savable map.

The prediction module estimates a person's biogeographic ancestry from a submitted genotype. The result is a set of geographic points with different probabilities of ancestral origin.

The cartographic module builds JPG maps of probable biogeographic ancestry using the geographic points received from the prediction module and hard-wired settings (cartographic base, scale types, parameters of probability interpolation). The module shows the area of probable ancestral origin predicted from the submitted genotype on the geographic map, which is very convenient for practical work. Besides, the accuracy of prediction can be improved using interpolation: the module will highlight the area where the submitted genotype occurs (**Figure 4**).

Validating the Method

To evaluate the prediction power of the model, we trained it on 1,241 population samples from our dataset and then tested it on the remaining 532 samples. The results are presented in **Table 2**. The EGG prediction heatmap is shown in **Supplementary Figure S6**. The Figure shows a bright diagonal reflecting the effectiveness of the model: most of the EGG predictions were correct (weighted average precision: 0.85; weighted average recall: 0.84; **Table 2**). The

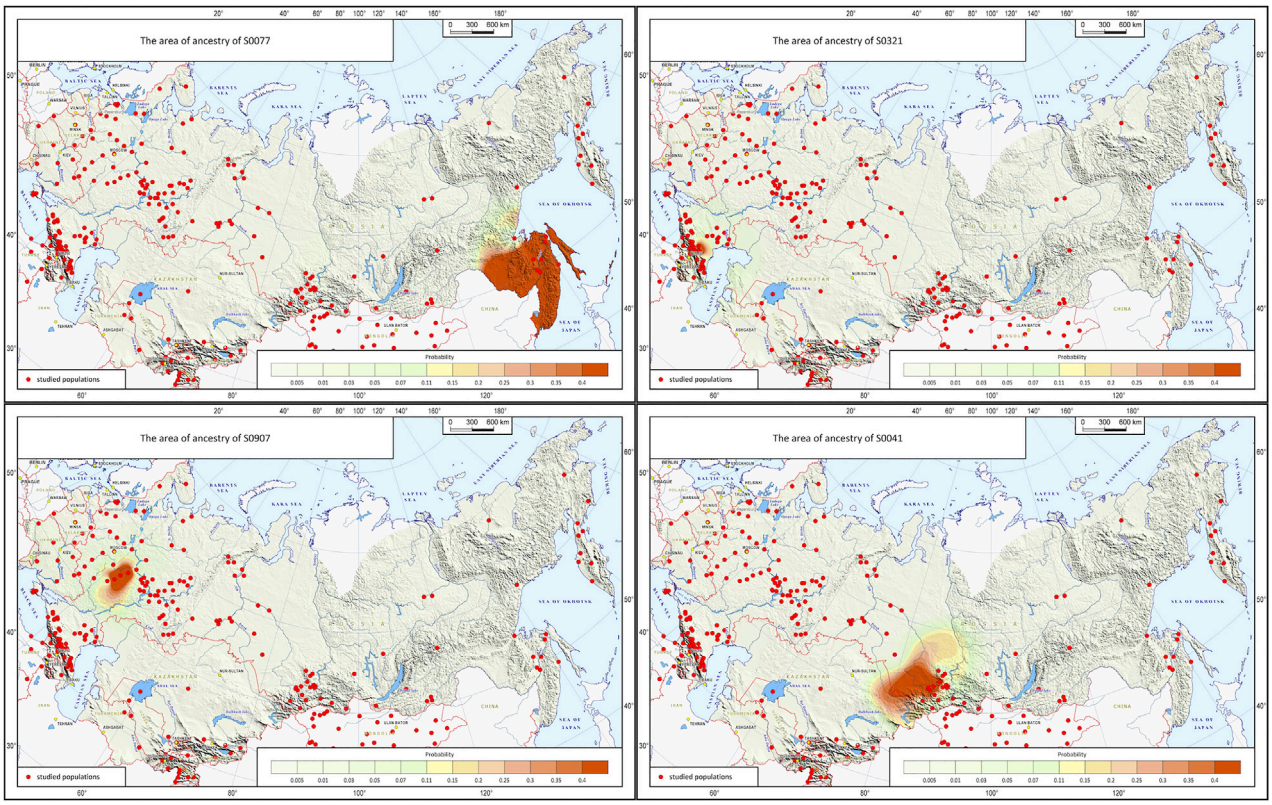


FIGURE 4 | Example of the map generated by the Homeland software.

model made correct predictions about the geographic ancestry of absolutely every individual sample from the following populations: Mari&Chuvash, Ukrainians, Khanty&Mansi&Nenets, Chukchi&Koryaks&Itelmen, Evenks&Evens, Amur_Nanais&Nivkhs&Orochi&Ulchi, Tuvinians&Tofalars, Shors&AltaiNorth (**Supplementary Figure S6**). There were a few cases when the sample was assigned to 2 EGGs (one correct + one false). These errors occurred with the following groups: Russians_VeryNorth, Karelians&Veps, Russians_North, Russians_Southern, Komi&Udmurts, Mordovians, Buryats&Khamnegan&Yakuts, Mongols&Kalmyks, Khakass&AltaiSouth, Tajiks&Pomiri&Yaghnobi, Ossets, Transcaucasia&Crimea.

False ancestry predictions occurred when the falsely predicted EGG was genetically or regionally close to the actual EGG. For example, 4 individuals from the Russians_Southern population were wrongly recognized by the model as Ukrainians, and Ukrainian ancestry was falsely predicted for 4 Mordovians, 3 Karelians&Veps, 4 Tatars and 3 representatives of the West_Caucasus group (**Supplementary Figure S6**). Following formal evaluation criteria, this could be interpreted as a reduction in precision. However, all of these falsely predicted EGGs either neighbor the Ukrainian group on the map (Russians_Southern) or inhabit the same region of Eastern Europe so that the false predictions may be due to the high frequency of genotypes inherited from the common ancestor protopopulaton and now spread across this region.

A reduction in sensitivity (low recall) was observed when the sample was assigned to the wrong EGG within the actual ancestral geographic region. Such errors most frequently occurred for the populations of Ural, West Siberia, Central Asia, and Caucasus (**Supplementary Figure S6**). According to earlier population genetics studies, these territories are highly genetically diverse, which is illustrated by the maps of genetic borders (Pagani et al., 2016; Jeong et al., 2019). This may be due to the vast variety of population sources for these regions. Their contribution differs significantly even between two neighboring populations: being dominant in one population, the contributing genetic component can be very low in another. Therefore, larger sample size and further division of heterogeneous EGGs into more homogenous groups may be needed to ensure more accurate predictions within these regions.

Notably, the highest number of EGG predictions with (almost) absolute accuracy and sensitivity was observed for South and Central Siberia, Far East, and Kamchatka (**Supplementary Figure S6**).

DISCUSSION

Forensic science may benefit from a tool for predicting the geographic area of a person's ancestral origin based on no more than a few thousand SNPs. Studies exploring the gene

TABLE 2 | Resulting metrics of predictions for each EGG.

	Precision	Recall	f1-Score	Support
Amur_Nanais&Nivkhs&Orochi&Ulchi	1.00	1.00	1.00	12
Bashkirs	0.71	0.77	0.74	13
Buryats&Khamnegan&Yakuts	0.79	0.88	0.83	17
Chechens&Ingush	1.00	0.63	0.77	8
Chukchi&Koryaks&Itelmen	1.00	1.00	1.00	20
Dagestan	0.90	0.90	0.90	20
Evenks&Evens	1.00	1.00	1.00	14
Karelians&Veps	1.00	0.73	0.84	11
Kazakh&Karakalpak&Uigur&Nogais	0.75	0.30	0.43	10
Khakass&AltaiSouth	1.00	0.92	0.96	13
Khanty&Mansi&Nenets	0.94	1.00	0.97	16
Komi&Udmurts	0.96	0.88	0.92	25
Kyrghyz	0.83	0.50	0.63	10
Mari&Chuvash	0.84	1.00	0.91	16
Mongols&Kalmyks	0.82	0.95	0.88	38
Mordovians	0.80	0.67	0.73	12
Ossets	0.86	0.55	0.67	11
Russians_North	0.80	0.35	0.48	23
Russians_Southern	0.75	0.93	0.83	59
Russians_VeryNorth	1.00	0.90	0.95	10
Shors&AltaiNorth	1.00	1.00	1.00	10
Siberian Tatars	1.00	0.65	0.79	20
Tajiks&Pomiri&Yaghnobi	0.81	0.95	0.88	22
Tatars	0.60	0.38	0.46	16
Transcaucasia&Crimea	0.92	0.96	0.94	25
Tuvinians&Tofalars	1.00	1.00	1.00	17
Ukrainians	0.57	1.00	0.73	24
Uzbeks&Turkmens	0.86	0.86	0.86	14
West_Caucasus	0.75	0.81	0.78	26
—	—	—	—	—
accuracy	—	—	0.84	532
macro avg	0.87	0.81	0.82	532
weighted avg	0.85	0.84	0.83	532

pools of the western (Western Europe) and eastern (Central and East Asia) poles of Eurasia have generated a massive body of evidence, which, unfortunately, only partly explains the characteristics of the North Eurasian gene pool. They could be better understood using data on the populations of North Eurasian countries that share a history of strong migration flows in the past and present. We determined the range of the most informative autosomal markers in this study that effectively characterize North Eurasian populations and developed a model and software for ancestry inference based on these markers.

Preliminary tests of the proposed model for ancestry prediction allowed us to quantitively evaluate its performance. The analysis of tables generated by the software revealed that the proportion of correct predictions (matches between the actual EGG and the most probable EGG) was 71%. On the maps, the proportion of correct predictions (the actual geographic location being within the most probable predicted region) reached 61% for more likely areas of origin and 81% for less likely areas of origin. Considering the plethora of ethnic geographic groups and the complex population structure of North Eurasia, the proposed method for biogeographic ancestry prediction has demonstrated very good performance.

Merging ethnic geographic groups into larger clusters or expanding the geographic area of probable ancestry improves the accuracy of the model (the proportion of correct predictions) but adversely affects the informative value of the method (geographic precision). This raises the need for further refinement that can be achieved by finding the right balance between accuracy and informative value. Almost absolute accuracy was demonstrated for the majority of EGGs from Siberia, Far East, and Kamchatka. Quite accurate ancestry predictions were achieved for the populations of East Europe, Ural, West Siberia, Caucasus, and Central Asia, and the observed minor deviations in accuracy suggest high genetic heterogeneity in these regions. In our opinion, improvements in prediction accuracy can be achieved by increasing the sample size of the training dataset.

In its current state, the proposed method can be employed to predict ancestries from the populations of Russia and its neighbor states. It can be used for the needs of forensic science and genetic genealogy.

Our method has two limitations: the genotyping approach is expensive, and the method itself has not been optimized for admixed individuals.

We do not propose a genotyping platform that could be used to genotype a DNA sample for the set of 5,000 SNPs. We assume that the sample that the end user has at their disposal has already been genotyped. We propose a method and software to estimate ancestries from the genotype. At the moment, the genotype can be obtained either by using the Illumina Infinium Omni5Exome-4 v1.3 BeadChip or through whole-genome sequencing. We collaborate with another research team that is currently developing a genotyping system for these and some other forensic SNPs. This system will be discussed in a separate publication. Another possible genotyping option is targeted sequencing.

Our method for biogeographic ancestry inference was developed and validated using the set of non-admixed individuals, so the algorithm tends to generate low probabilities of origin from every included ethnic geographic group for genotypes originated from admixed individuals. There are other methods suitable for admixed genetic profiles (Kozlov et al., 2015). Our primary goal was to achieve the highest possible geographic precision of ancestry prediction, and we intentionally focused on non-admixed individuals.

DATA AVAILABILITY STATEMENT

We followed the regulations of the Russian Federal Law on Personal Data (No. 152-FZ). The generated raw data can be shared *via* personal communication with the corresponding author with the following conditions, (i) the data can be only used for studying population history, (ii) the data will not be used for commercial purposes, (iii) the data will not be used to identify the sample donors, (iv) the data will not be used for studying natural/cultural selections, medical or other related studies. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the Research Centre for Medical Genetics, Moscow, Russia. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

Conceptualization, OB; Methodology, IG, OB, and EB; Software, OK and SK; Validation, IG and OB; Formal Analysis, OB; Investigation, IG and OB; Resources, MZ and EK; Data Curation, MZ, EK, AA, and VP; Writing–Original Draft, IG and OB; Writing–Review and Editing, MZ, AA, VP, and EB, Visualization, OK and SK; Supervision, OB and EB; Project Administration, EB; Funding Acquisition, EB.

FUNDING

This work was supported by the Russian Ministry of Science and Higher Education (Government Contract # 011–17 dated 26 September 2017). Genotyping and manuscript preparation were done as a part of the DNA-based identification Research and Technology Project of the Union State. Bioinformatic analysis and interpretation of the obtained results were carried out under the State Assignment of the Russian Ministry of Science and Higher Education for Vavilov Institute of General Genetics. The role of the geographic proximity factor in the origin of gene pools was determined as a part of the State Assignment of the Russian Ministry of Science and Higher Education for Research Centre for Medical Genetics. This research was funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP08855823).

ACKNOWLEDGMENTS

We thank all the donors who took part in this study and the Biobank of North Eurasia for the collection of DNA samples.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.902309/full#supplementary-material

Supplementary Figure S1 | A plot of the first and the third principal components based on the entire 4.5M SNP panel.

Supplementary Figure S2 | Preliminary analysis of predicting power of the models based on different numbers of SNPs.

Supplementary Figure S3 | Flowchart of SNP selection process.

Supplementary Figure S4 | A plot of the first and the second principal components based on the selected set of 5,229 SNPs.

Supplementary Figure S5 | A plot of the first and the third principal components based on the selected set of 5,229 SNPs.

Supplementary Figure S6 | The heatmap of prediction quality: The real EGG *vs*. predicted EGG for the testing dataset.

Supplementary Table S1 | Studied samples.

REFERENCES

Al-Asfi, M., McNevin, D., Mehta, B., Power, D., Gahan, M. E., and Daniel, R. (2018). Assessment of the Precision ID Ancestry Panel. *Int. J. Leg. Med.* 132 (6), 1581–1594. doi:10.1007/s00414-018-1785-9

Balanovska, E. V., Zhabagin, M. K., Agdzhoyan, A. T., Chukhryaeva, M. I., Markina, N. V., Balaganskaya, O. A., et al. (2016). Population Biobanks: Organizational Models and Prospects of Application in Gene Geography and Personalized Medicine. *Russ. J. Genet.* 52 (12), 1227–1243. doi:10.1134/s1022795416120024

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the Challenge of Larger and Richer Datasets. *GigaScience* 4, 7–15. doi:10.1186/s13742-015-0047-8

Balanovskaia, E. V., and Rychkov, I. G. (1990a). Ethnic Genetics: Ethnogeographic Diversity of the Gene Pool of Human Populations Around the World. *Genetika*, 26 (1), 114, 21.

Balanovskaia, E. V., and Rychkov, I. G. (1990b). Ethnogenetics: Adaptive Structure of the Gene Pool of the Mankind from the Data on Human Polymorphic Genetic Markers, *Genetika* 26 (4), 739, 48.

Gorin, I. O., Petrushenko, V. S., Zapisetskaya, Y. S., Koshel, S. M., and Balanovsky, O. P. (2020). Population-based Biobank for Analyzing the Frequencies of Clinically Relevant DNA Markers in the Russian Population: Bioinformatic Aspects. *Cardiovasc Ther. Prev.* 19 (6), 2732. doi:10.15829/1728-8800-2020-2732

Rychkov, I. G., and Balanovska, E. V., Genofond I Genogeografiia Naseleniia SSSR [Gene Pool and Gene Geography of the USSR Population], *Genetika* 28 (1) (1992) 52–75.

Jeong, C., Balanovsky, O., Lukianova, E., Kahbatkyzy, N., Flegontov, P., Zaporozhchenko, V., et al. (2019). The Genetic History of Admixture across Inner Eurasia. *Nat. Ecol. Evol.* 3, 966–976. doi:10.1038/s41559-019-0878-2

Jin, X.-Y., Wei, Y.-Y., Lan, Q., Cui, W., Chen, C., Guo, Y.-X., et al. (2019). A Set of Novel SNP Loci for Differentiating Continental Populations and Three Chinese Populations. *PeerJ* 7, e6508. doi:10.7717/peerj.6508

Kozlov, K., Chebotarev, D., Hassan, M., Triska, M., Triska, P., Flegontov, P., et al. (2015). Differential Evolution Approach to Detect Recent Admixture. *BMC Genomics* 16, S9. doi:10.1186/1471-2164-16-S8-S9

Lan, Q., Fang, Y., Mei, S., Xie, T., Liu, Y., Jin, X., et al. (2020). Next Generation Sequencing of a Set of Ancestry-Informative SNPs: Ancestry Assignment of Three Continental Populations and Estimating Ancestry Composition for Mongolians. *Mol. Genet. Genomics* 295 (4), 1027–1038. doi:10.1007/s00438-020-01660-2

Lan, Q., Shen, C., Jin, X., Guo, Y., Xie, T., Chen, C., et al. (2019). Distinguishing Three Distinct Biogeographic Regions with an In-house Developed 39-AIM-InDel Panel and Further Admixture Proportion Estimation for Uyghurs. *Electrophoresis* 40 (11), 1525–1534. doi:10.1002/elps.201800448

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., et al. (2008). Worldwide Human Relationships Inferred from Genome-wide Patterns of Variation. *Science* 319 (5866), 1100–1104. doi:10.1126/science.1153717

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust Relationship Inference in Genome-wide Association Studies. *Bioinformatics* 26, 2867–2873. doi:10.1093/bioinformatics/btq559

Mehta, B., Daniel, R., Phillips, C., and McNevin, D. (2017). Forensically Relevant SNaPshot Assays for Human DNA SNP Analysis: a Review. *Int. J. Leg. Med.* 131 (1), 21–37. doi:10.1007/s00414-016-1490-5

Pagani, L., Lawson, D. J., Jagoda, E., Mörseburg, A., Eriksson, A., Mitt, M., et al. (2016). Genomic Analyses Inform on Migration Events during the Peopling of Eurasia. *Nature* 538 (7624), 238–242. doi:10.1038/nature19792

Pakstis, A. J., Speed, W. C., Soundararajan, U., Rajeevan, H., Kidd, J. R., Li, H., et al. (2019). Population Relationships Based on 170 Ancestry SNPs from the Combined Kidd and Seldin Panels. *Sci. Rep.* 9 (1), 18874. doi:10.1038/s41598-019-55175-x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Available at: http://jmlr.org/papers/v12/pedregosa11a.html.

Pereira, V., Freire-Aradas, A., Ballard, D., Børsting, C., Diez, V., Pruszkowska-Przybylska, P., et al. (2019). Development and Validation of the EUROFORGEN NAME (North African and Middle Eastern) Ancestry Panel. *Forensic Sci. Int. Genet.* 42, 260–267. doi:10.1016/j.fsigen.2019.06.010

Phillips, C., McNevin, D., Kidd, K. K., Lagacé, R., Wootton, S., de la Puente, M., et al. (2019). MAPlex - A Massively Parallel Sequencing Ancestry Analysis Multiplex for Asia-Pacific Populations. *Forensic Sci. Int. Genet.* 42, 213–226. doi:10.1016/j.fsigen.2019.06.022

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal Components Analysis Corrects for Stratification in Genome-wide Association Studies. *Nat. Genet.* 38 (8), 904–909. doi:10.1038/ng1847

Stepanov, V. A., Balanovsky, O. P., Melnikov, A. V., Lash-Zavada, A. Y., Khar'kov, V. N., Tyazhelova, T. V., et al. (2011). Characteristics of Populations of the Russian Federation over the Panel of Fifteen Loci Used for DNA Identification and in Forensic Medical Examination. *Acta Naturae* 3, 56–67. doi:10.32607/20758251-2011-3-2-56-67

Xavier, C., de la Puente, M., Phillips, C., Eduardoff, M., Heidegger, A., Mosquera-Miguel, A., et al. (2020). Forensic Evaluation of the Asia Pacific Ancestry-Informative MAPlex Assay. *Forensic Sci. Int. Genet.* 48, 102344. doi:10.1016/j.fsigen.2020.102344

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

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

Copyright © 2022 Gorin, Balanovsky, Kozlov, Koshel, Kostryukova, Zhabagin, Agdzhoyan, Pylev and Balanovska. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 27 May 2022
doi: 10.3389/fgene.2022.880346

Forensic Efficiency Estimation of a Homemade Six-Color Fluorescence Multiplex Panel and In-Depth Anatomy of the Population Genetic Architecture in Two Tibetan Groups

*Yanfang Liu[1,2], Wei Cui[1], Xiaoye Jin[3], Kang Wang[4], Shuyan Mei[1], Xingkai Zheng[4] and Bofeng Zhu[1,5,6]**

[1]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, China, [2]Laboratory of Fundamental Nursing Research, School of Nursing, Guangdong Medical University, Dongguan, China, [3]Department of Forensic Medicine, Guizhou Medical University, Guiyang, China, [4]Ningbo Health Gene Technologies Co., Ltd., Ningbo, China, [5]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, China, [6]Microbiome Medicine Center, Department of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, China

The genetic information of the Chinese Tibetan group has been a long-standing research hotspot among population geneticists and archaeologists. Herein, 309 unrelated individuals from two Tibetan groups living in Qinghai Province, China (CTQ), and Tibet Autonomous Region, China (CTT), were successfully genotyped using a new homemade six-color fluorescence multiplex panel, which contained 59 autosomal deletion/insertion polymorphisms (au-DIPs), two mini short tandem repeats (miniSTRs), two Y-chromosomal DIPs, and one Amelogenin. The cumulative probability of matching and combined power of exclusion values for this new panel in CTQ and CTT groups were 1.9253E-27 and 0.99999729, as well as 1.5061E-26 and 0.99999895, respectively. Subsequently, comprehensive population genetic analyses of Tibetan groups and reference populations were carried out based on the 59 au-DIPs. The multitudinous statistical analysis results supported that Tibetan groups have close genetic affinities with East Asian populations. These findings showed that this homemade system would be a powerful tool for forensic individual identification and paternity testing in Chinese Tibetan groups and give us an important insight for further perfecting the genetic landscape of Tibetan groups.

Keywords: forensic efficiency estimation, population genetics, Tibetan group, genetic architecture dissection, deletion/insertion polymorphisms

OPEN ACCESS

Edited by:
Mengge Wang,
Sichuan University, China

Reviewed by:
Jianhui Xie,
Fudan University, China
Peng Chen,
Nanjing Medical University, China

****Correspondence:***
Bofeng Zhu
zhubofeng7372@126.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *21 February 2022*
Accepted: *06 April 2022*
Published: *27 May 2022*

Citation:
Liu Y, Cui W, Jin X, Wang K, Mei S, Zheng X and Zhu B (2022) Forensic Efficiency Estimation of a Homemade Six-Color Fluorescence Multiplex Panel and In-Depth Anatomy of the Population Genetic Architecture in Two Tibetan Groups. Front. Genet. 13:880346. doi: 10.3389/fgene.2022.880346

1 INTRODUCTION

Critical samples from crime scenes may contain only small amounts of highly fragmented and damaged DNA, so forensic scientists make every effort to address this complex problem. Numerous research approaches have emerged to amplify shorter amplicons of the target sequences in the damaged DNA samples from mass disasters or forensic caseworks (Senge et al., 2011). Short tandem repeats (STRs) have been authenticated to be highly sensitive, dependable, and discriminating for parentage testing and personal identification (Gymrek,

2017). In consideration of the same characteristics as the STRs, miniSTR typing is the representative analytical method of first choice for the degraded biological samples (Graham, 2005). More accurate DNA profiles can be achieved through miniSTRs, whose primers are located more closely to the repeat regions of STRs, so as to improve genotyping success ratio (Kun, Wictum, and Penedo, 2018). Currently, a series of studies have also suggested that single nucleotide polymorphism (SNP) are the efficacious genetic markers for forensic identification of degradation samples, but there is limited prevalence of SNP kits in forensic laboratories due to their high cost and tedious detection procedures based on the capillary electrophoresis analysis or other platforms (Fondevila et al., 2012). Deletion/insertion polymorphisms (DIPs) with smaller amplicon sizes and a time-saving genotyping process have been widely favored in the forensic molecular biology field (Fondevila et al., 2012). In our prophase research, we constructed a new homemade six-color fluorescence multiplex PCR system encompassing one Amelogenin gene, two Y-chromosome DIPs (Y-DIPs), 59 highly polymorphic autosomal DIPs (au-DIPs), and two mini short tandem repeats (miniSTRs) with amplicon lengths within 200 bp, and it was devoted to detect degradation samples for personal identification and paternity testing in the East Asian region (Liu et al., 2022). However, the genetic distributions and forensic applicability of these genetic markers in the novel panel have not yet been investigated and studied in most populations from China. The genetic polymorphisms of these genetic markers and forensic application efficiency estimation for Chinese Tibetan groups in different regions are important contents in this study.

Tibetan group is one of the nationalities with historic backgrounds, diligence, courage, and wisdom in China. The regions where Tibetans live are near the Tibetan Plateau, which is inlaid with snow-capped highlands and mountains for most of the year, resulting in a relatively geographical closure. Their pronounced historical traditions, distinctive culture, and genetic characteristics made them a long-term focus to explore hotspots among population geneticists and archeologists. In recent decades, many scholars' studies on archeology (Aldenderfer, 2011), language (Sagart et al., 2019), clothing (Li, Guo, and Wang, 2012), Tibetan medicine (Dakpa, 2014), genetics (Hu et al., 2017; Fan et al., 2021), and various aspects have remarkably enhanced our knowledge of Tibetan ethnicity. Nevertheless, these studies, other than the DNA studies, which veritably reflect human evolution, are susceptible to external factors such as the environment or subjective factors. In a previous study, genetic affinities among Tibetan groups and other worldwide populations were explored using 35 DIPs (Liu et al., 2020). But the currently available limited genetic information of Tibetan groups have also been made this topic still need to be explored sufficiently.

Thus, the present research intends to estimate the effectiveness of the new homemade six-color fluorescence multiplex system in Tibetan groups and further perfect the genetic landscape of the Tibetan groups from the perspective of the 59 au-DIPs. This study will be a significant contribution toward understanding the genetic background and diversity of Tibetan group.

2 MATERIALS AND METHODS

2.1 Characteristics of the Loci in the New Panel

The genetic markers have been carefully selected and constructed in a new homemade six-color fluorescence multiplex panel (Liu et al., 2022). The 2 Y-DIPs and 2 miniSTRs in the panel were selected from previously used loci. Physical distances among genetic markers located on autosomes are greater than 10 Mb. The 59 au-DIPs are located in intronic regions with allele lengths of 2–10 bp and the minor allele frequencies are greater than 0.2; and the *F*-statistics (F_{ST}) values for these loci among the five East Asian populations from the 1000 Genomes Project database were less than 0.06.

2.2 Sample Collection, PCR Amplification, and Genotyping

Bloodstain specimens of unrelated healthy individuals examined in the present study were gathered from the Tibetan group in Qinghai Province, China (CTQ), and the Tibetan group in the Tibet Autonomous Region, China (CTT), with the sample sizes of 155 and 154, respectively. The present research was conducted according to the ethical guidelines of the Southern Medical University and Xi'an Jiaotong University Health Science Center and further authorized by the Ethical Committee of the Xi'an Jiaotong University Health Science Center (approval number: 2019-1039).

Bloodstain samples were amplified via the GeneAmp PCR system 9700 (Thermo Fisher Scientific, Waltham, United States). About 1.0 mm^2 was used as template in a reaction mix with 10 μl total volume, comprising 2 μl of Master Mix (HEALTH Gene Technologies, Ningbo, China), 1 μl of Primer Mix, and 7 μl of nuclease-free water. The cycling conditions were as follows: an initial denaturation step of 5 min at 95°C; followed by 2 cycles of 94°C for 10 s and annealing at 63°C for 90 s; and then 23 cycles of 94°C for 10 s, annealing at 60°C for 90 s, and extension at 60°C for 90 s; and then final extension at 60°C for 15 min. Then, 1.0 μl PCR product was mixed with 0.5 μl of an internal lane size standard and 8.5 μl of Hi-Di deionized formamide. The mixtures were denatured at 95°C for 3 min, cooled immediately for 3 min, and then genotyped on the Applied Biosystems® 3500xL Genetic Analyzer (Thermo Fisher Scientific) for electrophoretic separation and detection. Finally, GeneMapper ID version 3.2 software (Thermo Fisher Scientific) was utilized to determine the genotyping results.

2.3 Comparison Populations and Data Analyses

The 26 comparison population data from five different geographic regions (Africa, Europe, America, East Asia, and South Asia) were downloaded from the 1000 genomes database. The statistical analysis information used in this study is summarized in **Table 1**.

3 RESULTS

3.1 Forensic Genetic Parameter Analyses of the New Panel

For the two studied Tibetan groups, *p* values of the LDs (**Figure 1A**) for all loci combinations of 59 au-DIPs and two

TABLE 1 | Statistical analysis information for forensic parameters and population genetic analyses.

Statistical parameter	Software	Description
Exact tests of Hardy–Weinberg equilibrium (HWE-*p*)	STRAF tool (Gouy and Zieger, 2017)	Sample representativeness, locus independence testing, genetic polymorphism, and forensic parameter analysis
Linkage disequilibrium (LD) analysis	-	-
Allele frequency	-	-
Power of exclusion (PE)	-	-
Power of discrimination (PD)	-	-
Probability of matching (PM)	-	-
Polymorphic information content (PIC)	-	-
Observed heterozygosity (H_{obs})	-	-
Allele frequency heatmap	TBtools version 0.665 (Chen et al., 2020)	Insertion allele frequency distribution characteristics of the 59 au-DIPs in the CTQ and CTT groups, and the other 26 reference populations
F_{ST} genetic distance	Arlequin version 3.5 (Excoffier and Lischer, 2010)	Population differentiation among the two studied Tibetan groups and other 26 reference populations
Nei's genetic distance (D_A distance)	DISPAN program	D_A distances among the two studied Tibetan groups and the other 26 reference populations, formed under the assumption that genetic differences originated from genetic drift and mutation events (Jin et al., 2019)
Phylogenetic tree reconstructions	MEGA version 7.0 (Kumar, Stecher, and Tamura, 2016)	Rooted evolutionary tree, which was built based on the F_{ST} values among the pairwise populations by the unweighted pair group method with the arithmetic mean (UPGMA) method
	Phylip version 3.697 (Shimada and Nishida, 2017)	Unrooted evolutionary tree, namely, radiation tree, which was established based on the allelic frequency data using the neighbor-joining method
Principal component analyses (PCA)	*R* Studio	Population level PCA based on the allele frequencies of the same loci
	Origin 2021	Individual level PCA based on the allelic genotyping raw data
	R version 3.5.3	Contribution quality correlation circle of the locus in the corresponding PCA plots
Population genetic structure analyses	STRUCTURE version 2.3.4 (Porras-Hurtado et al., 2013)	Each *K* value (*K* = 2-7) with ten replicates run under the admixture model includes 10,000 burn-in period length, followed by 10,000 Markov chain Monte Carlo steps
	CLUMPP version 1.1.2 (Jakobsson and Rosenberg, 2007)	The ancestor component bar graph drawing after the population genetic structure analysis
	Distruct version 1.1 (Rosenberg, 2004)	-
	Structure Harvester program (Earl and vonHoldt, 2012)	The optimum *K* value determination of STRUCTURE analysis
Population-specific divergence (PSD)	Snipper version 2.5 http://mathgene.usc.es/snipper/	The accumulated PSD values of all loci in distinguishing different geographical region populations
Informativeness for assignment (I_n) and F_{ST} values of locus-by-locus AMOVA	Infocalc version 1.1 https://rosenberglab.stanford.edu/infocalc.html	Determining the level of information about individual ancestry provided by each locus, and the locus-by-locus AMOVA F_{ST} values of each locus to pairs of regional populations

miniSTRs were higher than 0.0003 in the CTQ group (**Supplementary Table S1**) and higher than 0.0005 in the CTT group (**Supplementary Table S2**), which signified that no LDs for pairwise loci were discovered in these two groups after applying the Bonferroni correction ($p > 0.05/1830 = 0.00002732$). The *p* values of the HWE exact tests are displayed in **Supplementary Table S3** and **Figure 1B**. For the two studied Tibetan groups, the HWE *p* values ranged from 0.0300 to 1.0000 and from 0.0180 to 1.0000 in the CTQ and CTT groups, respectively. Some slight deviations for HWE were observed at loci rs10581929 (HWE-$p = 0.0320$), rs35828751 (HWE-$p = 0.0300$), rs3830338 (HWE-$p = 0.0450$), and rs3833559 (HWE-$p = 0.0300$) in the CTQ group, and at loci D1S1656 (HWE-$p = 0.0110$), rs3059221 (HWE-$p = 0.0310$), rs66739142 (HWE-$p = 0.0280$), and rs67365630 (HWE-$p = 0.0180$) in the CTT group. After conducting the Bonferroni correction (0.05/61 = 0.0008), there were no deviations from HWE at these loci.

Herein, the allele frequencies and other forensic parameters are summarized in **Supplementary Tables S3, S4** and **Figures 1B, C**. A total of 136 alleles were confirmed at the 59 au-DIPs and the two miniSTRs in the CTQ group. The insertion allele frequencies of 59 au-DIPs ranged from 0.2260 (rs3833559) to 0.6940 (rs5877451), with a median value of 0.4940. The allele frequencies of D1S1656 and D3S1358 loci ranged from 0.0060 to 0.4130. The PE, PD, PM, PIC, and H_{obs} values of 59 au-DIPs ranged from 0.0571 (rs3833559) to 0.2401 (rs10546179 and rs1611025); from 0.5126 (rs3833559) to 0.6547 (rs10581929); from 0.3453 (rs10581929) to 0.4874 (rs3833559); from 0.2885 (rs3833559) to 0.3750 (rs10540867, rs10649306, rs11283102, rs3043804, rs3083268, rs33928328, rs34899580, rs35713041, and rs4024564); and from 0.2839 (rs3833559) to 0.5548 (rs10546179 and rs1611025), with median values of 0.1738, 0.6199, 0.3801, 0.3732, and 0.4839, respectively. The PE, PD, PM, PIC, and H_{obs} values of the two miniSTR loci D1S1656 and D3S1358 were 0.6727, 0.9496, 0.0504, 0.8094, and 0.8387, as well as 0.3942, 0.8650, 0.1350, 0.6556, and 0.6774, respectively. The cumulative PM (CPM), cumulative PD (CPD), and combined PE (CPE) values of the 59 au-DIPs in the CTQ group were 2.8296E-25, 1-2.8296E-25, and 0.9999863, respectively. After

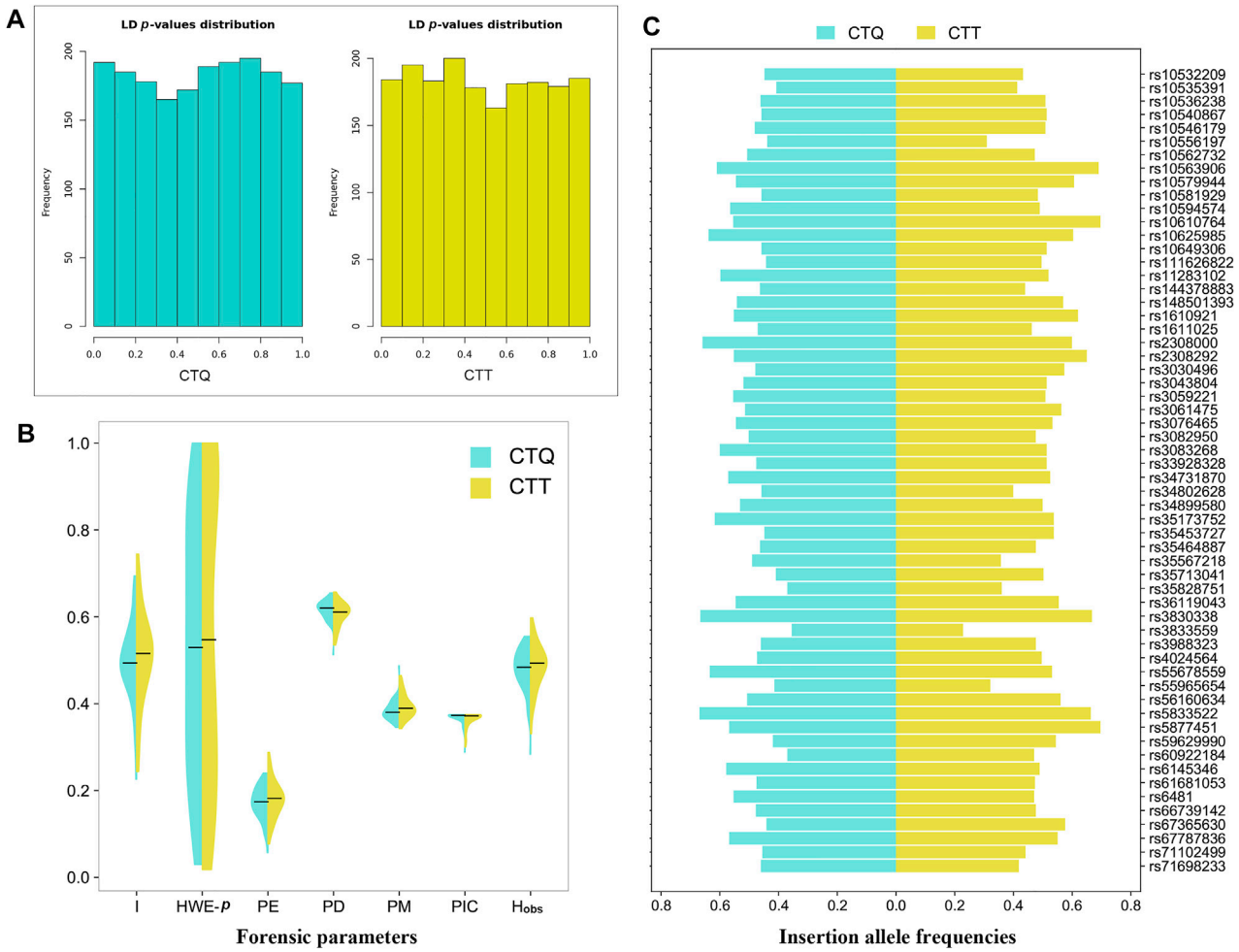


FIGURE 1 | Forensic parameters in the CTQ and CTT groups. **(A)** Bar chart of *p*-value distributions of pairwise linkage disequilibrium tests at the 59 au-DIPs and two miniSTRs. **(B)** Half violin of some forensic parameters for the 59 au-DIPs and two miniSTRs. I, insertion allele frequencies; HWE-*p*, exact tests of Hardy–Weinberg equilibrium; PE, power of exclusion; PD, power of discrimination; PM, probability of matching; PIC, polymorphic information content; H_{obs}, observed heterozygosity. **(C)** Half horizontal bar of insertion allele frequencies at the 59 au-DIPs.

adding the two miniSTRs, their CPM, CPD, and CPE values reached to 1.9253E-27, 1-1.9253E-27, and 0.99999729, respectively.

A total of 138 alleles were detected at the 59 au-DIPs and the 2 miniSTRs in the CTT group (**Supplementary Tables S3, S4**). The insertion allele frequencies of the 59 au-DIPs ranged from 0.2440 (rs10556197) to 0.7440 (rs3830338), with a median value of 0.5160. The allele frequencies of D1S1656 and D3S1358 loci were from 0.0030 to 0.3860. The PE, PD, PM, PIC, and H_{obs} values of the 59 au-DIPs were from 0.0772 (rs10556197) to 0.2878 (rs66739142); from 0.5351 (rs10556197) to 0.6570 (rs3059221); from 0.3430 (rs3059221) to 0.4649 (rs10556197); from 0.3006 (rs10556197) to 0.3750 (rs3061475, rs4024564, and rs61681053); and from 0.3312 (rs10556197) to 0.5974 (rs66739142), with median values of 0.1819, 0.6106, 0.3894, 0.3719, and 0.4935, respectively. The PE, PD, PM, PIC, and H_{obs} values of the D1S1656 and D3S1358 loci were 0.7349, 0.9297, 0.0703, 0.8022, and 0.8701, and 0.5153, 0.8441, 0.1559, 0.6482, and 0.7532, respectively. The CPM, CPD, and CPE values of the 59 au-DIPs in the CTT group were 1.3742E-24, 1-1.3742E-24, and 0.9999919, respectively. After adding the two miniSTR loci, their CPM, CPD, and CPE values reached to 1.5061E-26, 1-1.5061E-26, and 0.99999895, respectively.

3.2 Allele Frequency Divergences of the 59 au-DIPs

Supplementary Figure S1 lists the insertion allele frequency values of the 59 au-DIPs in two studied Tibetan groups and 26 reference populations involved in this study and intuitively revealed the allele frequency divergences among these 28 populations. The color gradation from turquoise to yellow and then to purple meant the transition of insertion allele frequency values from the lowest to the highest. Except that the insertion

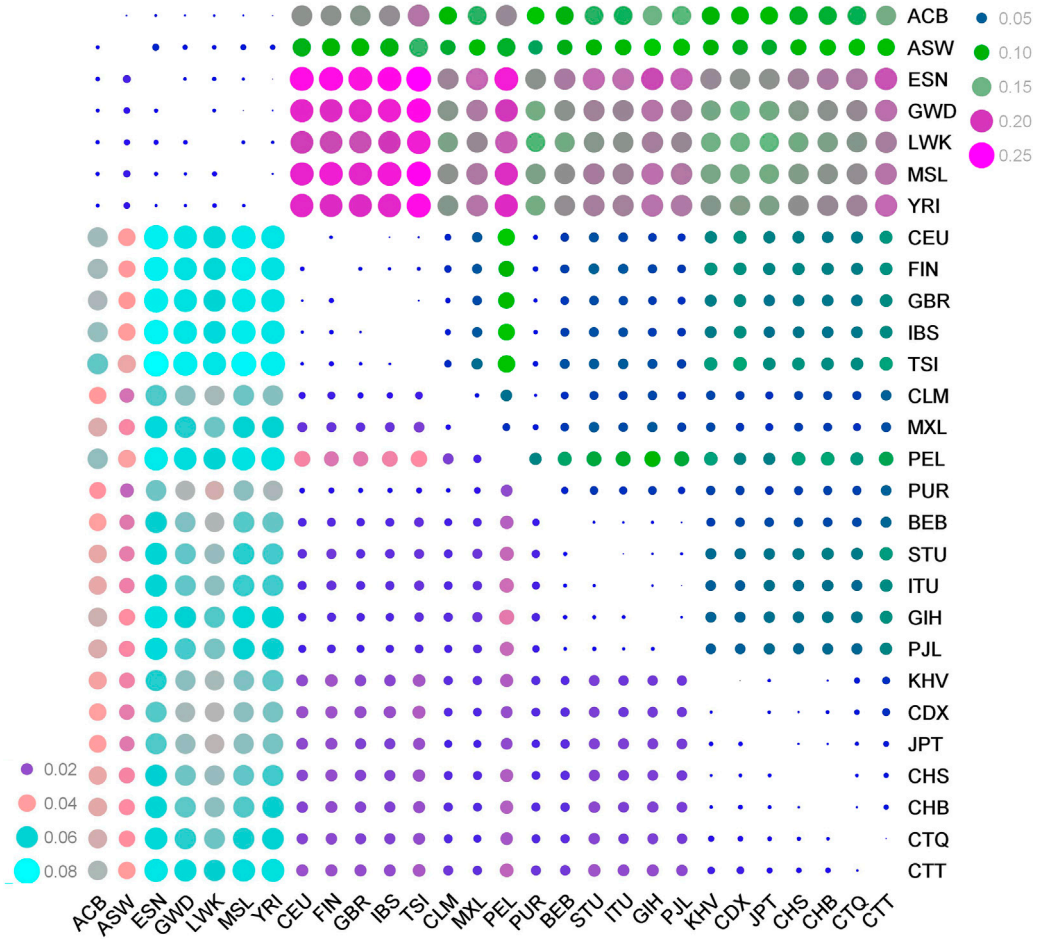


FIGURE 2 | Distribution heatmaps of the pairwise F_{ST} values (the upper right part) and D_A values (the bottom left part) among the CTQ and CTT groups and other 26 comparison populations. CTQ, Tibetan in Qinghai Province, China (n = 155); CTT, Tibetan in Tibet Autonomous Region, China (n = 154); CDX, Chinese Dai in Xishuangbanna, China (n = 93); CHB, Han Chinese in Beijing, China (n = 103); CHS, Southern Han Chinese, China (n = 105); KHV, Kinh in Ho Chi Minh City, Vietnam (n = 99); JPT, Japanese in Tokyo, Japan (n = 104); PJL, Punjabi in Lahore, Pakistan (n = 96); GIH, Gujarati Indian in Houston, TX (n = 103); ITU, Indian Telugu in the United Kingdom (n = 102); STU, Sri Lankan Tamil in the United Kingdom (n = 102); BEB, Bengali in Bangladesh (n = 86); CLM, Colombian in Medellin, Colombia (n = 94); MXL, Mexican Ancestry in Los Angeles, California (n = 64); PEL, Peruvian in Lima, Peru (n = 85); PUR, Puerto Rican in Puerto Rico (n = 104); CEU, Utah residents with Northern and Western European ancestry (n = 99); FIN, Finnish in Finland (n = 99); GBR, British in England and Scotland (n = 91); IBS, Iberian populations in Spain (n = 107); TSI, Toscani in Italy (n = 107); ACB, African Caribbean in Barbados (n = 96); ASW, African Ancestry in Southwest US (n = 61); ESN, Esan in Nigeria (n = 99); GWD, Gambian in Western Division, The Gambia (n = 113); LWK, Luhya in Webuye, Kenya (n = 99); MSL, Mende in Sierra Leone (n = 85); and YRI, Yoruba in Ibadan, Nigeria (n = 108).

allele frequency values of rs3833559 in the CTQ and CTT groups, and of rs10556197 and rs55965654 in the CTT group were less than 0.3, the insertion allele frequency values of the 59 au-DIPs were relatively balanced in the CTQ and CTT groups, and the other five populations from East Asia with the frequency values ranging from 0.3 to 0.7. From the aspect of heatmap clustering (**Supplementary Figure S1**), the branch on the left represented the cluster based on the insertion allele frequencies of each locus in these populations, and the branch above indicated the cluster pattern of the 28 populations based on the insertion allele frequencies of the whole 59 DIPs. By and large, two major cluster classifications (African population cluster and non-African population cluster) appeared in all 28 populations; and the loci in the same small branch reflected similar allele frequency distributions in different populations from the same geographic region.

3.3 Population Genetic Distance Measures

The F_{ST} values and D_A distances of pairwise populations based on the 59 au-DIPs were used to perform genetic relationship analyses from these diverse genetic distances. The results showed (**Supplementary Tables S5, S6**, **Figure 2**) that the minimal F_{ST} value (F_{ST} = 0.0016) and D_A distance (D_A = 0.0012) were between the CTQ and CTT groups. The second smallest F_{ST} value and D_A distance were between the CTQ group and the CHB population (F_{ST} = 0.0035, D_A = 0.0019) from East Asia, and then between the CTT group and the CHS population (F_{ST} = 0.0107, D_A = 0.0038) from East Asia. The maximum F_{ST} values (0.1334 in

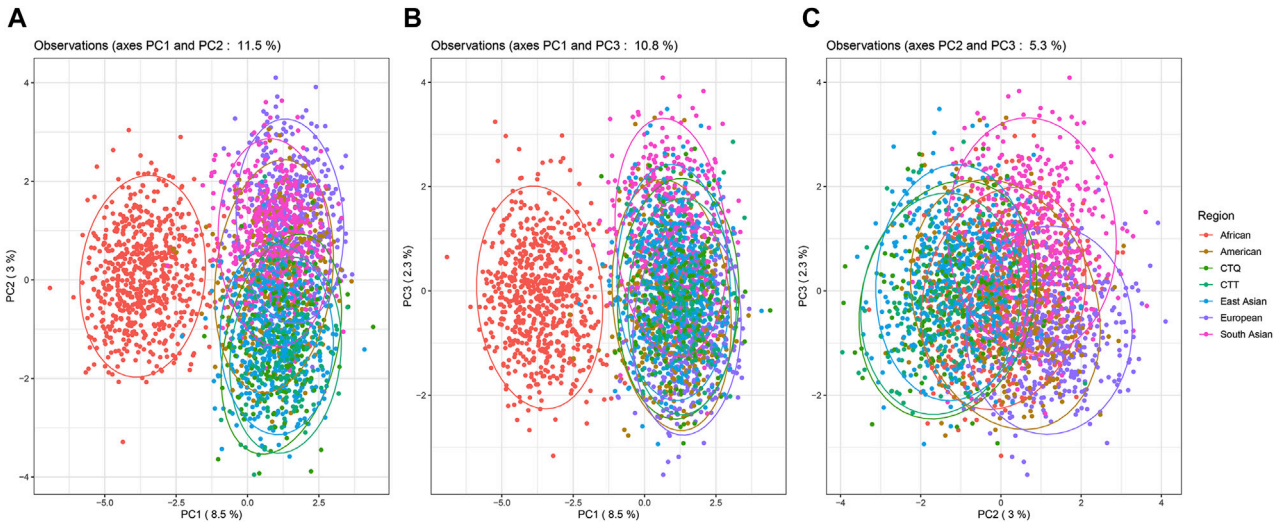


FIGURE 3 | PCA results on the individual level among 2813 individuals from the CTQ and CTT groups, and other five different geographic region populations. **(A)** PCA results based on PC1 and PC2. **(B)** PCA results based on PC1 and PC3. **(C)** PCA results based on PC2 and PC3.

CTQ, 0.1439 in CTT) and D_A distances (0.0439 in CTQ, 0.0469 in CTT) were found between the Tibetan groups and Esan in Nigeria (ESN) from Africa. In the heatmaps of F_{ST} (the upper right part of the picture) and D_A distance (the bottom left part of the picture) values (**Figure 2**), the color degree ranged from purple to pink and then to turquoise, as well as the color degree from blue to green to rose-red denoted the changes from minimum to maximum population hereditary distances. Obviously, pairwise comparisons among the Tibetan groups and the five East Asian populations preferred the small bubbles representing relatively close genetic distances, while the pairwise comparison populations among the Tibetan groups and the African populations were inclined toward the big bubbles illustrating the farther hereditary distances.

3.4 Phylogenetic Reconstructions

Subsequently, phylogenetic trees constructed using two methods (UPGMA and neighbor-joining methods) also revealed the genetic relationships among the CTQ group, the CTT group, and 26 other reference populations. The various color modules in the trees represented the populations from different geographic regions. **Supplementary Figure S2A** presented a circular rooted phylogenetic tree built by the UPGMA method based on the F_{ST} values of the pairwise populations at the same 59 au-DIPs in the novel panel. The 28 populations involved were divided into two primary branches, one of which was dominated by seven populations from the African region and another included the other 19 populations from East Asia, South Asia, America, and Europe regions, as well as the CTQ and CTT groups. The CTQ and CTT groups formed a sister clade with the other five populations from East Asia. The basically analogous phylogenetic branch distributions also were observed in the radiation tree, which was established based on the allelic frequency data applying the Neighbor-Joining method (**Supplementary Figure S2B**).

3.5 Principal Component Analyses

The PCA plots based on the population level (**Supplementary Figure S3**) and individual level (**Figure 3**) were conducted using the allele frequency data and allele genotyping raw data, respectively. In **Supplementary Figure S3**, the top three principal components (PCs) could explain 72.9% of genetic variance. PC1 (46.8%) and PC2 (16.1%) dispersed the European, African, and East Asian populations into three domains which were isolated from each other (**Supplementary Figure S3A**). PC1 (48.8%) and PC3 (10%) could separate the African and South Asian populations from other intercontinental populations (**Supplementary Figure 3B**). PC2 (16.1%) and PC3 (10%) mainly distinguished the South Asian populations from other intercontinental populations (**Supplementary Figure S3C**). In general, populations from the same geographic region clustered closer together, and the CTQ and CTT groups gathered with other populations from East Asia all the time. The contribution qualities for PC1 and PC2 of the 59 au-DIPs in the PCA plot were shown in a correlation circle (**Supplementary Figure S3D**), which was acquired with the square cosine values (Cos2, which is calculated as the squared coordinates), representing the contribution degrees to the PCs. The length from the center point to each variable represented the proportion of the variable in this dimension. In other words, the closer the locus to the circumference of the circle, the more important its contribution quality degree to PCA and the more effective it is in distinguishing these populations.

In **Figure 3**, the PCA distribution plots contained all individuals (sample size = 2813) from the CTQ and CTT groups and the other five different geographical regions. The first three PCs interpreted 13.8% of the total variation. The PC1

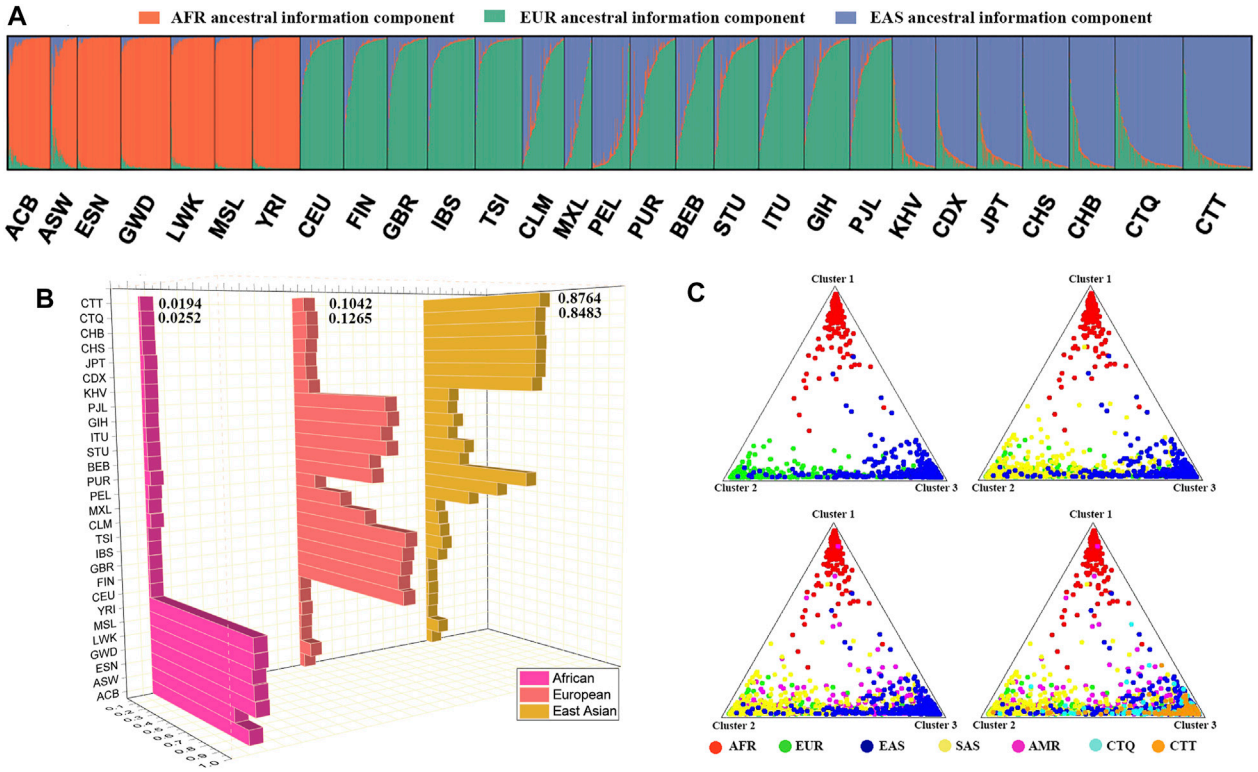


FIGURE 4 | (A) Inferred population genetic structure by STRUCTURE analysis under a model with three ancestral components ($K = 3$). **(B)** Ancestral composition proportions of these 28 populations at $K = 3$; **(C)** triangle plots for individual ancestry estimations among the CTQ and CTT groups, and other five different geographic region populations. Dots with diverse colors represent individuals from different continental regions. EAS, East Asian; SAS, South Asian; AMR, American; EUR, European; AFR, African.

(8.5%) and PC2 (3%) plots (**Figure 3A**), and PC1 and PC3 (2.3%) plots (**Figure 3B**) significantly separated the African individuals and the individuals from other geographic regions. The PC2 and PC3 plots (**Figure 3C**) could not distinguish individuals from all geographic regions. Individuals from the CTQ and CTT groups tended to cluster together with individuals from East Asia in these three PCA results.

3.6 Population Genetic Architectures

The STRUCTURE diagram via the Bayesian algorithm displayed the ancestral information components of all individuals. The predefined ancestral information components (K = 2–7) were distinguished by different colors (**Supplementary Figure S4**). The orange and green ancestral compositions were extracted at K = 2. The orange ancestral component was the African population, ranging from 98.39% (ESN) to 83.07% (African Ancestry in Southwest US, ASW) in the African seven populations. The predefined appropriate ancestral information component was three, which was determined by the online Structure Harvester program (**Supplementary Figure S5**). Herein, green, orange, and violet colors were defined as African, European, and East Asian ancestral compositions, respectively (**Figure 4**), and their proportions spanned from 97% (ESN) to 79.29% (ASW), from 89.47% (Toscani in Italy, TSI) to 84.01% (FIN), and from 87.64% (CTT) to 82.58% (CHB). The South Asian and American populations were considered mixed groups composing mainly of East Asian and European ancestral information components. The East Asian ancestral component proportion had the highest values in the CTQ and CTT groups, and the proportion values were 84.83% and 87.64%, respectively. The rest included the small amounts of European ancestral ingredient (12.65% in CTQ, 10.42% in CTT) and the negligible African ancestral composition (2.52% in CTQ, 1.94% in CTT). With the further addition of the K values, other ancestral compositions from South Asia and America regions manifested continuously. When defining five ancestral populations, these ancestral components corresponded to five different geographical regions one by one. The main ancestral component ratios of African, European, East Asian, South Asian populations ranged from 85.36% (ESN) to 66.53% (ASW), from 74.68% (TSI) to 65.87% (GBR), from 68.85% (CTT) to 54.67% (CDX), and from 68.92% (Sri Lankan Tamil in the United Kingdom, STU) to 59.4% (Bengali in Bangladesh, BEB), whereas the American ancestral composition had the largest value in Peruvian in Lima, Peru (PEL, 68.96%), followed by the Mexican Ancestry in Los Angeles, California (MXL); Colombian in Medellin, Colombia (CLM); and Puerto Rican in Puerto Rico (PUR) populations (46.65%, 30.52%, and 21.56%, respectively). At present, the East Asian ancestral component proportions in the CTQ and CTT groups were 61.98% and 68.85%, respectively (**Supplementary Table S7**).

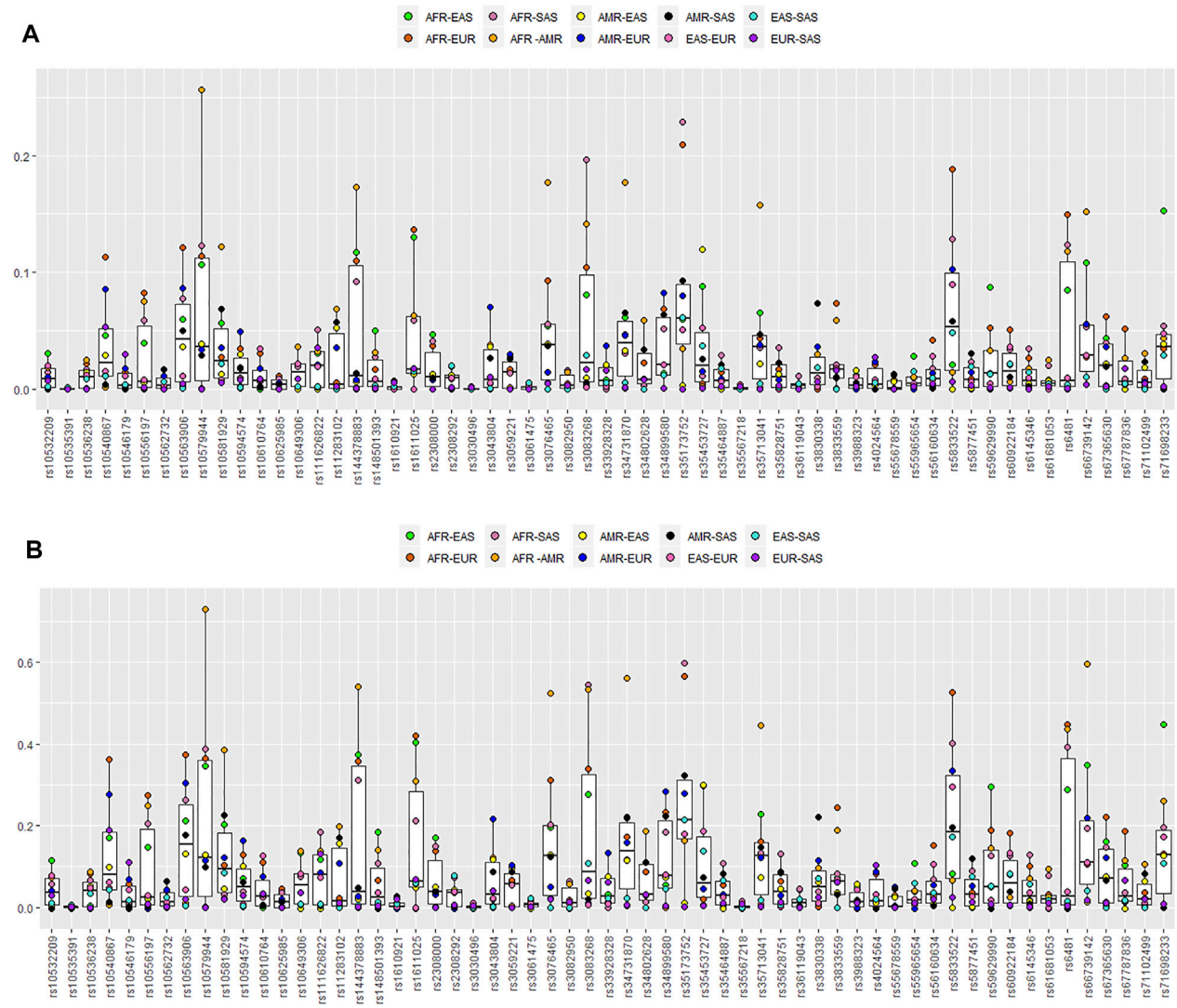


FIGURE 5 | Box plots of pairwise I_n and locus-by-locus of F_{ST} values. **(A)** Pairwise I_n values for the 59 au-DIPs among the five different geographical region populations. **(B)** Pairwise F_{ST} values for the 59 au-DIPs among the five different geographical region populations.

In the meantime, as shown in the triangle plot (**Figure 4C**), dots with diverse colors represented individuals from different intercontinental regions. These four triangle plots simulated the individual distributions from different geographic regions with the increase in the included populations. The first triangular plot included the African, European, and East Asian individuals, which were distributed in three different corners, representing the three main ancestral components. For the second triangle plot included African, European, East Asian, and South Asian populations, where South Asian individuals (yellow dots) were mainly distributed in the green and blue corners. Similarly, the AMR individuals were shown in the third triangle plot. The fourth triangle plot contained the African, European, East Asian, South Asian, and American individuals, as well as individuals from the CTQ and CTT groups. The results signified that the CTQ and CTT individuals were marked with turquoise and orange dots, respectively, and overlapped principally with the East Asian individuals, which were labeled with dark blue dots.

3.7 Mining Ancestry Informative Markers

3.7.1 Population-Specific Divergence

The PSD values of these 59 au-DIPs among the five geographical populations, Africa, Europe, East Asia, South Asia, and America (represented by the PEL population), were summarized in **Supplementary Table S8**. The accumulated PSD value of the 59 au-DIPs in distinguishing the populations from the five geographical regions was 3.0359. The maximum PSD was found at rs35173752 (PSD = 0.1855), rs10579944 (PSD = 0.1526), and then rs5833522 (PSD = 0.1509). The cumulative PSD values of these loci in African, European, American, South Asian, and East Asian populations were 2.7545, 1.0462, 1.3185,

0.6476, and 0.7029, respectively. These loci could assign 98.49% African individuals, 86.08% European individuals, 92.94%, American individuals, 79.55% South Asian individuals, and 90.08% East Asian individuals to the intercontinental regions where they originated from (**Supplementary Table S9**). The closer the individual was to the corner, the more accurately it could be assigned to its original geographic region (**Supplementary Figure S6**).

3.7.2 Informativeness for Assignment and *F*-Statistics

The I_n values and locus-by-locus of F_{ST} values of these 59 au-DIPs for pairs of regional populations are summarized in **Supplementary Table S10** and **Figure 5**. Herein, the threshold proposed for being a high-efficiency ancestry informative marker was the I_n value >0.131 (Rosenberg et al., 2003). The largest I_n value (I_n = 0.2563) was found in the African and American combination at the rs10579944 locus with the F_{ST} value of 0.7298. The rs144378883 (I_n = 0.1732), rs3076465 (I_n = 0.1767), rs3083268 (I_n = 0.1415), rs34731870 (I_n = 0.1770), rs35713041 (I_n = 0.1579), and rs66739142 (I_n = 0.1517) loci also showed higher I_n values (>0.131) with higher F_{ST} values of 0.5389, 0.5248, 0.5336, 0.5617, 0.4462, and 0.5957 between the African and American populations, respectively. In the same way, higher I_n values and F_{ST} values were also found at the rs71698233 (I_n = 0.1530, F_{ST} = 0.4481) locus between the African and East Asian populations and the rs1611025 (I_n = 0.1366, F_{ST} = 0.4189), rs35173752 (I_n = 0.2092, F_{ST} = 0.5650), rs5833522(I_n = 0.1822, F_{ST} = 0.5261), and rs6481 (I_n = 0.1495, F_{ST} = 0.4484) loci between the African and European populations and the rs3083268 (I_n = 0.1968, F_{ST} = 0.5454) and rs35173752 (I_n = 0.2283, F_{ST} = 0.5984) loci between the African and South Asian populations, respectively. These loci were thus considered as the promising marker to differentiate African and non-African populations.

4 DISCUSSION

4.1 Forensic Application Efficacy Evaluation of the Novel System

The genetic polymorphisms of these DIPs included in this new homemade panel and its forensic efficiency for personal identification and paternity testing should be validated in relevant populations. So far, these forensic parameters have not been evaluated in Tibetan groups. Here, 155 CTQ and 154 CTT samples were genotyped by this new panel, respectively. The LD analysis and HWE exact testing illustrated that all pairwise loci were independent of each other, and the samples in this survey were representative. The PD and PIC values of the autosomal loci ranged from 0.5126 to 0.9496, and 0.2885 to 0.8094; and 0.5351 to 0.9297, and 0.3006 to 0.8022 in the CTQ and CTT groups, respectively. The autosomal loci in this new panel were also reasonably informative (PIC >0.25) in the two Tibetan groups (Botstein et al., 1980). This new panel also provided substantially lower CPM (1.9253E-27 in CTQ, 1.5061E-26 in CTT) than those obtained from the previously developed systems incorporating 35 DIPs (5.5662E-15 in CTQ, 9.6907E-15 in CTT) (Liu et al., 2020), and including 50 DIPs (about E-19–E-20) (Wang et al., 2020), indicating that this new system was more efficient for individual identification in Tibetan groups. Additionally, the CPE values generated in this system were higher than 0.9999 (0.999997 in CTQ, 0.99999895 in CTT), which was an appreciable improvement compared with other existing DIP kits such as the kits containing 30 DIPs (Li et al., 2019), 35 DIPs (Liu et al., 2020), and 50 DIPs (Wang et al., 2020) with CPE values less than 0.9999.

4.2 Genetic Affinity Comparison and Population Structure Inquisition

A robust population genetic analysis is a prerequisite for understanding the genetic background of Chinese nationalities. To deeply determine the potential admixture level of the Tibetan groups and the genetic relationships with other reference populations, a comprehensive population genetic analysis was performed employing multitudinous bioinformatics analysis softwares, including STRAF, Arlequin, DISPAN program, Genepop, TBtools, PHYLIY, MEGA, Origin, STRUCTURE, *R*, CLUMPP, Distruct, Infocal, Snipper, and so on.

First, the allele frequency distributions of the CTQ and CTT groups were most similar to other populations in the East Asia region, and these seven East Asian populations aggregated together in the different subordinate branches of East Asia in the allele frequency heatmap. The main usage of genetic distance is to measure the genetic divergence among species or populations across the world, including the ancestral relationship or differentiation degree (Jakobsson, Edge, and Rosenberg 2013; Goswami and Dagla 2017). Different types of genetic distances have been developed, of which the F_{ST} value and D_A distance are still used to summarize genetic variations within and among populations. They have their own unique evolutionary and statistical properties, and the F_{ST} value is generally considered to be the population differentiation caused by the difference in genetic structure (Weir and Cockerham 1984; Jakobsson, Edge, and Rosenberg 2013), while the D_A distance is developed under the hypothesis that genetic differentiation originates from genetic drift and mutation event (Nei, Tajima, and Tateno 1983; Jin et al., 2019). In this study, genotyping data of the 59 au-DIPs were used to analyze the hereditary distances among these 28 populations. The obtained results were basically consistent with a previously published study (Liu et al., 2020) and demonstrated that the genetic relationship between the CTQ and CTT groups was the closest, and then the CTQ and CTT groups had relatively closer affinities with Han Chinese populations than other populations from East Asia. Unsurprising, given that linguistics (Lu et al., 2016) and archeological data (Yang et al., 2017) have already suggested that the Tibetan group and Han populations split from their shared ancestral population roughly 6,000–4,725 years ago, and increasing numbers of Han Chinese also began to settle down in Tibet or the northern part of the Tibetan Plateau (Qinghai Province) after the peaceful liberation of Tibet in 1951 (Moore et al., 2000). While the CTQ group had more genetic similarity with CHB than the CTT group, it was inevitably related to the special geographic location of Qinghai on the northeastern

margin of the Tibetan Plateau, which is a vital geographical corridor for population migrations and admixtures (Liu et al., 2021), and the CTQ group might thus gain more hereditary effects from other lowland populations.

Whereafter, the tree model is usually applied to explore the biological evolution history in the bioinformatics field, and genetic affinities among populations are generally displayed through phylogenetic trees (Takezaki and Nei 2008). Here, the populations from the same geographical regions were concentrated in the same branch (except for the American populations) by the phylogenetic reconstructions with different tree methods. The genetic affinities of the CTQ and CTT groups were significantly close and gathered first. The CTQ and CTT groups and other East Asian populations belonged to different sub-branches in the East Asian branch. PCA is used to extract important information by reducing the multivariate data dimensionality to two or three principal components, which can be visualized graphically (Lever, Krzywinski, and Altman 2017). We also conducted PCAs at both the population and individual levels among the two Tibetan groups and other global reference populations, these PCA plots reflected that the CTQ and CTT groups always converged with the subpopulations or individuals from the East Asia region under the first three PCs. Additionally, the correlation circle of PCA contribution quality offered the thought for the follow-up on the ancestry informative marker excavation. STRUCTURE analysis is applied using the model-based Bayesian iterative estimation algorithm to infer the origins of individuals with unknown population characteristics (Porras-Hurtado et al., 2013). Here, STRUCTURE analysis was ultimately carried out to infer the detailed genetic structures of the CTQ and CTT groups, and the result that the two Tibetan groups had the most similar genetic structures to the East Asian populations was confirmed once again, which was consistent with previous findings based on the other 35 au-DIPs. Finally, the three commonly used indicators, PSD, I_n, and locus-by-locus of F_{ST} values, were selected to determine the level of informativeness provided by these markers about distinguishing different continental populations. Results revealed that there were more loci that could differentiate between African populations and other non-African populations. These DIPs would be used as potential ancestral information markers to be further explored in future forensic application.

5 CONCLUSION

In this study, the forensic application efficacy of this new self-made six-color fluorescence multiplex PCR system was validated in the CTQ and CTT groups. The obtained results showed that the novel system could be applied well to the individual identification and paternity testing in two Tibetan groups, especially for forensic applications of degraded biological materials. Furthermore, combining the previous research achievements of these two Tibetan groups, the comprehensive overview for the genetic relatedness and population structure of the Chinese Tibetan groups was explored in-depth again, and these findings were of great significance to further reveal the genetic background and structure of Chinese Tibetan groups in different regions.

DATA AVAILABILITY STATEMENT

The datasets for this article are not publicly available due to concerns regarding participant anonymity. Requests to access the datasets should be directed to the corresponding author.

ETHICS STATEMENT

The study involving human participants were reviewed and approved by the ethical committee of the Xi'an Jiaotong University Health Science Center (approval number: 2019-1039). The participants provided their written informed consents to participate in this study.

AUTHOR CONTRIBUTIONS

YL conducted the experiment, performed the statistical analysis, wrote the original draft, and edited the manuscript; WC conducted some statistical analysis; KW and XZ participated in some experiments; WC, XJ and SM revised the manuscript; and BZ designed the work, provided the conception, and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant number: 81930055).

ACKNOWLEDGMENTS

We would like to thank the volunteers who contributed samples for this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.880346/full#supplementary-material

REFERENCES

Aldenderfer, M. (2011). Peopling the Tibetan Plateau: Insights from Archaeology. *High Alt. Med. Biol.* 12 (2), 141–147. doi:10.1089/ham.2010.1094

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *Am. J. Hum. Genet.* 32 (3), 314–331.

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* 13 (8), 1194–1202. doi:10.1016/j.molp.2020.06.009

Dakpa, T. (2014). Unique Aspect of Tibetan Medicine. *Acupunct. Electrother. Res.* 39 (1), 27–43. doi:10.3727/036012914x13966138791145

Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a Website and Program for Visualizing STRUCTURE Output and Implementing the Evanno Method. *Conserv. Genet. Resour.* 4 (2), 359–361. doi:10.1007/s12686-011-9548-7

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin Suite Ver 3.5: a New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. *Mol. Ecol. Resour.* 10 (3), 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Fan, G.-Y., Zhang, Z.-Q., Tang, P.-Z., Song, D.-L., Zheng, X.-K., Zhou, Y.-J., et al. (2021). Forensic and Phylogenetic Analyses of Populations in the Tibetan-Yi Corridor Using 41 Y-STRs. *Int. J. Leg. Med.* 135 (3), 783–785. doi:10.1007/s00414-020-02453-3

Fondevila, M., Phillips, C., Santos, C., Pereira, R., Gusmão, L., Carracedo, A., et al. (2012). Forensic Performance of Two Insertion-Deletion Marker Assays. *Int. J. Leg. Med.* 126 (5), 725–737. doi:10.1007/s00414-012-0721-7

Goswami, D., and Dagla, H. R. (2017). Standardization of DNA Extraction and Genetic Diversity Analysis of Haloxylon Salicornicum : An Underutilized Species of Extreme Arid Environment. *Plant gene.* 12, 66–71. doi:10.1016/j.plgene.2017.08.001

Gouy, A., and Zieger, M. (2017). STRAF-A Convenient Online Tool for STR Data Evaluation in Forensic Genetics. *Forensic Sci. Int. Genet.* 30, 148–151. doi:10.1016/j.fsigen.2017.07.007

Graham, E. A. M. (2005). Mini-STRs. *Fsmp* 1 (1), 065–068. doi:10.1385/fsmp:1:1:065

Gymrek, M. (2017). A Genomic View of Short Tandem Repeats. *Curr. Opin. Genet. Dev.* 44, 9–16. doi:10.1016/j.gde.2017.01.012

Hu, H., Petousi, N., Glusman, G., Yu, Y., Bohlender, R., Tashi, T., et al. (2017). Evolutionary History of Tibetans Inferred from Whole-Genome Sequencing. *PLoS Genet.* 13 (4), e1006675. doi:10.1371/journal.pgen.1006675

Jakobsson, M., Edge, M. D., and Rosenberg, N. A. (2013). The Relationship between FST and the Frequency of the Most Frequent Allele. *Genetics* 193 (2), 515–528. doi:10.1534/genetics.112.144758

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a Cluster Matching and Permutation Program for Dealing with Label Switching and Multimodality in Analysis of Population Structure. *Bioinformatics* 23 (14), 1801–1806. doi:10.1093/bioinformatics/btm233

Jin, X. Y., Wei, Y. Y., Cui, W., Chen, C., Guo, Y. X., Zhang, W. Q., et al. (2019). Development of a Novel Multiplex Polymerase Chain Reaction System for Forensic Individual Identification Using Insertion/deletion Polymorphisms. *Electrophoresis* 40 (12-13), 1691–1698. doi:10.1002/elps.201800412

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33 (7), 1870–1874. doi:10.1093/molbev/msw054

Kun, T. J., Wictum, E. J., and Penedo, M. C. T. (2018). A Mini-STR Typing System for Degraded Equine DNA. *Anim. Genet.* 49 (5), 464–466. doi:10.1111/age.12716

Lever, J., Krzywinski, M., and Altman, N. (2017). Principal Component Analysis. *Nat. Methods* 14 (7), 641–642. doi:10.1038/nmeth.4346

Li, J., Guo, X., and Wang, Y. (2012). Temperature Rating Prediction of Tibetan Robe Ensemble Based on Different Wearing Ways. *Appl. Ergon.* 43 (5), 909–915. doi:10.1016/j.apergo.2011.12.015

Li, L., Ye, Y., Song, F., Wang, Z., and Hou, Y. (2019). Genetic Structure and Forensic Parameters of 30 InDels for Human Identification Purposes in 10 Tibetan Populations of China. *Forensic Sci. Int. Genet.* 40, e219–e227. doi:10.1016/j.fsigen.2019.02.002

Liu, Y., Jin, X., Mei, S., Xu, H., Zhao, C., Lan, Q., et al. (2020). Insights into the Genetic Characteristics and Population Structures of Chinese Two Tibetan Groups Using 35 Insertion/deletion Polymorphic Loci. *Mol. Genet. Genomics* 295 (4), 957–968. doi:10.1007/s00438-020-01670-0

Liu, Y., Mei, S., Jin, X., Zhao, M., and Zhu, B. (2022). Independent Development and Validation of a Novel Six-Color Fluorescence Multiplex Panel Including 61 Diallelic DIPs and 2 miniSTRs for Forensic Degradation Sample. *Electrophoresis.*

Liu, Y., Wang, M., Chen, P., Wang, Z., Liu, J., Yao, L., et al. (2021). Combined Low-/High-Density Modern and Ancient Genome-wide Data Document Genomic Admixture History of High-Altitude East Asians. *Front. Genet.* 12, 582357. doi:10.3389/fgene.2021.582357

Lu, D., Lou, H., Yuan, K., Wang, X., Wang, Y., Zhang, C., et al. (2016). Ancestral Origins and Genetic History of Tibetan Highlanders. *Am. J. Hum. Genet.* 99 (3), 580–594. doi:10.1016/j.ajhg.2016.07.002

Moore, L. G., Fernando Armaza, V., Villena, M., and Vargas, E. (2002). Comparative Aspects of High-Altitude Adaptation in Human Populations. *Adv. Exp. Med. Biol.* 475, 45–62. doi:10.1007/0-306-46825-5_6

Nei, M., Tajima, F., and Tateno, Y. (1983). Accuracy of Estimated Phylogenetic Trees from Molecular Data. *J. Mol. Evol.* 19 (2), 153–170. doi:10.1007/bf02300753

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., and Lareu, M. V. (2013). An Overview of STRUCTURE: Applications, Parameter Settings, and Supporting Software. *Front. Genet.* 4, 98. doi:10.3389/fgene.2013.00098

Rosenberg, N. A. (2004). Distruct: a Program for the Graphical Display of Population Structure. *Mol. Ecol. Notes* 4, 137–138. doi:10.1046/j.1471-8286.2003.00566.x

Rosenberg, N. A., Li, L. M., Ward, R., and Pritchard, J. K. (2003). Informativeness of Genetic Markers for Inference of Ancestry*. *Am. J. Hum. Genet.* 73 (6), 1402–1422. doi:10.1086/380416

Sagart, L., Jacques, G., Lai, Y., Ryder, R. J., Thouzeau, V., Greenhill, S. J., et al. (2019). Dated Language Phylogenies Shed Light on the Ancestry of Sino-Tibetan. *Proc. Natl. Acad. Sci. U.S.A.* 116 (21), 10317–10322. doi:10.1073/pnas.1817972116

Senge, T., Madea, B., Junge, A., Rothschild, M. A., and Schneider, P. M. (2011). STRs, Mini STRs and SNPs - A Comparative Study for Typing Degraded DNA. *Leg. Med.* 13 (2), 68–74. doi:10.1016/j.legalmed.2010.12.001

Shimada, M. K., and Nishida, T. (2017). A Modification of the PHYLIP Program: A Solution for the Redundant Cluster Problem, and an Implementation of an Automatic Bootstrapping on Trees Inferred from Original Data. *Mol. Phylogenetics Evol.* 109, 409–414. doi:10.1016/j.ympev.2017.02.012

Takezaki, N., and Nei, M. (2008). Empirical Tests of the Reliability of Phylogenetic Trees Constructed with Microsatellite DNA. *Genetics* 178 (1), 385–392. doi:10.1534/genetics.107.081505

Wang, M., Du, W., He, G., Wang, S., Zou, X., Liu, J., et al. (2020). Revisiting the Genetic Background and Phylogenetic Structure of Five Sino-Tibetan-speaking Populations: Insights from Autosomal InDels. *Mol. Genet. Genomics* 295 (4), 969–979. doi:10.1007/s00438-020-01673-x

Weir, B. S., and Cockerham, C. C. (1984). Estimating F -Statistics for the Analysis of Population Structure. *Evolution* 38 (6), 1358–1370. doi:10.1111/j.1558-5646.1984.tb05657.x

Yang, J., Jin, Z.-B., Chen, J., Huang, X.-F., Li, X.-M., Liang, Y.-B., et al. (2017). Genetic Signatures of High-Altitude Adaptation in Tibetans. *Proc. Natl. Acad. Sci. U.S.A.* 114 (16), 4189–4194. doi:10.1073/pnas.1617042114

Conflict of Interest: Authors KW and XZ are employed by Ningbo Health Gene Technologies Co., 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.

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

Copyright © 2022 Liu, Cui, Jin, Wang, Mei, Zheng and Zhu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 31 May 2022
doi: 10.3389/fgene.2022.897650

Developmental Validation of the Novel Five-Dye-Labeled Multiplex Autosomal STR Panel and Its Forensic Efficiency Evaluation

*Shimei Huang[1†], Xiaoye Jin[1†], Hongling Zhang[1], Haiying Jin[2], Zheng Ren[1], Qiyan Wang[1], Yubo Liu[1], Jingyan Ji[1], Meiqing Yang[1], Han Zhang[1], Xingkai Zheng[2], Danlu Song[2], Bingjie Zheng[2] and Jiang Huang[1]**

[1]Department of Forensic Medicine, Guizhou Medical University, Guiyang, China, [2]Ningbo Health Gene Technologies Co.,Ltd, Ningbo, China

Short tandem repeats (STRs) are the most frequently used genetic markers in forensic genetics due to their high genetic diversities and abundant distributions in the human genome. Currently, the combined DNA index system is commonly incorporated into various commercial kits for forensic research. Some novel STRs that are different from the combined DNA index system were not only used to assess complex paternity cases but also could provide more genetic information and higher forensic efficiency in combination with those commonly used STRs. In this study, we validated forensic performance of a novel multiplex amplification STR panel to evaluate its sensitivity, species specificity, forensic application values, and so on. Obtained results revealed that the kit showed high sensitivity, and the complete allelic profile could be observed at 0.125 ng DNA sample. In addition, the kit possessed high species specificity, good tolerance to common inhibitors, and accurate genotyping ability. More importantly, STRs out of the kit displayed high discrimination power and probability of exclusion. To sum up, the novel kit presented in this study can be viewed as a promising tool for forensic human identification and complex paternity analysis.

Keywords: STRs, CODIS, forensic research, developmental validation, complex paternity analysis

OPEN ACCESS

Edited by:
Ryan Lan-Hai Wei,
Inner Mongolia Normal University, China

Reviewed by:
Zheng Wang,
Sichuan University, China
Xue-Ling Ou,
Sun Yat-sen University, China

****Correspondence:***
Jiang Huang
mmm_hj@126.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *16 March 2022*
Accepted: *21 April 2022*
Published: *31 May 2022*

Citation:
Huang S, Jin X, Zhang H, Jin H, Ren Z, Wang Q, Liu Y, Ji J, Yang M, Zhang H, Zheng X, Song D, Zheng B and Huang J (2022) Developmental Validation of the Novel Five-Dye-Labeled Multiplex Autosomal STR Panel and Its Forensic Efficiency Evaluation. Front. Genet. 13:897650. doi: 10.3389/fgene.2022.897650

INTRODUCTION

Short tandem repeats (STRs), also known as microsatellites, are repeat sequences of 2–6 bp nucleotides and common genetic variants in the human genome (Edwards et al., 1992). STRs are also viewed as gold-standard genetic markers for forensic identity testing and parentage analysis owing to their high diversities and wide distributions in the human genome (Cheng et al., 2021; Vullo et al., 2021; Yang et al., 2021; Zhang et al., 2021). In 1998, Budowle et al. (1998) proposed a combined DNA index system (CODIS) that included 13 core STRs. Subsequently, a large number of DNA databases consisting of these 13 STRs were developed to aid in identifying suspects related to criminal cases. One potential problem is that adventitious matches of DNA typing may occur with the increase of DNA databases. In 2015, Hares (2015) selected seven additional STRs and added them to the original CODIS; they stated that the expanded CODIS could provide high discrimination power (PD) and reduce falsely matching rates of suspects. In the meantime, most of these STRs are also integrated into some commercial kits (Qu et al., 2021; Yin et al., 2021; Zhang et al., 2021). However,

previous studies found that some STRs exhibited relatively low genetic diversities that went against forensic individual identification (Xie et al., 2015; Xiao et al., 2016; Tan et al., 2017). Recently, forensic researchers screened some novel STRs that exhibited high genetic polymorphisms in Chinese populations (Zhu et al., 2015; Li et al., 2017). On the one hand, these novel STRs could possess high cumulative PD. On the other hand, they could be used as additional loci for paternity testing when mutations of CODIS loci occur. More importantly, these novel STRs are also good for analyzing complex kinships like half-siblings.

In this study, 1 amelogenin gene and 26 autosomal STRs (D10S1248, D10S1435, D11S2368, D12S391, D13S325, D14S1434, D15S659, D16S539, D17S1301, D18S1364, D19S253, D1S1656, D20S482, D21S2055, D22GATA198B05, D22S1045, D2S441, D3S1744, D3S3045, D4S2366, D5S2800, D6S474, D6S477, D7S3048, D8S1132, and D9S1122) were developed into a multiplex panel named STRtyper-27 comp kit (HEALTH Gene Technologies, Zhejiang, China). Most of STRs in the novel kit are not overlapped with the expanded CODIS loci. Therefore, the kit is expected to provide more genetic information in combination with the extant upgraded CODIS loci. We also conducted validation studies of the kit to evaluate its overall performance based on the guideline of the Scientific Working Group on DNA Analysis Methods (https://1ecb9588-ea6f-4feb-971a-73265dbf079c.filesusr.com/ugd/4344b0_813b241e8944497e99b9c45b163b76bd.pdf). Furthermore, genetic distribution and forensic application value of the kit were assessed in the Guizhou Han population.

MATERIAL AND METHODS

Sample Information

We collected 312 bloodstain samples from unrelated healthy Guizhou Han individuals after obtaining their written informed consent. The 9948 and 9947A positive samples (1 ng/μL) were obtained from Promega Corporation (WI, United States). DNA samples of common species including dog, pig, cow, sheep, chicken, mouse, rabbit, fish, and colibacillus were collected from the Animal Laboratory Center of Guizhou Medical University to assess species specificity of the STRtyper-27 comp kit. This research was performed in line with the guidelines of Guizhou Medical University and warranted by the Ethic Commission of Guizhou Medical University.

DNA Amplification, Electrophoresis, and STR Typing

DNA sample of 1 ng was used to conduct the multiplex PCR of 27 loci according to the following specification. First, we prepared 10 μL PCR cocktail comprising 5 μL STRtyper-27 comp Master Mix, 2.5 μL STRtyper-27 comp Primer Mix, 2.5 μL ddH_2O, and 1 ng DNA sample. Second, PCR was conducted on the GeneAmp PCR System 9700 (Applied Biosystems, Foster City, CA, United States) under reaction conditions of initial denaturation at 95°C for 5 min; 28 cycles of 94°C for 10s, 61°C for 60s, and 70°C for 30 s; and 60°C for 15 min. Third, we mixed 1 μL amplified product/STRtyper-27 comp Allelic Ladder Mix with 8.75 μL deionized HiDi Formamide and 0.25 μL ILS-500 (HEALTH Gene Technologies) and then denatured the mixture at 95°C for 3 min, followed by chilling at 4°C for 3 min. Finally, the mixture was electrophoresed and separated by the 3500xL Genetic Analyzer (Thermo Fisher Scientific). STR typing of each locus was determined by the GeneMapper® ID-X Software v1.5 (Thermo Fisher Scientific) in comparison with the allelic ladder.

PCR Condition Studies

Different annealing temperatures (56, 57, 58, 59, 60, 61, 62, 63, 64, 65, and 66°C), extension temperatures (65, 66, 67, 68, 69, 70, 71, 72, 73, 74, and 75°C), final extension time (5, 10, 15, 20, 25, and 30min), cycle numbers (25, 26, 27, 28, 29, and 30), and reaction volumes (5, 10, 15, 20, and 25 μL) were set to assess the performance of the kit in various PCR conditions. In addition, we also adjusted the concentrations of master mix (0.7×, 0.8×, 0.9×, 1.0×, 1.1×, 1.2×, and 1.3×) and primer mix (0.7×, 0.8×, 0.9×, 1.0×, 1.1×, 1.2×, and 1.3×) to explore the robustness of the kit for reagent fluctuations. The aforementioned experiments were performed by the 9948 DNA sample by adjusting the testing condition.

Species Specificity and Sensitivity

Cross-reaction of the kit with non-human samples were evaluated by amplifying DNA samples of dog, pig, cow, sheep, chicken, mouse, rabbit, fish, and colibacillus.

The 9948 DNA sample was serially diluted to explore the detection lower limit of the kit: 1ng, 500, 250, 125, and 62.5 pg/μL. In addition, we also assessed the detection upper limit of the kit: 1, 2, 5, and 10 ng/μL.

Size Precision and Mixture Studies

To evaluate size precision of the kit, we injected and detected the STRtyper-27 comp Allelic Ladder Mix by the 3500xL Genetic Analyzer 24 times.

Different ratios of 9948 and 9947A mixtures (1:1, 1:2, 1:4, 1:8, 1:19, 19:1, 8:1, 4:1, and 2:1) were constructed to assess the power of the kit to detect the mixture.

Stability Study

In total, seven inhibitors including heme (1.0, 1.2, 1.4, 1.6, and 1.8 mM), humic acid (0.4, 0.43, 0.45, and 0.47 mg), EDTA (1, 5, 10, 15, 20, and 25 mM), melanin (0.5, 0.6, 0.7, 0.8, and 0.9 mg), Ca^{2+} (16, 18, 20, 22, and 24 mM), and tannin (1, 3, 5, and 7 mg) were collected to evaluate the tolerance of the kit to these inhibitors.

Degraded and Case-Type Sample Studies

We used the ultraviolet (wave length: 254nm; power: 28W) to treat the positive DNA sample 9948 (1 ng/μL) to simulate the degraded sample at different time periods (0, 15, 30, 45, and 60 min). Then, these mocked samples were detected by the developed kit in triplicate. Here, the 50-relative fluorescence unit was used as the detecting threshold to determine the allele peak.

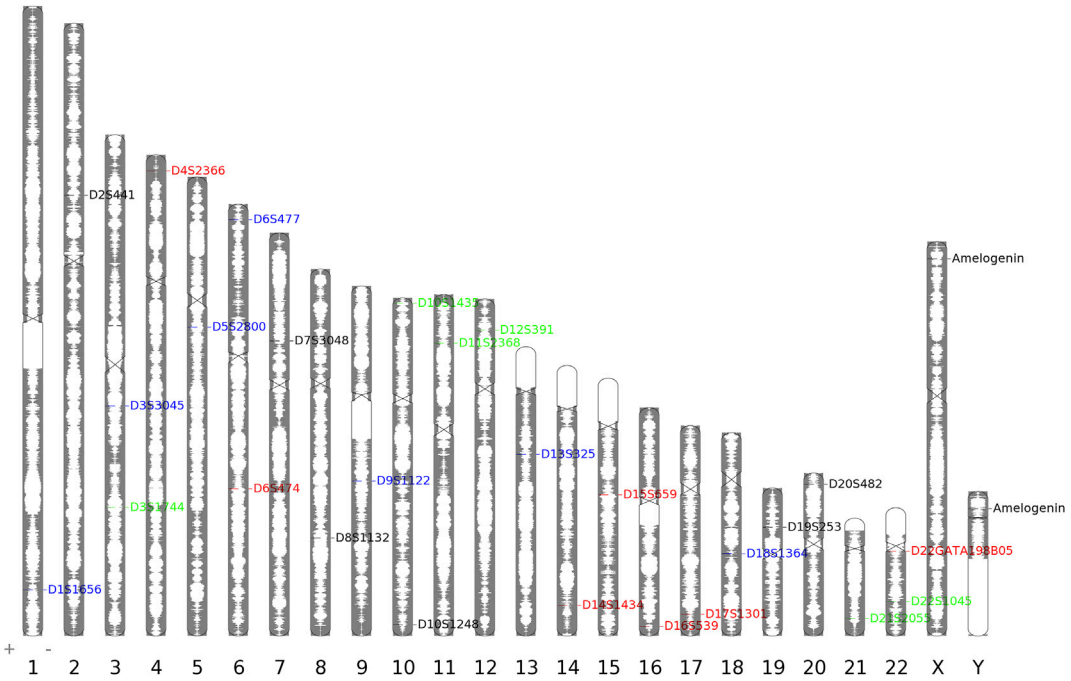


FIGURE 1 | Loci information of 26 autosomal STRs and the amelogenin gene.

Common samples found in the forensic scene including cigarette, bloodstain, seminal stain, fingerprint swab, and blood swab were collected and detected by the developed kit. First, we extracted DNA samples from these biomaterials by the ML ultrafine magnetic bead extraction kit (Changchun Bokun Biotech Corporation, Jilin, China). Next, these DNA samples were detected by the developed kit and the STRtyper-32G kit (HEALTH Gene Technologies).

Statistical Analysis

Allelic frequencies and forensic-related parameters including expected heterogeneity (He), observed heterogeneity (Ho), polymorphism information content (PIC), match probability (PM), PD, power of exclusion (PE), and typical paternity index (TPI) of 26 STR loci in the Guizhou Han population were estimated by the STRAF online tool v1.0.5 (Gouy and Zieger, 2017). Furthermore, Hardy–Weinberg equilibrium (HWE) and linkage disequilibrium (LD) analysis of these STRs in the Guizhou Han population were also assessed by the STRAF online tool v1.0.5.

RESULTS AND DISCUSSION

Loci Information

As shown in **Figure 1**, 27 loci of the STRtyper-27 comp kit were located on all chromosomes. These 27 loci were classified into four groups labeled by four dyes, respectively: D1S1656, D3S3045, D5S2800, D6S477, D9S1122, D13S325, and D18S1364 (FAM); D3S1744, D10S1435, D11S2368, D12S391, D21S2055, and D22S1045 (HEX); D4S2366, D6S474, D14S1434, D15S659, D16S539, D17S1301, and D22GATA198B05 (ROX); Amelogenin, D2S441, D7S3048, D8S1132, D10S1248, D19S253, and D20S482 (TAMRA). The allelic profile of 9947A positive sample is also given in **Figure 2**. The results showed that amplicon lengths of these loci distributed from 90 to 500 bp. Compared to other commercial STR kits (Hares, 2015; Zhu et al., 2015; Li et al., 2017; Ludeman et al., 2018; Zhong et al., 2019), we found that the kit in this study showed the most number (15) of overlapped loci using the Microreader 23SP kit (**Supplementary Table S1**). Even so, there were more than 10 novel STRs available in the developed kit. More importantly, the majority of loci presented in the kit were different from the expanded CODIS set. In addition, we found that physical distances between these novel STRs and those STRs (the expanded CODIS) on the same chromosomes were larger than 10 Mb (**Supplementary Table S2**), implying that these STRs could be viewed as independent loci from each other for forensic research. Even so, LD analyses of these STRs should be performed in the future. Anyway, we proposed that the kit in this study could be utilized as a high-efficient supplementary system for complex paternity analysis in parallel with the extant CODIS kits.

PCR-Based Studies

The annealing temperature is the key factor for PCR because it determines whether the primer binds to the DNA template. Therefore, annealing temperature variations may exert some effects on the performance of the multiplex detection assay. We assessed amplification efficiency of the developed kit at

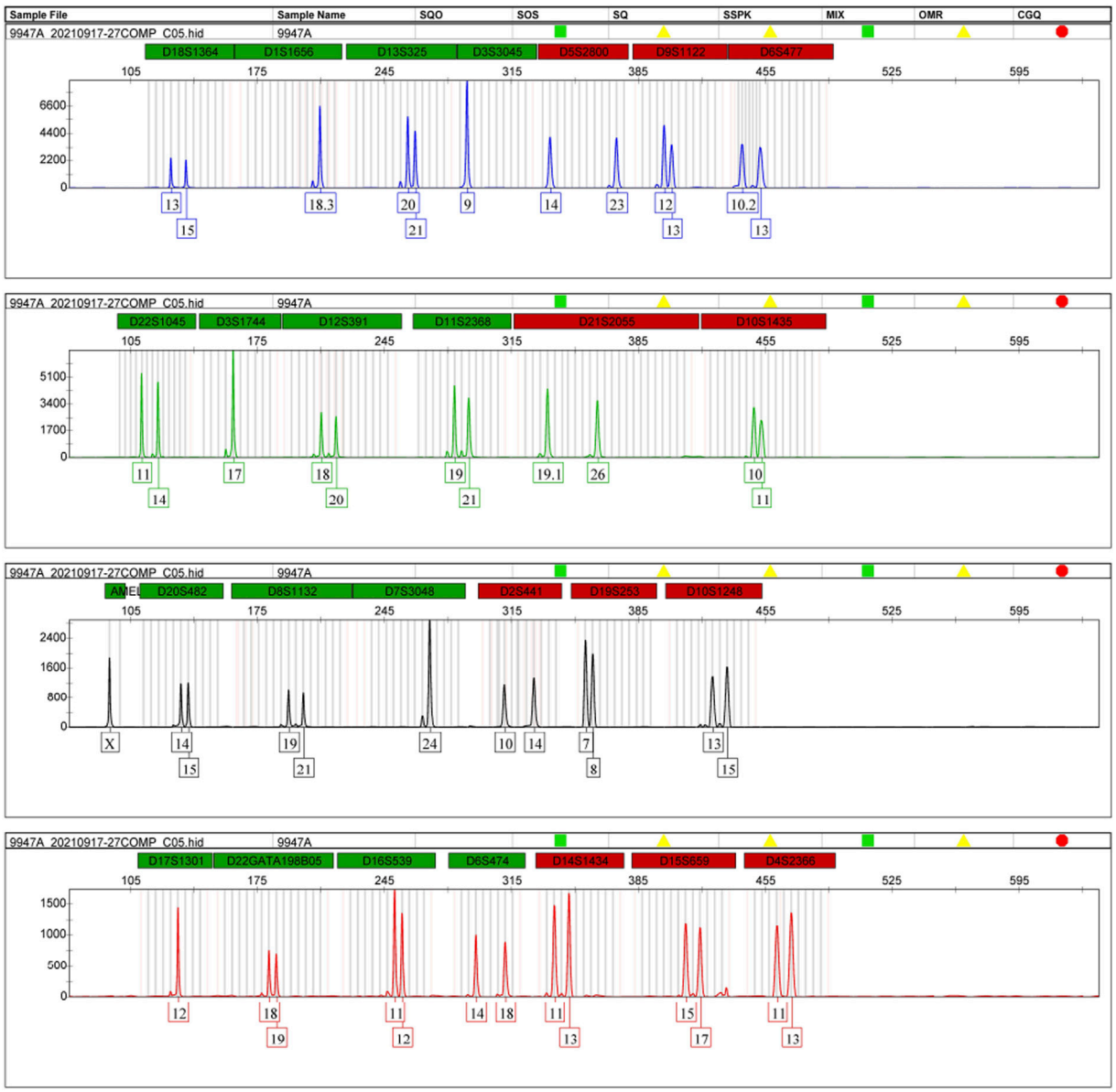


FIGURE 2 | Allelic profile of the positive sample 9947A by the STRtyper-27 comp kit.

different annealing temperatures, as shown in **Supplementary Figure S1**. We found that the kit displayed comparable amplification performance at 56–64°C. However, some alleles begun to drop out at 65°C. In addition, allele peak height showed significant decrease at 65 and 66°C. Consequently, researchers are not suggested to set higher annealing temperature than our recommended temperature in practical application.

For the extension temperature, it is also crucial for PCR performance of the STR kit because it is related to the DNA template extension reaction. We evaluated the influences of different extension temperatures on the amplification performance of the developed kit, as presented in **Supplementary Figure S2**. The results revealed that all alleles could be correctly typed at different extension temperatures. Thus, extension temperature variations did not show any negative effects on amplification efficiency of the STR system.

The final elongation reaction can be used to avoid non-template–depended adenylation that may give rise to minus A or shoulder peaks. Different final elongation times at 60°C were set to evaluate amplification performance of the developed kit, as given in **Supplementary Figure S3**. The results demonstrated that all alleles showed normal electrophoretic peaks at different elongation times, indicating that the kit was tolerant to elongation time variations.

Different cycle numbers were tested to explore the optimal condition for the developed kit. As shown in **Supplementary**

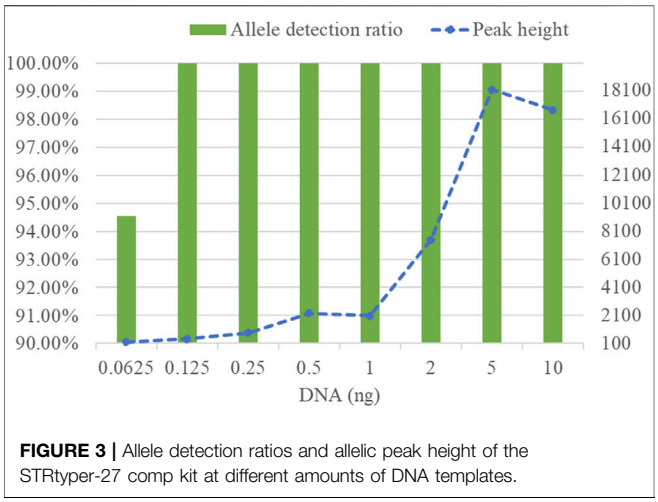


FIGURE 3 | Allele detection ratios and allelic peak height of the STRtyper-27 comp kit at different amounts of DNA templates.

Figure S4, all alleles could be observed at different cycle numbers. In addition, allele peak height gradually increased with the augment of cycle numbers. Some non-specific amplification products were also observed at higher cycle numbers. Given that more balanced peak height was seen at 29 cycle numbers, we suggested that 29 is the optimal cycle number. Even so, researchers may explore the best condition for samples of interest in their studies.

Primer, Taq DNA polymerase, and PCR buffer are indispensable components in PCR. The fluctuations of PCR reagents may occur due to pipetting errors, which may have negative impacts on amplification efficiency. A series of concentrations of the primer mix and master mix comprising Taq DNA polymerase, PCR buffer, and other essential components were tested to evaluate the robustness of the developed kit. As shown in **Supplementary Figure S5**, a full allelic profile could be observed at different concentrations of the primer mix. Moreover, more balanced peak heights among different alleles were observed at 1.0× primer mix. For different concentrations of the master mix, we also found that all alleles could be detected (**Supplementary Figure S6**). However, some noise peaks were also observed at higher concentrations of the master mix. Thus, we recommended 1.0× master mix as the optimal concentration.

In forensic practice, researchers may reduce PCR reaction volume for trace samples. Thus, we also assessed the impact of different reaction volumes on the amplification performance of the developed kit. As given in **Supplementary Figure S7**, a full allelic profile could be obtained at different reaction volumes. In addition, allele peak height decreased with the increase of reaction volume. From the aforementioned results, we proposed that the developed kit is robust for different reaction volumes.

Sensitivity Studies

To determine detection limit of the developed kit, different quantities of DNA samples was amplified by the kit. Obtained results revealed that the complete allelic profile could be obtained from these diluted DNA samples except for 0.0625 ng DNA sample of which nearly 6% of alleles dropped out (**Figure 3**). In addition, as DNA quantity increased, allele peak height also gradually rose. Consequently, the developed kit is not recommended to detect those samples in which the amount of DNA was less than 0.0625 ng.

Stability Studies

To evaluate the tolerance of the developed kit to common inhibitors, we added different concentrations of inhibitors to the PCR reagents. Obtained results are given in **Figure 4**; **Supplementary Figures 8–13**. For heme, we found all loci could be detected at 1.0 mM. In addition, alleles of some loci began to drop out at larger concentrations of heme, especially for 1.4–1.8 mM. For tannin, nearly 30% of 27 loci missed at 3mg, and more loci dropped out at 5–7 mg. For humic acid, we found that the majority of 27 loci could be observed at different concentrations of humic acid. Similar results could be seen from different concentrations of EDTA. For melanin and Ca^{2+}, the loci detection rate of the kit decreased with the increasing of melanin and Ca^{2+} concentrations. Overall, we stated that the kit performed relatively good tolerances for these inhibitors.

Size Precision

Allele size variations may occur between different runs even on the same equipment, which affects reliable and accurate typing. Accordingly, it is vital to evaluate size precision of the kit. As shown in **Figure 5**, standard deviations of all alleles ranged from 0.02 to 0.08, indicating relatively subtle size variations of the kit. Thus, we proposed that the kit could provide accurate and reliable allelic typing.

Mixture Analysis

DNA mixtures are common biological samples in forensic research. Consequently, it is necessary to assess the efficiency of the developed kit to detect mixtures. Two positive samples (9948 and 9947A) were mixed at different ratios (1:1, 1:2, 1:4, 1:8, 1:19, 19:1, 8:1, 4:1, and 2:1) and detected by the developed kit in triplicate. The allelic profile of 9948 and 9947A samples are presented in **Supplementary Table S3**. We found that nearly all alleles could be observed at different mixed ratios (**Figure 6** and **Supplementary Figure S14**). Even so, one allele of D6S477 locus dropped out when mixed ratios were 1:4, 1:8, 1:19, 2:1, 8:1, and 19:1. One allele of D10S1435 locus was also missing at 1:19 ratio. Moreover, an extra allele was observed for D6S477 locus at 19:1 ratio. Given that the extra allele was less than one repeat unit than targeted alleles, we postulated that it might be stutter peaks. Anyway, we proposed that the developed kit could be employed to dissolve mixtures of two individuals given that alleles of most loci could be detected at different ratios.

Species Specificity

Non-human DNA may present in the forensic scene. It is critical to evaluate species specificity of the developed kit. As shown in **Supplementary Figure S15**, no allele peaks were seen at nine common species including dog, pig, cow, sheep, chicken, mouse, rabbit, fish, and colibacillus, suggesting that the kit was human-specific and could be used to detect human samples without the interfering from other non-human samples.

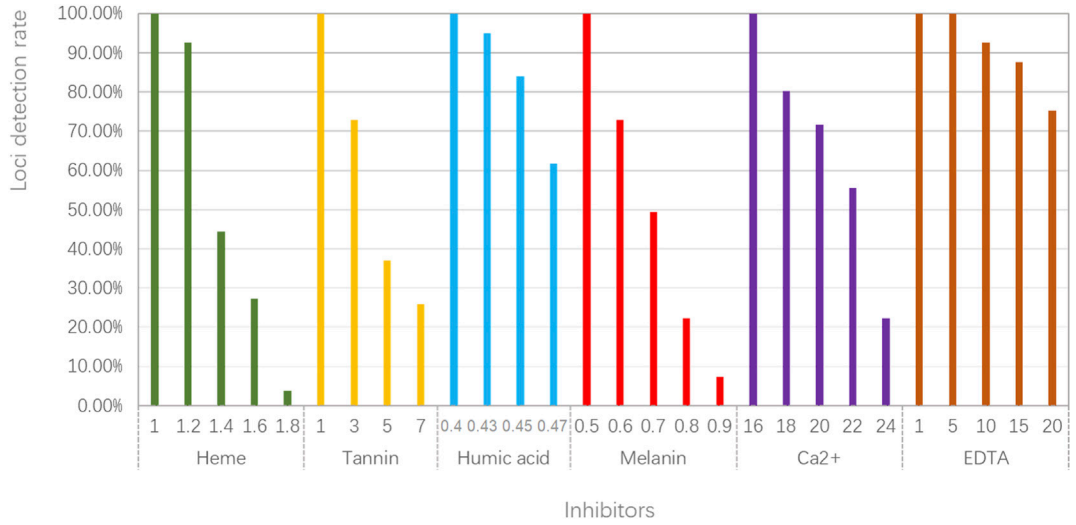


FIGURE 4 | Stability studies of the STRtyper-27 comp kit for six common inhibitors.

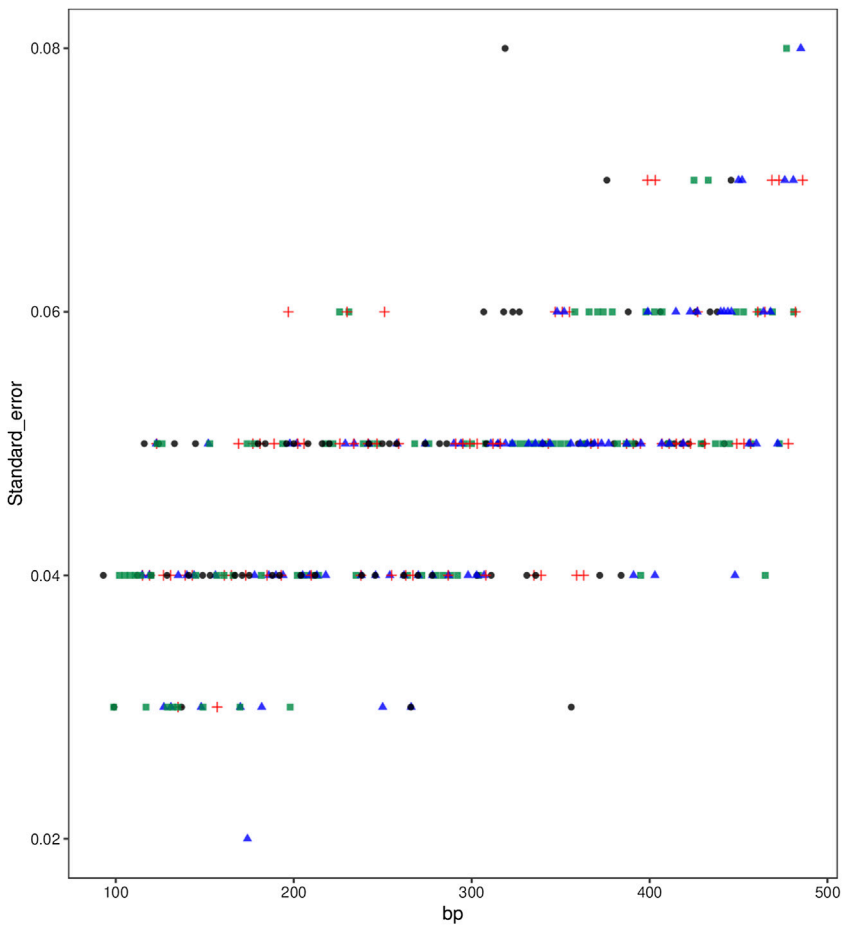


FIGURE 5 | Size precision of the STRtyper-27 comp kit.

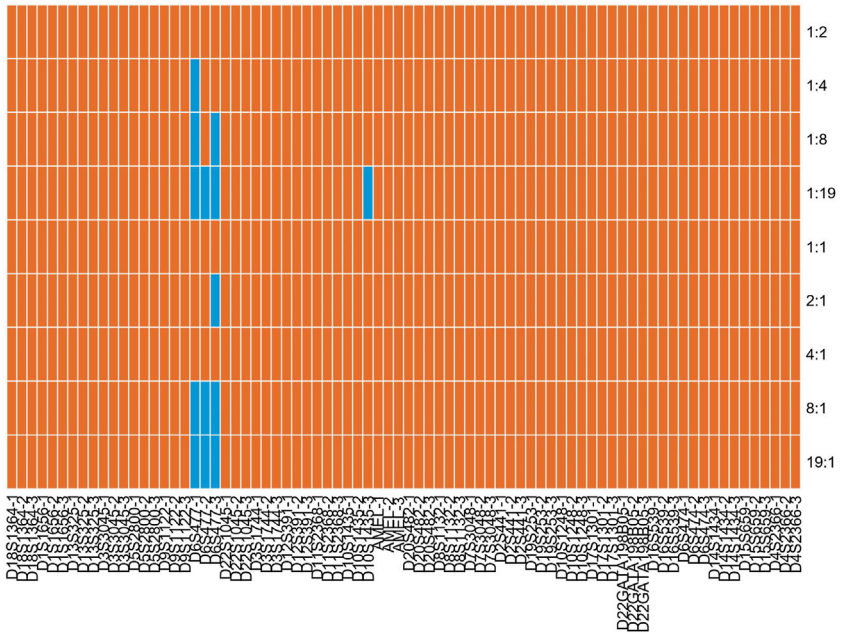


FIGURE 6 | Allele detection results of the STRtyper-27 comp kit for different ratios of 9948 and 9947A samples. Orange and blue blocks indicated complete and partial alleles detected, respectively.

TABLE 1 | Allele detection ratio and average peak height of 27 loci of the 9948 DNA sample exposed at different time periods of ultraviolet.

Time (min)	Allele detection ratio (%)	Average peak height
0	100.00	3,518
15	100.00	4,200
30	82.99	2,219
45	63.27	1762
60	28.57	171

Degraded and Case-Type Sample Studies

Allele detection results of 27 loci for mocked degraded samples are presented in **Table 1**; **Supplementary Figure S16**. A complete allelic profile of these 27 loci could be obtained from degraded samples at 15 min exposure time. When the exposure time increased to 30min, more than 80% alleles of these 27 loci could be detected. In addition, more than 60% alleles of these 27 loci were observed at 45 min exposure time. More importantly, approximately 30% alleles of these 27 loci were still obtained even though the exposure time increased to 60 min, indicating the kit displayed relatively good tolerance to these mocked degraded samples.

For five collected samples (cigarette, bloodstain, seminal stain, fingerprint swab, and blood swab), we found that all alleles of 26 STRs could be observed. Furthermore, the same allelic profile of 14 overlapping STRs between the developed kit and the STRtyper-32G kit were also discerned from these samples. Thus, we proposed that the developed kit was suitable for detecting these common case samples.

Population Studies

First, we conducted HWE and LD tests of 26 STRs in the Guizhou Han population, as given in Supplementary Tables 4–5. The results demonstrated that these loci did not deviate from HWE in the Guizhou Han population ($p < 0.05$). For LD of pairwise loci (**Supplementary Table S5**), we found that nine pairs deviated from linkage equilibrium after applying Bonferroni correction ($p < 0.000154$). There are some factors that may lead to LD of pairwise loci, like genetic linkage, mutation, genetic drift, and population inbreeding (Slatkin, 2008). However, these nine pairs that deviated from linkage equilibrium in the current study were located on different chromosomes. It was less likely that genetic linkage resulted in LD of these loci. Accordingly, it might be other reasons like mutation and genetic drift leading to LD of these nine pairwise loci, which needed to be further evaluated.

Next, we investigated allelic distributions and forensic parameters of 26 STRs in the Guizhou Han population, as shown in **Figure 7**; **Supplementary Table S4**. A total of 259 alleles were observed at these loci with the number of alleles per locus ranging from 6 (D6S474) to 21 (D21S2055). For forensic parameters, we found that these 26 loci displayed small PM values less than 0.15; their PE, PD, and TPI values ranged from 0.3740 to 0.7771, 0.8599 to 0.9727, and 1.4857 to 4.5882, respectively. For diversity indexes, the mean Ho, He, and PIC values were 0.7839, 0.7872, and 0.7549. In addition, PIC values of all loci were larger than 0.6, indicating that these loci possessed high genetic diversities in the Guizhou Han population. The cumulative PD and PE values of 26 STRs in the Guizhou Han population were 0.999 999 999 999 999 999 999 999 999 996961 and 0.999 999 999 90297, respectively. Compared to other commercial kits (Hares, 2015; Zhu et al., 2015; Li et al., 2017;

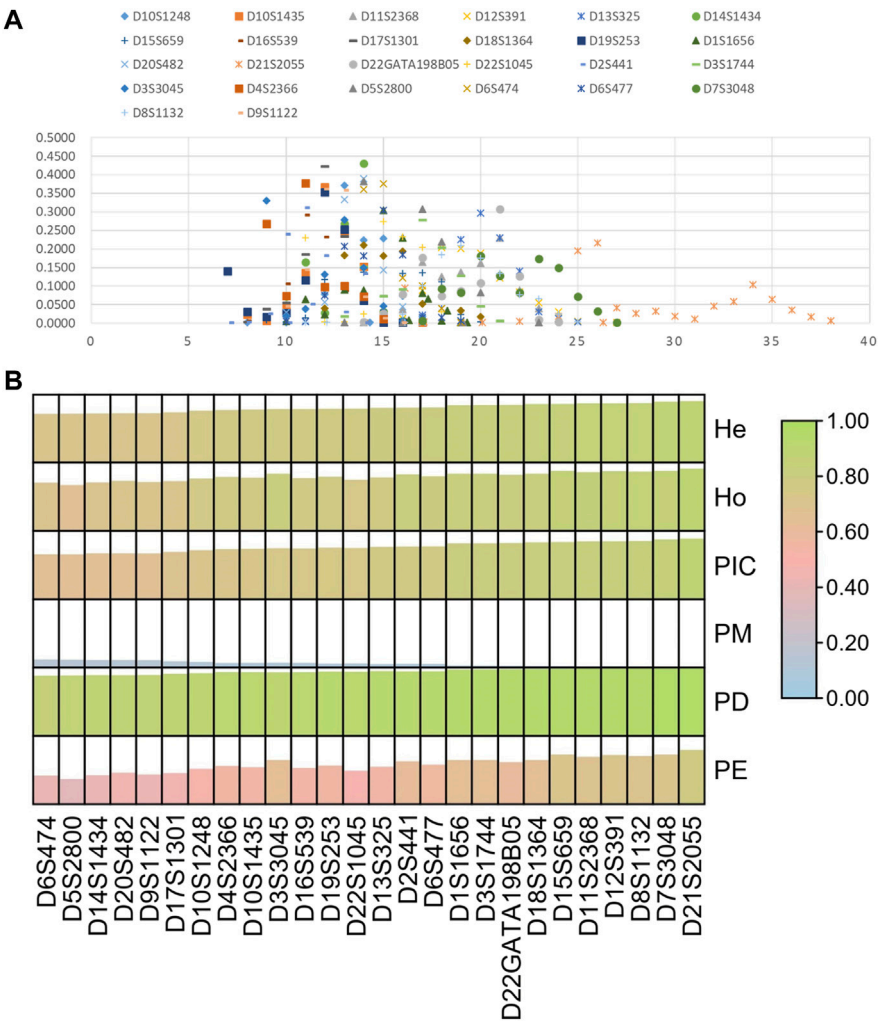


FIGURE 7 | Allelic frequencies **(A)** and forensic parameters **(B)** of 26 STRs in the Guizhou Han population.

Ludeman et al., 2018; Zhong et al., 2019), the developed kit could obtain better cumulative PD and PE values, implying that the kit could be viewed as a high-performance system for forensic identity testing and paternity analyses in the Guizhou Han population. Of note, most loci out of the kit were different from the expanded CODIS. Therefore, not only did these 26 STRs enhance discrimination efficiency for unrelated individuals by combining with the available commonly used STRs, but they could also be used to assess complex kinships.

CONCLUSION

In this study, we validated the performance of the novel kit according to the specification of the Scientific Working Group on DNA Analysis Methods. The kit showed good species specificity, high sensitivity, and tolerance to six common inhibitors. In addition, the kit possessed good compatibility for the variations of PCR reagents and PCR conditions. More importantly, the kit displayed high forensic application values for forensic human identification and paternity testing. In conclusion, the developed kit could be viewed as a valuable tool for forensic research.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethic Commission of Guizhou Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SH and XJ wrote the main text. HZ, ZR, and QW conducted data analysis. YL and JJ conducted statistical analysis. MY and HZ collected samples. SH, XJ, HJ, XZ, DS, and BZ performed the experiment. JH designed the work and provided the conception.

FUNDING

This study was supported by the Guizhou Provincial Science and Technology Projects (ZK (2022) General 355); Guizhou Education Department Young Scientific and Technical Talents Project (Qian Education KY (2022)215); Guizhou Scientific Support Project, Qian Science Support (2021) General 448; Shanghai Key Lab of Forensic Medicine, Key Lab of Forensic Science, Ministry of Justice, China (Academy of Forensic Science); Open Project, KF202009; Guizhou Province Education Department, Characteristic Region Project, Qian Education KY No. (2021)065; Guizhou "Hundred" High-level Innovative Talent Project, Qian Science Platform Talents (2020) 6012; Guizhou Scientific Support Project, Qian Science Support (2020)4Y057; Guizhou Science Project, Qian Science Foundation (2020) 1Y353; Guizhou Scientific Support Project, Qian Science Support (2019)2825; Guizhou Scientific Cultivation Project, Qian Science Platform Talent (2018)5779-X; Guizhou Engineering Technology Research Center Project, Qian High-Tech of Development and Reform Commission No. (2016)1345; and Guizhou Innovation Training Program for College Students (2019)5200926.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.897650/full#supplementary-material

REFERENCES

Budowle, B., Budowle, B., Moretti, T. R., Moretti, T. R., Niezgoda, S. J., Niezgoda, S. J., et al. (1998). "CODIS and PCR-Based Short Tandem Repeat Loci: Law Enforcement Tools," in The Second European Symposium on Human Identification, 73–88. Available at: https://www.promega.com/~/media/files/resources/conference proceedings/ishi 02/oral presentations/17.pdf.

Cheng, J., Song, B., Fu, J., Zheng, X., He, T., and Fu, J. (2021). Genetic Polymorphism of 19 Autosomal STR Loci in the Yi Ethnic Minority of Liangshan Yi Autonomous Prefecture from Sichuan Province in China. *Sci. Rep.* 11. doi:10.1038/s41598-021-95883-x

Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T., and Chakraborty, R. (1992). Genetic Variation at Five Trimeric and Tetrameric Tandem Repeat Loci in Four Human Population Groups. *Genomics* 12, 241–253. doi:10.1016/0888-7543(92)90371-X

Gouy, A., and Zieger, M. (2017). STRAF-A Convenient Online Tool for STR Data Evaluation in Forensic Genetics. *Forensic Sci. Int. Genet.* 30, 148–151. doi:10.1016/j.fsigen.2017.07.007

Hares, D. R. (2015). Selection and Implementation of Expanded CODIS Core Loci in the United States. *Forensic Sci. Int. Genet.* 17, 33–34. doi:10.1016/j.fsigen.2015.03.006

Li, J., Luo, H., Song, F., Zhang, L., Deng, C., Yu, Z., et al. (2017). Validation of the Microreader 23sp ID System: A New STR 23-plex System for Forensic Application. *Forensic Sci. Int. Genet.* 27, 67–73. doi:10.1016/j.fsigen.2016.12.005

Ludeman, M. J., Zhong, C., Mulero, J. J., Lagacé, R. E., Hennessy, L. K., Short, M. L., et al. (2018). Developmental Validation of GlobalFiler PCR Amplification Kit: a 6-dye Multiplex Assay Designed for Amplification of Casework Samples. *Int. J. Leg. Med.* 132, 1555–1573. doi:10.1007/s00414-018-1817-5

Qu, Y., Tao, R., Yu, H., Yang, Q., Wang, Z., Tan, R., et al. (2021). Development and Validation of a Forensic Six-dye Multiplex Assay with 29 STR Loci. *Electrophoresis* 42, 1419–1430. doi:10.1002/elps.202100019

Slatkin, M. (2008). Linkage Disequilibrium - Understanding the Evolutionary Past and Mapping the Medical Future. *Nat. Rev. Genet.* 9, 477–485. doi:10.1038/nrg2361

Tan, B., Zhao, Z., Zhang, Z., Li, S., and Li, S. C. (2017). Search for More Effective Microsatellite Markers for Forensics with Next-Generation Sequencing. *IEEE Trans.on Nanobioscience* 16, 375–381. doi:10.1109/TNB.2017.2712795

Vullo, C. M., Catelli, L., Ibarra Rodriguez, A. A., Papaioannou, A., Merino, J. C. Á., Lopez-Parra, A., et al. (2021). Second GHEP-ISFG Exercise for DVI: "DNA-Led" Victims' Identification in a Simulated Air Crash. *Forensic Sci. Int. Genet.* 53, 102527. doi:10.1016/j.fsigen.2021.102527

Xiao, C., Zhang, W., Wei, T., Pan, C., and Huang, D. (2016). Population Data of 21 Autosomal STR Loci in Chinese Han Population from Hubei Province in Central China. *Forensic Sci. Int. Genet.* 20, e13–e14. doi:10.1016/j.fsigen.2015.11.002

Xie, B., Chen, L., Yang, Y., Lv, Y., Chen, J., Shi, Y., et al. (2015). Genetic Distribution of 39 STR Loci in 1027 Unrelated Han Individuals from Northern China. *Forensic Sci. Int. Genet.* 19, 205–206. doi:10.1016/j.fsigen.2015.07.019

Yang, Q., Yao, Y., Shao, C., Zhou, Y., Li, H., Li, C., et al. (2021). Calculation of the Paternity Index for STR with Tri-allelic Patterns in Paternity Testing. *Forensic Sci. Int.* 324, 110832. doi:10.1016/j.forsciint.2021.110832

Yin, L., Zhu, J., Qu, S., Li, Y., Liu, Y., Yu, Z., et al. (2021). Validation of the Microreader 28A ID System: A 6-dye Multiplex Amplification Assay for Forensic Application. *Electrophoresis* 42, 1928–1935. doi:10.1002/elps.202100110

Zhang, K., Song, F., Wang, S., Wei, X., Gu, H., Xie, M., et al. (2021). Evaluation of the AGCU Expressmarker 30 Kit Composed of 31 Loci for Forensic Application. *Forensic Sci. Int.* 324, 110849. doi:10.1016/j.forsciint.2021.110849

Zhong, C., Gopinath, S., Norona, W., Ge, J., Lagacé, R. E., Wang, D. Y., et al. (2019). Developmental Validation of the Huaxia Platinum PCR Amplification Kit: A 6-dye Multiplex Direct Amplification Assay Designed for Chinese Reference Samples. *Forensic Sci. Int. Genet.* 42, 190–197. doi:10.1016/j.fsigen.2019.07.001

Zhu, B.-F., Zhang, Y.-D., Shen, C.-M., Du, W.-A., Liu, W.-J., Meng, H.-T., et al. (2015). Developmental Validation of the AGCU 21+1 STR Kit: A Novel Multiplex Assay for Forensic Application. *Electrophoresis* 36, 271–276. doi:10.1002/elps.201400333

Conflict of Interest: Authors HJ, XZ, DS, and BZ were employed by Ningbo Health Gene Technologies Co., Ltd.

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

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

Copyright © 2022 Huang, Jin, Zhang, Jin, Ren, Wang, Liu, Ji, Yang, Zhang, Zheng, Song, Zheng 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 | Frontiers in Genetics

ORIGINAL RESEARCH
published: 21 June 2022
doi: 10.3389/fgene.2022.919063

Genetic Diversity Analysis of the Chinese Daur Ethnic Group in Heilongjiang Province by Complete Mitochondrial Genome Sequencing

Mansha Jia[1†], *Qiuyan Li*[2,3,4†], *Tingting Zhang*[2,3], *Bonan Dong*[2,3], *Xiao Liang*[2,3], *Songbin Fu*[2,3*] *and Jingcui Yu*[1,3*]

[1]*Scientific Research Centre, The Second Affiliated Hospital of Harbin Medical University, Harbin, China,* [2]*Laboratory of Medical Genetics, Harbin Medical University, Harbin, China,* [3]*Key Laboratory of Preservation of Human Genetic Resources and Disease Control in China, Harbin Medical University, Ministry of Education, Harbin, China,* [4]*Editorial Department of International Journal of Genetics, Harbin Medical University, Harbin, China*

OPEN ACCESS

Edited by:
Guanglin He,
Nanyang Technological University, Singapore

Reviewed by:
Zheng Ren,
Guizhou Medical University, China
Varun Sharma,
NMC Healthcare (NMC Genetics), India

***Correspondence:**
Songbin Fu
fusb@ems.hrbmu.edu.cn
Jingcui Yu
yujingcui@ems.hrbmu.edu.cn

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *13 April 2022*
Accepted: *12 May 2022*
Published: *21 June 2022*

Citation:
Jia M, Li Q, Zhang T, Dong B, Liang X, Fu S and Yu J (2022) Genetic Diversity Analysis of the Chinese Daur Ethnic Group in Heilongjiang Province by Complete Mitochondrial Genome Sequencing. Front. Genet. 13:919063. doi: 10.3389/fgene.2022.919063

Mitochondrial DNA (mtDNA) has the characteristics of maternal inheritance, high mutation rate, high copy number, and no recombination**.** As the most powerful tool for studying the origin and evolution of modern humans, mtDNA has great significance in the research of population genetics and evolutionary genetics. Here, we provide new insights into the maternal genetic history of the Daur ethnic group by generating complete mitochondrial genomes from a total of 146 Daur individuals in China. We also collected the published complete mitochondrial genome sequences of 5,094 individuals from 56 worldwide populations as reference data to further explore the matrilineal genetic landscape of the Daur ethnic group. First, the haplotype diversity was 0.9943 ± 0.0019 and nucleotide diversity was 0.0428 ± 0.0210. The neutrality tests of the Daur group showed significant negative values and the mismatch distribution curve was obviously distributed in a unimodal pattern. The results showed that the Daur ethnic group has high genetic diversity and may have experienced recent population expansion. In addition, the main haplogroups of the Daur population were haplogroup D (31.51%), M* (20.55%), C (10.28%), F (7.53%), and B (6.85%), all of which were prevalent in northern China. It probably implies the northern Chinese origin of the Daur population. The PCA, F_{ST}, and phylogenetic analysis results indicated that the Daur group formed a cluster with East Asian populations, and had few genetic differences with the populations in northern China. More importantly, we found that disease-related mutation sites of the mitochondrial genome may be related to ethnic groups, which may have important implications for the prevention and occurrence of specific diseases. Overall, this study revealed the complexity and diversity of the matrilineal genetic background of the Daur ethnic group. Meanwhile, it provided meaningful data for the research on the diversity of the human genome.

Keywords: Daur ethnic group, mitochondrial DNA, genetic diversity, maternal inheritance, population genetics

1 INTRODUCTION

Mitochondrial DNA (mtDNA) is the only DNA that exists outside the nucleus of human cells. MtDNA has lower molecular weight and higher mutation rate than nuclear DNA (Chatterjee et al., 2006). Mitochondria have unique cell dynamics to ensure their correct distribution in dividing cells and high fidelity of genomic inheritance through maternal transmission (Mishra and Chan, 2014). Moreover, mtDNA also has a high copy number and lack of recombination properties. MtDNA reveals regional and ethnic genetic differences, and it is widely used in the fields of population genetics, forensic science, and evolutionary anthropology (Zheng et al., 2011; Chaitanya et al., 2016; Font-Porterias et al., 2018). The study of mtDNA genetic markers reflects the evolutionary history of a population, which is helpful to infer the maternal origin of the population and analyze the migration trajectory. Meanwhile, it also reflects the genetic relationship among different populations.

The Daur ethnic group is one of the minority nationalities in northern China, mainly distributed in the Daur Autonomous Banner of Morin Dawa, Inner Mongolia Autonomous Region, and Qiqihar, Heilongjiang Province. The Daur language belongs to the Altaic language family. The Daur nationality is sparsely populated, and there are few studies on its mtDNA polymorphism. The few previous studies available have never performed complete mitochondrial genome sequencing of the Daur population. Therefore, the research on this subject is extremely necessary and significant. In recent years, there have been increasing studies on the genetic polymorphism of East Asian populations, especially Chinese ethnic groups (Park et al., 2017; Trejaut et al., 2019; Wei et al., 2020). However, there are still some ethnic groups that have rarely been studied. It has led to incomplete mtDNA databases for some populations around the world, which has greatly restricted the study of human evolution and origin. The Daur population is a rare ethnic minority in China; therefore, our research samples are extremely precious, and it is necessary to study their genetic diversity. It may be of great significance to study the historical migration and evolution of the East Asian population.

Many researchers pay more attention to the analysis of mitochondrial hypervariable region sequences. However, the mutations in the coding region also make an important contribution to the construction of maternal lineages. Therefore, sequencing of the whole mitochondrial genome will significantly improve the resolution for distinguishing differences between individuals or groups (Seo et al., 2015). On the other hand, the genetic information in the mitochondria can be obtained more accurately and comprehensively. MtDNA has an extremely crucial significance in the related research of population genetics. In this study, we chose the whole mitochondrial sequencing to research the mtDNA diversity of the Daur ethnic group, which would clarify the distribution characteristics of polymorphism sites and haplogroups for the Daur ethnic group and would reveal its maternal genetic structure. The genetic discrepancy between the Daur ethnic group and other populations would also be studied. It would provide powerful genetic information for understanding the history of changes among groups.

2 MATERIALS AND METHODS

2.1 Sample Collection

A total of 146 samples were collected from unrelated healthy individuals of the Daur ethnic group in Qiqihar, Heilongjiang Province, including 71 males and 75 females. All of them have lived in Heilongjiang region for at least three generations according to the narrative. Written informed consent was obtained from all participants. This research was approved by the ethics committee of the Second Affiliated Hospital of Harbin Medical University (Approval Number: KY2020-250). All methods were performed in a manner consistent with the approved protocols and in accordance with the relevant guidelines and regulations for human subjects research.

2.2 DNA Extraction, Long-PCR Amplification, and Sequencing

DNA was extracted from sample blood using the QIAamp DNA Blood Mini Kit (QIAGEN) according to the manufacturer's protocol.

At first, six pairs of primers were used for long-PCR amplification to enrich the mitochondrial genome. The primers are described in detail elsewhere (Xu et al., 2021). A total volume of 20 μL in the PCR reaction system containing 2.4 μL 2.5 mm dNTP, 1 μL each of reverse and forward primers (1 μm), 1 μL template DNA (2 ng/μL), 10.2 μL ddH2O, 4 μL 5× TransStart FastPfu Fly Buffer, and 0.4 μL DNA polymerase. PCR was performed under the following cycle conditions: 95°C for 10 min; followed by 28 cycles of 94°C for 20 s, 68°C for 6 min; a final extension at 72°C for 12 min, and hold at 4°C. Amplification products were purified and fragmented. End-repair, end tail, and adapter ligation for fragmented DNA were performed using the NEBNext® DNA Library Prep Reagent Set for Illumina®. Library fragment selection and quality assessment were done using Agilent 2100 Bioanalyzer. Finally, the libraries were sequenced in a 2 bp × 150 bp paired-end mode on the Illumina Hiseq platform.

2.3 Sequencing Data Processing

The original data was obtained by high-throughput sequencing. By using the MEM algorithm of BWA software (http://bio-bwa.sourceforge.net/) (Li and Durbin, 2010) to compare the original data of each sample with the reference genome, acquired the preliminary mapped results in the BAM file. The human reference genome of this research was the revised Cambridge Reference Sequence (rCRS) of hg38 at UCSC. Picard software (https://broadinstitute.github.io/picard/) was used to analyze the mapped information of each sample, including the ratio of duplicate reads resulting from PCR amplification and the average sequencing depth, etc. GATK (https://software.broadinstitute.org/gatk/best-practices/) (McKenna et al., 2010) was used to calibrate the preliminary mapped results obtained by BWA software, which greatly reduced the false positives and false negatives generated during the sequencing and mapping process. Detected the mutation sites of the complete mitochondrial genome by the GATK Mutect2+HaplotypeCaller method.

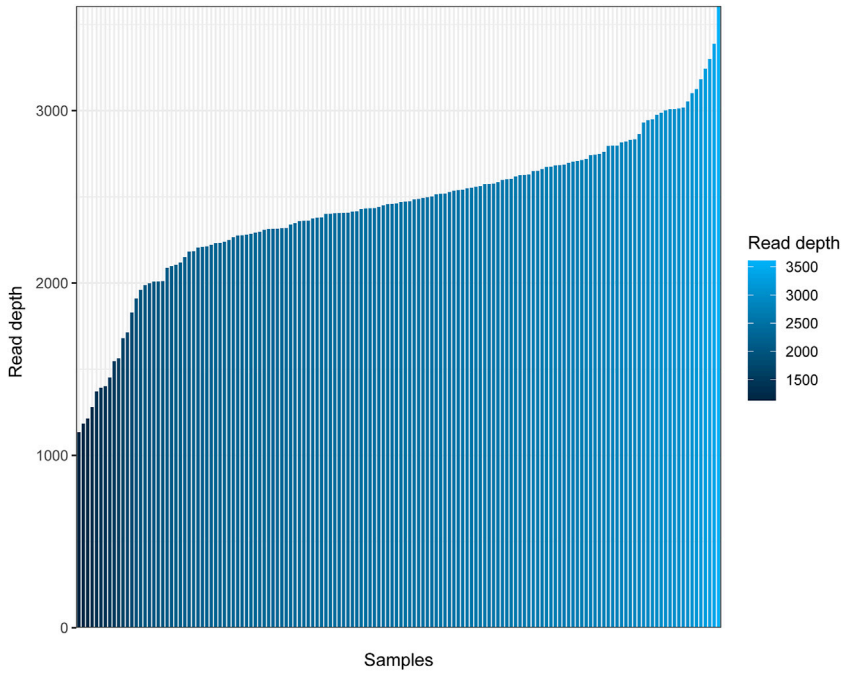


FIGURE 1 | Read depth for the complete mitochondrial genome of 146 Daur individuals. The horizontal axis represents the different individuals sorted from small to large according to the mean sequencing depth, and the vertical axis represents the average read depth.

Information annotation for all variant sites by ANNOVAR (http://annovar.openbioinformatics.org/en/latest/) (Wang et al., 2010). Meanwhile, deep filtering by Perl scripts was performed to obtain detailed mutation information of all samples. Finally, the mitochondrial sequences in FASTA format were generated.

2.4 Analysis of Mitochondrial Sequences

To further describe the complex matrilineal genetic landscape of the Daur ethnic group, this study sequenced and generated 146 mitochondrial sequences of the Daur group. In addition, we also searched a total of 5,094 complete mitochondrial sequences from 56 populations as reference data by two researchers. Among them, a total of 2,503 individuals from 26 populations were collected from the 1000 Genome Project. The whole mitochondrial sequences of the 30 populations were screened from published studies and then downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Ultimately, there are 5,240 mitochondrial genomes from 57 populations included in our research. Detailed information on the worldwide populations and cited references are listed in **Supplementary Table S1**.

All of the complete mtDNA sequences in the FASTA format were aligned with rCRS using BioEdit software (Anderson et al., 1981; Andrews et al., 1999). The genetic diversity indexes containing a number of polymorphic sites (S), the total number of mutations (Eta), and the number of haplotypes (h) were calculated by DnaSp v6 software. Nucleotide diversity (Pi), haplotype diversity (HD), the mean number of pairwise differences, neutrality tests including Tajima's D and Fu's Fs, and the values of mismatch distribution analysis were estimated using Arlequin ver 3.5.2.2. Analysis of molecular variance (AMOVA) and pairwise fixation index (F_{ST}) were also generated by Arlequin ver 3.5.2.2. The haplogroups of the complete mitochondrial genome sequences for the Daur group were classified using HaploGrep2 based on PhyloTree build 17 (http://www.phylotree.org/index.htm). In addition, the haplogroups of 56 reference populations worldwide involved in this study have also been redefined. The frequencies of the mitochondrial haplogroups were calculated by direct counting. To reveal the relationship between mitochondrial polymorphism sites and disease of the Daur population, we annotated disease information based on MITOMAP (https://www.mitomap.org/) for variant sites. The chart of sequencing quality, pheatmap, and principal component analysis (PCA) were generated by R 4.0.3. The R packages used for PCA were "tidyr," "dplyr," and "ggplot2". The phylogenetic tree was produced by MEGA. To reconstruct the demographic history for Daur samples, we performed a Bayesian skyline plot (BSP) using BEAST 1.8.4. The plot was visualized with Tracer v1.7.2.

3 RESULTS

3.1 Sequencing Quality Analysis

In the present study, 146 Daur individuals were sequenced successfully. To observe the sequencing quality clearly, we plotted a bar chart to show the depth of sequencing for all individuals. As shown in **Figure 1**, the sequencing depth of all individuals was higher than 1,100 ×, and approximately ranged from 1,134 × to 3,607 ×. The average read depth was 2,439 × ± 434× (mean ± SD) for each individual. Q20 and Q30 values of 146

TABLE 1 | Genetic diversity indexes of the Daur ethnic group.

Index	Value
Number of polymorphic sites (S)	490
Total number of mutations (Eta)	497
Nucleotide diversity (Pi)	0.0428 ± 0.0210
Number of haplotypes (h)	111
Haplotype diversity (HD)	0.9943 ± 0.0019
Mean number of pairwise differences	20.9604 ± 9.3042

sequencing samples are displayed in **Supplementary Table S2**. The sequencing performance was excellent for the whole mitochondrial genome in our research.

3.2 Genetic and Variation of Mitochondrial DNA

To research the genetic and variation characteristics of the Daur ethnic group, we calculated the related genetic diversity indexes (**Table 1**). A total of 497 variants were observed at 490 positions in the range of the complete mitochondrial genomes, including 77 transversions and 420 transitions. Meanwhile, among 146 Daur individuals analyzed in this study, 111 different haplotypes were detected and 91 of them were unique. It was worth noting that the most frequent haplotype occurred 6 times in all individuals, the following haplotype appeared 5 times and another appeared 4 times. 11 haplotypes occurred 2 times, and six haplotypes occurred 3 times. The haplotype diversity (HD = 0.9943 ± 0.0019) and the nucleotide diversity (Pi = 0.0428 ± 0.0210) were also significant genetic parameters we focused on. The mean number of pairwise differences in the studied Daur population was 20.9604 ± 9.3042, respectively.

The values of Tajima's D (−2.503) and Fu's Fs (−23.820) of the Daur group were calculated. Analysis of mismatch distribution was performed to reflect the historical dynamics of the Daur group, as shown in **Figure 2**. The mismatch distribution curve of the Daur was obviously distributed in a unimodal pattern. Moreover, we detected the observation model was basically consistent with the expected expansion model. The results implied that the group has experienced expansion or continued growth in the past. The reliability of the result of the mismatch distribution analysis was evaluated through two parameters: sum of squared deviations (SSD) and Harpending's Raggedness index (HRI). The SSD (0.0003, $p = 0.940$) and HRI (0.0007, $p = 0.990$) of the Daur group showed the statistical test was not significant; it suggested that the hypothesis of group expansion cannot be rejected.

3.3 The Population Expansion Time of Daur Ethnic Group

In order to further verify whether population expansion occurred in the Daur group, we performed the Bayesian skyline plot (BSP) to clarify the population expansion time. **Figure 3** shows the effective population size of the Daur group over the past 80 kya. The results showed that the Daur group experienced a significant population expansion around 70 kya, resulting in a sharp increase in population size. The effective population size was relatively stable from 67 kya to 26.6 kya, showing a slow upward trend. After this period of time, the Daur ethnic group experienced a small population expansion at 26.6 kya. After stabilizing for a while, there was a small shrinkage in population size at 7.5 kya.

3.4 Mitochondrial Haplogroup Distribution

The distribution of each haplogroup reflects the basic composition of the genetic structure of a population. A total of 88 sub-haplogroups were classified among 146 complete mitochondrial genomes of the Daur group based on PhyloTree build 17. The detailed haplogroup classification results of each individual are shown in **Supplementary Table S3**. Haplogroup D (31.51%) was the most common haplogroup and D4 (26.71%)

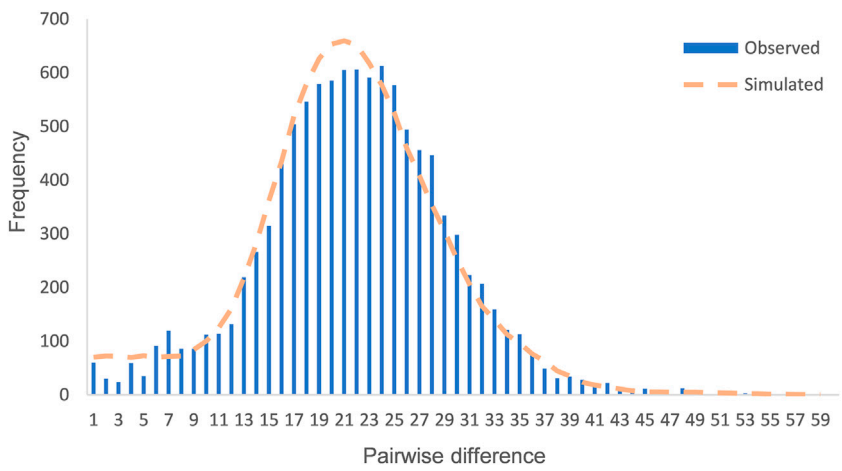


FIGURE 2 | Mismatch distribution of the Daur ethnic group. The orange dotted line represents the simulated model, and the blue bar graph represents the observed model of the Daur ethnic group.

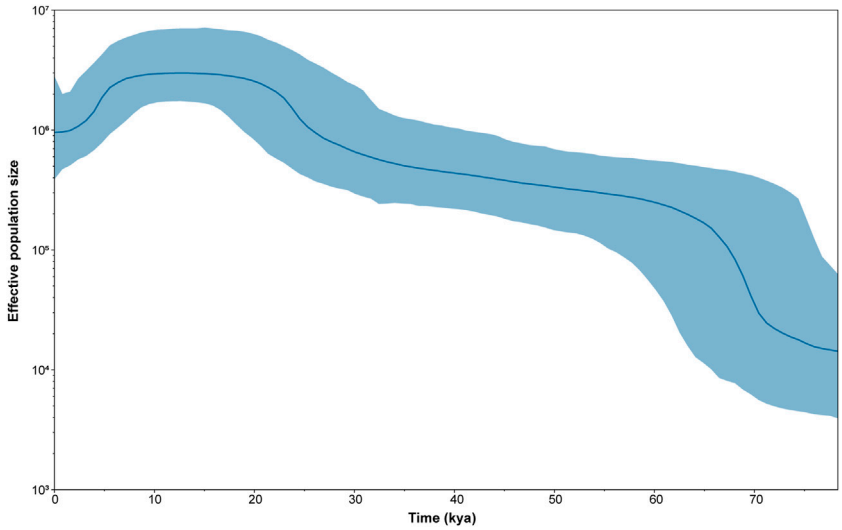


FIGURE 3 | The Bayesian skyline plot (BSP) of changes in effective population size through time for the Daur ethnic group. The dark blue line represents the median population, and the blue line demarcates the boundaries of the 95% highest posterior density.

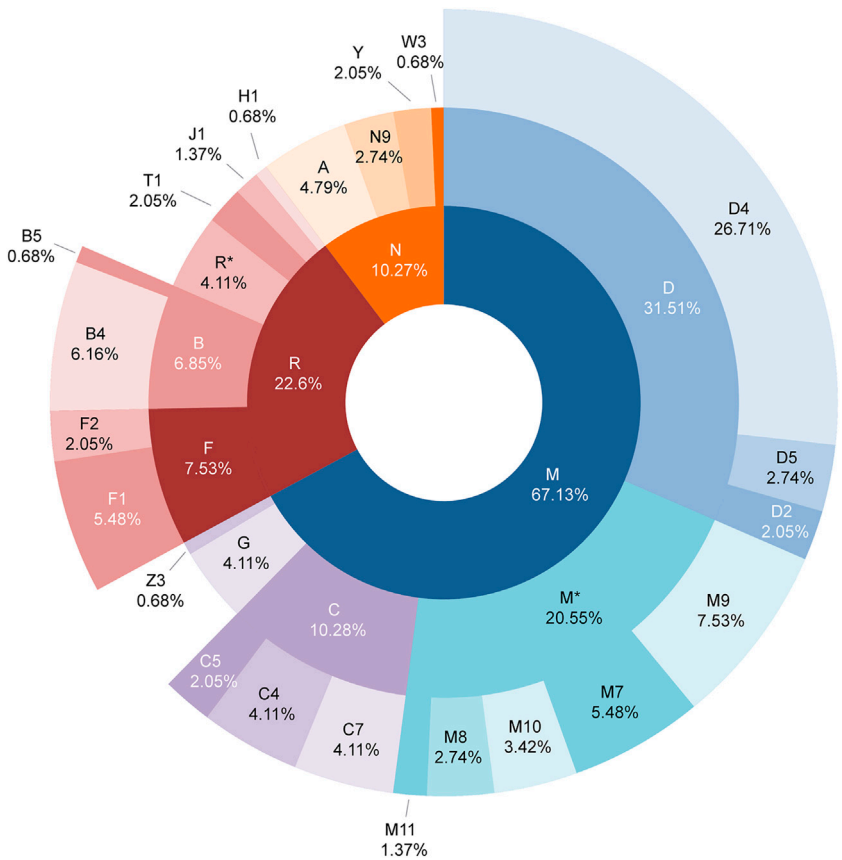


FIGURE 4 | Distribution of mitochondrial haplogroups in the Daur ethnic group. The macro haplogroups M, N, and R displayed in the innermost circle are represented by different colors. The circle in the middle represents the distribution and proportion of each haplogroup belonging to macro haplogroups M, N, and R. The outermost circle shows the distribution and proportion of sub-haplogroups in more detail.

TABLE 2 | 17 disease-related sites in the mitochondrial genome of the Daur ethnic group.

Gene	Position	Disease	Count	Frequency
ND3	A10398G	PD protective factor/longevity/altered cell pH/metabolic syndrome/breast cancer risk/ADHD	109	0.7466
CYTB	G15043A	MDD-associated	98	0.6712
CR	T310TC	Melanoma	55	0.3767
ND2	C4883T	Glaucoma	46	0.3151
ND2	C5178A	Longevity; extraversion MI/AMS protection; blood iron metabolism	46	0.3151
RNR2	G3010A	Cyclic vomiting syndrome with migraine	45	0.3082
ATP8	C8414T	Longevity	42	0.2877
ND6	C14668T	Depressive disorder associated	41	0.2808
CR	T16189C	Diabetes/cardiomyopathy/cancer risk/mtDNA copy nbr/metabolic syndrome/melanoma	29	0.1986
CR	G16129A	Cyclic vomiting syndrome with migraine	25	0.1712
CR	A16183C	Melanoma	21	0.1438
CR	C150T	Longevity/cervical carcinoma/HPV infection risk	20	0.1370
tRNA-Pro	T16093C	Cyclic vomiting syndrome	17	0.1164
CR	T195C	BD-associated/melanoma	14	0.0959
ND1	T3394C	LHON/diabetes/CPT deficiency/high-altitude adaptation	11	0.0753
ND4	G11696A	LHON/LDYT/DEAF/hypertension helper mut	9	0.0616
COX1	G6962A	Possible helper variant for 15927A	9	0.0616

accounted for the largest proportion among them, then followed by haplogroup M* (20.55%), haplogroup C (10.28%), haplogroup F (7.53%), and haplogroup B (6.85%), which accounted for 76.71% of the haplogroups in the Daur population, while haplogroups A (4.79%), G (4.11%), R* (4.11%), N9 (2.74%), Y (2.05%), and Z3 (0.68%) accounted for a relatively small proportion. Interestingly, European-specific haplogroups such as T1, J1, H1, and W3 also contributed 4.78% to the maternal genetic structure of the Daur population. We have generated a sunburst chart (**Figure 4**) to show the distribution of Daur haplogroups more intuitively. It can be observed all of the haplogroups belonged to macro haplogroups M, N, and R, macrohaplogroup M occupied the largest proportion among them. Additionally, we reconstructed the haplogroup tree for the tested Daur group, reflecting the evolution of specific branches during defining sub-haplogroups in detail (**Supplementary Figure S1**).

3.5 Specific Disease-Related Mutation Sites in the Daur Mitochondrial Genome

In order to research the distribution of known disease sites in the entire mitochondrial genome of the Daur population, and to grasp the specific maternal genetic markers of the Daur population more accurately, we screened out disease-related sites with a higher mutation frequency in the Daur population. According to disease information annotation based on MITOMAP, we found 71 reported mitochondrial genome loci related to disease in the Daur population, of which 17 loci with a minimum allele frequency higher than 0.05, accounting for 23.94% of the total detected known disease sites. The information of the 17 loci is listed in **Table 2**. Most of the mutations at these sites occurred in the coding region, and mutation at six sites occurred in the control region. The most common disease-related locus in the Daur ethnic group was A10398G, with a frequency of 0.7466. On the contrary, G11696A and G6962A were the disease-related locus with the lowest mutation frequency, accounting for only 0.0616.

3.6 Genetic Discrepancy for Daur and Other Populations

3.6.1 Analysis of Molecular Variance

To determine the factors that may play a role in the mtDNA diversity, we performed an analysis of molecular variance (AMOVA) by grouping the 57 studied worldwide populations according to the different classifications (including language dialects and geographic regions) as shown in **Table 3**. We have observed that whether it is classification of language families or geographic regions, within the populations, variation occupied an extremely major proportion. Variations among groups accounted for the minimum contribution. In the groups by geographic distributions of worldwide populations, variation within populations was 87.35%, among populations within groups was 6.93%, and among groups was 5.73%. For the linguistic family groups of worldwide populations, variation within populations was 87.65%, among populations within groups was 9.65%, and among groups was 2.70%. Moreover, the values of geographic regions were lower than those of the language dialects groups for the variation within populations and among populations within groups. Populations separated by geographic regions contained a higher percentage of variation compared to the groups of language families after grouping.

3.6.2 Principal Component Analysis Based on Haplogroup Frequency

To investigate the genetic discrepancy between the Daur ethnic group and 56 other worldwide populations. Principal component analysis (PCA) based on haplogroup frequency was performed (**Figure 5**, **Supplementary Figure S2**, **Supplementary Figure S3**). Due to the higher variation for geographical grouping based on the AMOVA results, we divided the tested populations into seven groups according to the geographical regions. The first

TABLE 3 | The AMOVA results based on 57 worldwide populations.

Grouping	Number of populations	Number of groups	Among groups	Among populations within groups	Within populations
Geographic distributions of worldwide populations	57	7	5.73	6.93	87.35
Linguistic families of worldwide populations	57	11	2.70	9.65	87.65

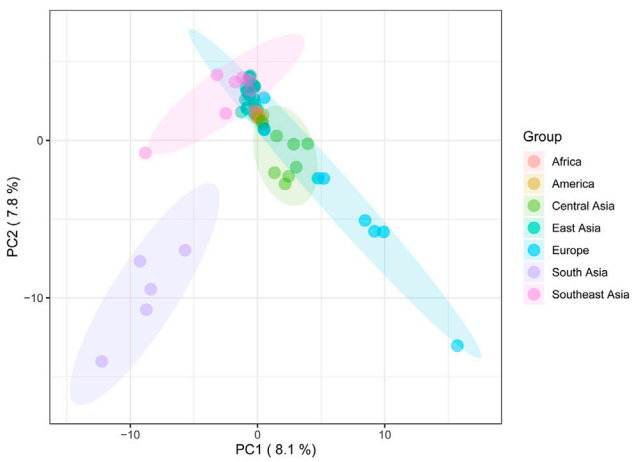


FIGURE 5 | The principal component analysis (PCA) plot for the Daur group and 56 worldwide populations. Daur belongs to the East Asian cluster, indicated in dark green.

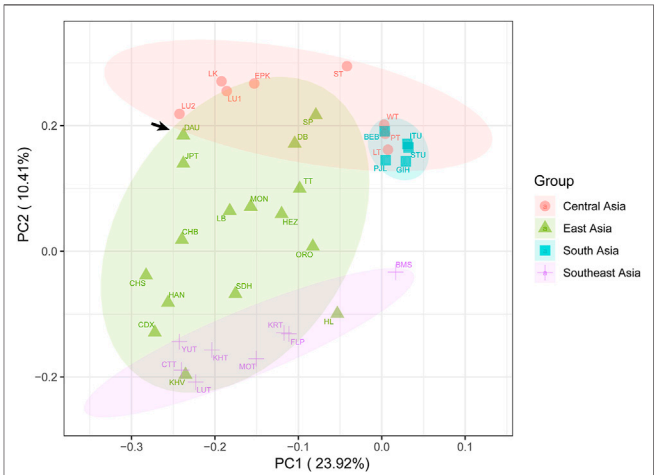


FIGURE 6 | The principal component analysis (PCA) plot for the Daur group and other Asian populations. The Daur group (DAU) is indicated with arrows.

three principal components explained 23.7% of the variation, of which PC1, PC2, and PC3 accounted for 8.1%, 7.8%, and 7.8%, respectively. The PCA results showed East Asian populations clustered very tightly in the context of analysis of worldwide populations.

To observe the genetic relationship more clearly between the Daur group and other East Asian populations, we further performed PCA in the context of the entire Asian mtDNA (**Figure 6**, **Supplementary Figure S4**, **Supplementary Figure S5**). A total of 43.73% of genetic variations were extracted by the first three components (PC1: 23.92%, PC2: 10.41%, and PC3: 9.40%). In the PC1 and PC2, the point representing the Daur population was relatively closer to JPT and LU2 (Lowland Uyghur). In the PC1 and PC3, the Daur population was also relatively closer to JPT and LU2. The plot of PC2 and PC3 illustrated the Daur population was closer to some populations living in the Tibetan Autonomous Region, such as DB (Deng), TT (Tingri Tibetan), and SP (Sherpa).

3.6.3 Pairwise Fixation Index (F_{ST}) Values Reveal Population Genetic Distance

To reveal the discrepancy in the matrilineal genetic landscape between the Daur ethnic group and other populations, the pairwise F_{ST} values for the complete mitochondrial sequences between the Daur group and 56 reference populations were calculated as shown in **Supplementary Table S4**. Our results showed that the F_{ST} values between the Daur group and the reference population ranged from 0.00402 to 0.46263. The lowest pairwise F_{ST} value was between the Daur group and TT (F_{ST} = 0.00402), followed by JPT (F_{ST} = 0.00464). The Daur group also showed lower F_{ST} values with SP (F_{ST} = 0.01728) and MON (Mongola, F_{ST} = 0.01914). In the comparison with the Daur group, the largest value was compared with IBS (Iberian Population in Spain, F_{ST} = 0.46263), followed by CEU (Utah Residents with Northern and Western European Ancestry, F_{ST} = 0.43967). Meanwhile, all the pairwise F_{ST} values were visualized by a heatmap to show the genetic distance more clearly, as shown in **Figure 7**. Populations were grouped according to the geographic distribution.

3.6.4 Phylogenetic Analysis

We conducted a phylogenetic analysis to further clarify the genetic relationship between the Daur ethnic group and other populations and generated a phylogenetic tree based on the F_{ST} values (**Figure 8**). We still divided the 57 populations into seven groups according to their geographical distribution: Africa cluster, America cluster, Central Asia cluster, East Asia cluster, Europe cluster, South Asia cluster, and Southeast Asia cluster. We observed that the Daur population and JPT, MON, and TT gathered on the same subbranch, especially clustering closer to JPT.

4 DISCUSSION

Mitochondria, as a useful genetic marker, reflect the characteristics of maternal inheritance and are suitable for

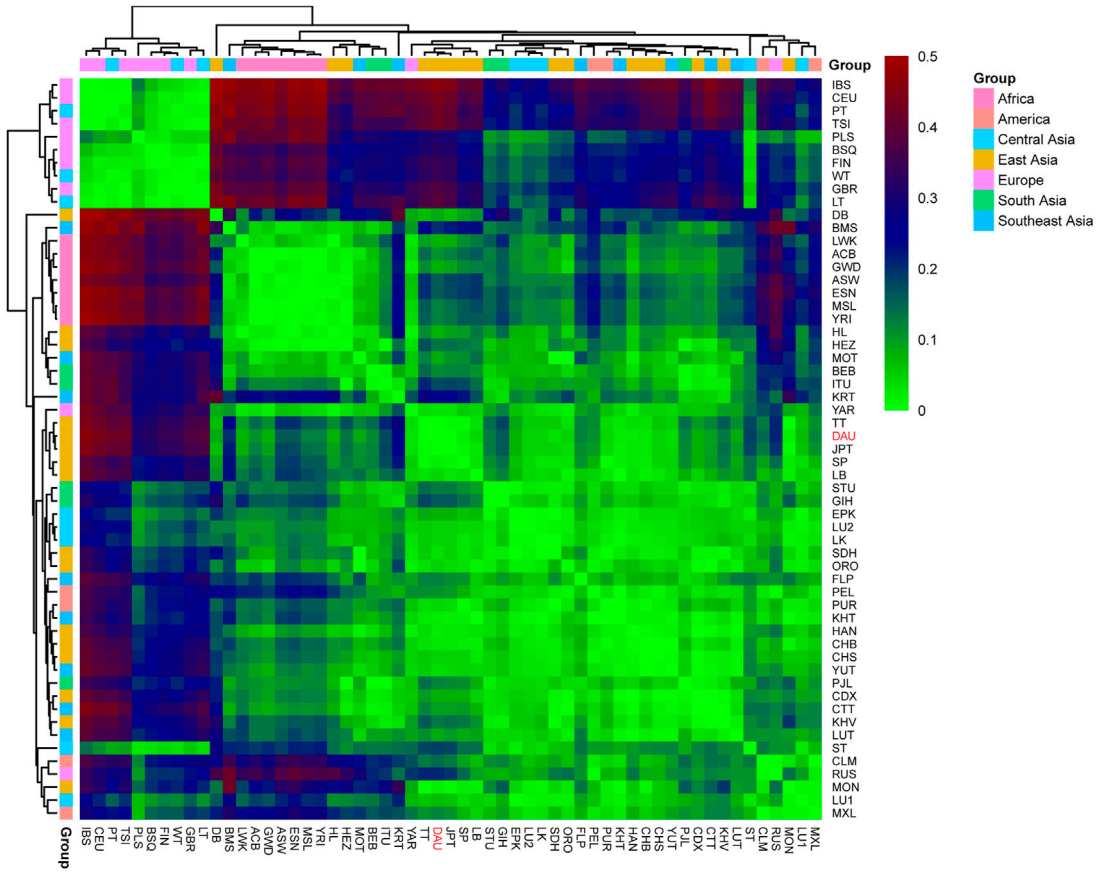


FIGURE 7 | Heatmap for genetic distance among the Daur ethnic group (DAU) and 56 worldwide populations. Visualizing the F_{ST} values with different colors. The color red represents the high F_{ST} values, and the green represents the low F_{ST} values. DAU is highlighted in red.

population evolution analysis. The main purpose of the present study was to understand the maternal genetic diversity of the Daur ethnic group and to provide important genetic information for the study of human genome diversity.

We analyzed the genetic variation in the complete mitochondrial genomes of the Daur ethnic group from 146 Daur individuals. The haplotype diversity (Nei and Tajima, 1981) is a measure of the uniqueness of a particular haplotype in a certain population, rendering a high gene diversity value (HD = 0.9943 ± 0.0019) in the Daur group. Meanwhile, the value of nucleotide diversity (Nei and Li, 1979) (Pi = 0.0428 ± 0.0210) also revealed that the Daur ethnic group had high genetic diversity and rich genetic resources. The neutrality tests can be used to detect natural selection among the nucleotide sequence variants in a population. In this study, the significant negative values of the neutrality tests (Tajima's D and Fu's Fs) mainly reflected that the Daur population experienced the population expansion after a bottleneck recently or indicated an excess of rare variation (Tajima, 1989; Fu and Li, 1993; Carlson et al., 2005). At the same time, the results of the neutrality tests deviated from the neutral mutation significantly. According to the results of the mismatch distribution analysis, on the other hand, it was speculated that the Daur group underwent recent population expansion potentially (Mousset et al., 2004). BSP also proved that Daur had experienced a significant population expansion at 70 kya and a small population expansion at 26.6 kya. Therefore, all of the results reached a consistent conclusion, which further supported the speculation of population expansion powerfully.

Mitochondrial dysfunction and defects caused by mitochondrial DNA polymorphism are related to many diseases. It has been reported that the A10398G variant was probably related to metabolic syndrome (Yan et al., 2014), attention deficit and hyperactivity disorder (ADHD) (Hwang et al., 2017), breast cancer susceptibility (Tengku Baharudin et al., 2012), and Parkinson's disease (PD) (Jiang et al., 2004). Meanwhile, A10398G may lead to the reduction of the function of the complex I, and the level of reactive oxygen species (ROS) in the cell increased subsequently (Mohamed Yusoff et al., 2018). It will further accumulate more damage to mtDNA to promote the occurrence and development of the disease. It is reported that the generation of ROS may be related to type 2 diabetes mellitus (T2DM) risk (Chalkia et al., 2018). Furthermore, studies have shown that the G11696A mutation may be related to Leber's hereditary optic neuropathy (LHON) (De Vries et al., 1996; Dai et al., 2018). The mutation frequency of Daur in G11696A was only 0.0616,

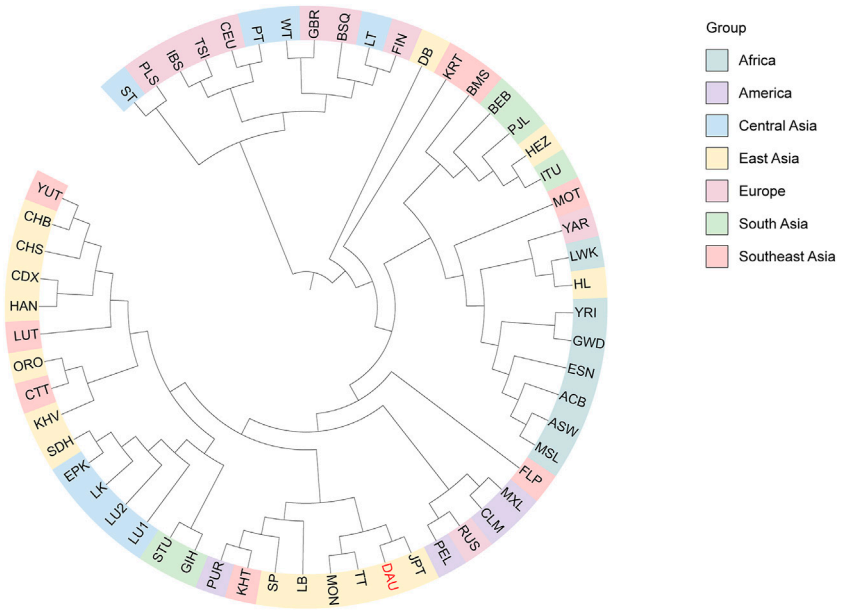


FIGURE 8 | The phylogenetic UPGMA tree for the Duar ethnic group (DAU) and 56 reference populations. The tree is colored according to geographic origins. DAU is highlighted in red.

it is speculated that the Daur population was less susceptible to LHON caused by the mutation of the locus.

The results showed that the Daur ethnic group in this study produced 88 specific sub-haplogroups. The frequency of haplogroups varies with varying degrees among populations in different regions. Previous studies have shown that haplogroup D maintained a very high overall frequency among East Asian, North Asian, and Central Asian populations (Derenko et al., 2007; Derenko et al., 2010). Haplogroup D4 clades more likely reside in the north of East Asia (Zheng et al., 2011). It is also prevalent in northern and northeastern China, implying a potential northern China origin of this haplogroup (Li et al., 2019). Haplogroup M was initially thought to be an ancient marker of East Asian origin. The geographic distribution of M9 is in Central and East Asia, with the highest frequency in Tibet (Chandrasekar et al., 2009). There are significant differences in the frequency of haplogroup M9a between high-altitude Tibetan populations and low-altitude populations, and it has the highest frequency in the Tibetan population (Li et al., 2016). Haplogroup C7 mainly existed in East Asia (Derenko et al., 2010). Haplogroup B4 was a typical haplogroup in southern China (Li et al., 2019). It was worth noting that the most common European mitochondrial haplogroups T, J, H, and W have also been detected in the Daur ethnic group (Grignani et al., 2009; Kozin et al., 2020). These results suggested that most of the haplogroups of the Daur ethnic group were popular in East Asia, and our results proved that the Daur population belongs to East Asian lineage and originated from northern China. In addition, due to the emergence of European-specific haplogroups in the results, we speculated that European ancestry also contributed a small proportion to the maternal inheritance pool for the Daur ethnic group.

AMOVA was used to detect significant variation in the genetic structure of mtDNA among populations (Bodner et al., 2011). We found that when the dominant variation occurred within populations, it revealed more genetic discrepancy within populations. The variation among groups based on geographic regions was slightly higher than that based on linguistic families. The results indicated that geographical grouping might provide a better explanation for the genetic divergence of complete mitochondrial genomes among groups than linguistic grouping.

As for the PCA results, the close clustering between the Daur group and East Asian populations meant that there were almost no genetic differences between them. It confirmed the previous conclusion that the Daur ethnic group belongs to the East Asian branch. In the PCA for all Asian populations, we found the Chinese Heilongjiang Daur group had a close genetic relationship with JPT, TT, SP, DB, and LU2. The population genetic differences between East Asia and Southeast Asia seem to be less obvious to be detected. On the contrary, Daur and South Asian groups performed the farthest genetic relationship among all Asian populations. Meanwhile, the genetic structure of the Daur ethnic group was also well expressed by PCAs.

F_{ST} provides important insights into the evolutionary processes that influence the structure of genetic variation within and among populations, and it is among the most widely used descriptive statistics in population and evolutionary genetics. The small F_{ST} value means that the allele frequencies in each population are similar. If the value is larger, that means the allele frequencies are different, indicating that the genetic distance is farther (Holsinger and Weir, 2009). According to our results, the Daur

population showed a close genetic distance with TT, JPT, SP, and MON. Among them, Daur had the closest genetic distance with TT in the Tibetan region. In addition, the Daur ethnic group with IBS showed the farthest genetic distance among all studied populations. Overall, the Daur ethnic group showed a closer genetic relationship with the vast majority of East Asians, especially the north Chinese populations. However, the Daur group showed obvious genetic divergence with European populations.

At the same time, the phylogenetic tree generated based on the F_{ST} values also revealed consistent results. The Daur group apparently congregated with East Asian populations and distributed in the nearest sub-branch with JPT. It revealed that there was little genetic difference between the Daur group and JPT; they had a close maternal genetic relationship. Our results may be due to the Daur sampling site in this study being located in northeast China, which is geographically close to Japan, and the recent introduction of the Daur genes into Japanese or their common maternal origin. In addition, the results revealed that the genetic distance between the Daur group and TT was relatively close. It may be due to the gene exchange and fusion between the Daur population and TT during historical development. The Daur ethnic group and MON also showed a close genetic relationship; they both belong to the Altaic language family. It is reported that the Daur ethnic group originated from the Mongolian ethnic group. The results of this study may further explain the view from the perspective of maternal genetics.

Our research provided a complex and comprehensive maternal genetic landscape of the Daur ethnic group. First, we found that the Daur ethnic group has a high genetic diversity and may have experienced recent population expansion. According to the results, most of the haplogroups of Daur are prevalent in East Asia. It is confirmed that the Daur group belongs to the East Asian lineage and originated from north China. All results of PCA, F_{ST}, and phylogenetic tree revealed that the Daur group was closely clustered with East Asian populations, especially in northern China. The Daur ethnic group showed a closer genetic relationship with TT, MON, JPT, and SP. We found that the specific disease-related mutation sites of the mitochondrial genome may be ethnic-related. Overall, the mitochondrial genome generated in this study would enrich the existing mtDNA database, actively promoting the research on the genetic diversity and population historical dynamics of the Daur ethnic group.

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; BankIt2567580: ON127701—ON127846.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of the Second Affiliated Hospital of Harbin Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MJ wrote the manuscript; QL and MJ conceived and designed the study; MJ and QL ran analyses; MJ, QL, TZ, BD, and XL conducted experiments; JY, SF, and QL revised the manuscript; all authors contributed to critically revising the manuscript. All authors read and approved the final version.

FUNDING

This work was supported by the research fund of a key laboratory for the preservation of human genetic resources and disease control in China (Harbin Medical University), Ministry of Education, China.

ACKNOWLEDGMENTS

The authors thank all volunteers who provided blood samples for this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.919063/full#supplementary-material

REFERENCES

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., et al. (1981). Sequence and Organization of the Human Mitochondrial Genome. *Nature* 290 (5806), 457–465. doi:10.1038/290457a0

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N. (1999). Reanalysis and Revision of the Cambridge Reference Sequence for Human Mitochondrial DNA. *Nat. Genet.* 23 (2), 147. doi:10.1038/13779

Bodner, M., Zimmermann, B., Röck, A., Kloss-Brandstätter, A., Horst, D., Horst, B., et al. (2011). Southeast Asian Diversity: First Insights into the Complex mtDNA Structure of Laos. *BMC Evol. Biol.* 11, 49. doi:10.1186/1471-2148-11-49

Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., et al. (2005). Genomic Regions Exhibiting Positive Selection Identified from Dense Genotype Data. *Genome Res.* 15 (11), 1553–1565. doi:10.1101/gr.4326505

Chaitanya, L., van Oven, M., Brauer, S., Zimmermann, B., Huber, G., Xavier, C., et al. (2016). High-quality mtDNA Control Region Sequences from 680 Individuals Sampled across the Netherlands to Establish a National Forensic mtDNA Reference Database. *Forensic Sci. Int. Genet.* 21, 158–167. doi:10.1016/j.fsigen.2015.12.002

Chalkia, D., Chang, Y.-C., Derbeneva, O., Lvova, M., Wang, P., Mishmar, D., et al. (2018). Mitochondrial DNA Associations with East Asian Metabolic Syndrome. *Biochim. Biophys. Acta (BBA) - Bioenerg.* 1859 (9), 878–892. doi:10.1016/j.bbabio.2018.07.002

Chandrasekar, A., Kumar, S., Sreenath, J., Sarkar, B. N., Urade, B. P., Mallick, S., et al. (2009). Updating Phylogeny of Mitochondrial DNA Macrohaplogroup M in India: Dispersal of Modern Human in South Asian Corridor. *PloS one* 4 (10), e7447. doi:10.1371/journal.pone.0007447

Chatterjee, A., Mambo, E., and Sidransky, D. (2006). Mitochondrial DNA Mutations in Human Cancer. *Oncogene* 25 (34), 4663–4674. doi:10.1038/sj.onc.1209604

Dai, Y., Wang, C., Nie, Z., Han, J., Chen, T., Zhao, X., et al. (2018). Mutation Analysis of Leber's Hereditary Optic Neuropathy Using a Multi-Gene Panel. *Biomed. Rep.* 8 (1), 51–58. doi:10.3892/br.2017.1014

De Vries, D. D., Went, L. N., Bruyn, G. W., Scholte, H. R., Hofstra, R. M., Bolhuis, P. A., et al. (1996). Genetic and Biochemical Impairment of Mitochondrial Complex I Activity in a Family with Leber Hereditary Optic Neuropathy and Hereditary Spastic Dystonia. *Am. J. Hum. Genet.* 58 (4), 703–711.

Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Dambueva, I., Perkova, M., et al. (2007). Phylogeographic Analysis of Mitochondrial DNA in Northern Asian Populations. *Am. J. Hum. Genet.* 81 (5), 1025–1041. doi:10.1086/522933

Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Rogalla, U., Perkova, M., et al. (2010). Origin and Post-glacial Dispersal of Mitochondrial DNA Haplogroups C and D in Northern Asia. *PLoS One* 5 (12), e15214. doi:10.1371/journal.pone.0015214

Font-Porterias, N., Solé-Morata, N., Serra-Vidal, G., Bekada, A., Fadhlaoui-Zid, K., Zalloua, P., et al. (2018). The Genetic Landscape of Mediterranean North African Populations through Complete mtDNA Sequences. *Ann. Hum. Biol.* 45 (1), 98–104. doi:10.1080/03014460.2017.1413133

Fu, Y. X., and Li, W. H. (1993). Statistical Tests of Neutrality of Mutations. *Genetics* 133 (3), 693–709. doi:10.1093/genetics/133.3.693

Grignani, P., Turchi, C., Achilli, A., Peloso, G., Alù, M., Ricci, U., et al. (2009). Multiplex mtDNA Coding Region SNP Assays for Molecular Dissection of Haplogroups U/K and J/T. *Forensic Sci. Int. Genet.* 4 (1), 21–25. doi:10.1016/j.fsigen.2009.04.001

Holsinger, K. E., and Weir, B. S. (2009). Genetics in Geographically Structured Populations: Defining, Estimating and Interpreting F(ST). *Nat. Rev. Genet.* 10 (9), 639–650. doi:10.1038/nrg2611

Hwang, I. W., Hong, J. H., Kwon, B. N., Kim, H. J., Lee, N. R., Lim, M. H., et al. (2017). Association of Mitochondrial DNA 10398 A/G Polymorphism with Attention Deficit and Hyperactivity Disorder in Korean Children. *Gene* 630, 8–12. doi:10.1016/j.gene.2017.08.004

Jiang, Y., Ellis, T., and Greenlee, A. R. (2004). Genotyping Parkinson Disease-Associated Mitochondrial Polymorphisms. *Clin. Med. Res.* 2 (2), 99–106. doi:10.3121/cmr.2.2.99

Kozin, M. S., Kulakova, O. G., Kiselev, I. S., Boyko, A. N., and Favorova, O. O. (2020). Variability of the Mitochondrial Genome and Development of the Primary Progressing Form of Multiple Sclerosis. *Mol. Biol. Mosk.* 54 (4), 596–602. doi:10.31857/s0026898420040084

Li, H., and Durbin, R. (2010). Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 26 (5), 589–595. doi:10.1093/bioinformatics/btp698

Li, Q., Lin, K., Sun, H., Liu, S., Huang, K., Huang, X., et al. (2016). Mitochondrial Haplogroup M9a1a1c1b Is Associated with Hypoxic Adaptation in the Tibetans. *J. Hum. Genet.* 61 (12), 1021–1026. doi:10.1038/jhg.2016.95

Li, Y.-C., Ye, W.-J., Jiang, C.-G., Zeng, Z., Tian, J.-Y., Yang, L.-Q., et al. (2019). River Valleys Shaped the Maternal Genetic Landscape of Han Chinese. *Mol. Biol. Evol.* 36 (8), 1643–1652. doi:10.1093/molbev/msz072

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 (9), 1297–1303. doi:10.1101/gr.107524.110

Mishra, P., and Chan, D. C. (2014). Mitochondrial Dynamics and Inheritance during Cell Division, Development and Disease. *Nat. Rev. Mol. Cell Biol.* 15 (10), 634–646. doi:10.1038/nrm3877

Mohamed Yusoff, A. A., Zulfakhar, F. N., Mohd Khair, S. Z. N., Abdullah, W. S. W., Abdullah, J. M., and Idris, Z. (2018). Mitochondrial 10398A>G NADH-Dehydrogenase Subunit 3 of Complex I Is Frequently Altered in Intra-Axial Brain Tumors in Malaysia. *Brain Tumor Res. Treat.* 6 (1), 31–38. doi:10.14791/btrt.2018.6.e5

Mousset, S., Derome, N., and Veuille, M. (2004). A Test of Neutrality and Constant Population Size Based on the Mismatch Distribution. *Mol. Biol. Evol.* 21 (4), 724–731. doi:10.1093/molbev/msh066

Nei, M., and Li, W. H. (1979). Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases. *Proc. Natl. Acad. Sci. U.S.A.* 76 (10), 5269–5273. doi:10.1073/pnas.76.10.5269

Nei, M., and Tajima, F. (1981). DNA Polymorphism Detectable by Restriction Endonucleases. *Genetics* 97 (1), 145–163. doi:10.1093/genetics/97.1.145

Park, S., Cho, S., Seo, H. J., Lee, J. H., Kim, M.-Y., and Lee, S. D. (2017). Entire Mitochondrial DNA Sequencing on Massively Parallel Sequencing for the Korean Population. *J. Korean Med. Sci.* 32 (4), 587–592. doi:10.3346/jkms.2017.32.4.587

Seo, S. B., Zeng, X., King, J. L., Larue, B. L., Assidi, M., Al-Qahtani, M. H., et al. (2015). Underlying Data for Sequencing the Mitochondrial Genome with the Massively Parallel Sequencing Platform Ion Torrent™ PGM. *BMC Genomics* 16 (Suppl. 1), S4. doi:10.1186/1471-2164-16-S1-S4

Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* 123 (3), 585–595. doi:10.1093/genetics/123.3.585

Tengku Baharudin, N., Jaafar, H., and Zainuddin, Z. (2012). Association of Mitochondrial DNA 10398 Polymorphism in Invasive Breast Cancer in Malay Population of Peninsular malaysia. *Malays J. Med. Sci.* 19 (1), 36–42.

Trejaut, J. A., Muyard, F., Lai, Y.-H., Chen, L.-R., Chen, Z.-S., Loo, J.-H., et al. (2019). Genetic Diversity of the Thao People of Taiwan Using Y-Chromosome, Mitochondrial DNA and HLA Gene Systems. *BMC Evol. Biol.* 19 (1), 64. doi:10.1186/s12862-019-1389-0

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Res.* 38 (16), e164. doi:10.1093/nar/gkq603

Wei, Y.-Y., Ren, Z.-P., Jin, X.-Y., Cui, W., Chen, C., Guo, Y.-X., et al. (2020). Haplogroup Structure and Genetic Variation Analyses of 60 Mitochondrial DNA Markers in Southern Shaanxi Han Population. *Biochem. Genet.* 58 (2), 279–293. doi:10.1007/s10528-019-09942-0

Xu, L., Yang, K., Fan, Q., Zhao, D., Pang, C., and Ren, S. (2021). Whole Mitochondrial Genome Analysis in Chinese Patients with Keratoconus. *Mol. Vis.* 27, 270–282.

Yan, R., Luan, Q. X., Liu, L. S., Wang, X. Y., Li, P., and Sha, Y. Q. (2014). Association between Chronic Periodontitis and Metabolic Syndrome Related Mitochondria Single Nucleotide Polymorphism. *Beijing Da Xue Xue Bao Yi Xue Ban.* 46 (2), 264–268.

Zheng, H.-X., Yan, S., Qin, Z.-D., Wang, Y., Tan, J.-Z., Li, H., et al. (2011). Major Population Expansion of East Asians Began before Neolithic Time: Evidence of mtDNA Genomes. *PloS one* 6 (10), e25835. doi:10.1371/journal.pone.0025835

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

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

Copyright © 2022 Jia, Li, Zhang, Dong, Liang, Fu and Yu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 07 July 2022
doi: 10.3389/fgene.2022.954586

Genetic Structure and Forensic Utility of 23 Autosomal STRs of the Ethnic Lao Groups From Laos and Thailand

*Khaing Zin Than[1], Kanha Muisuk[2], Wipada Woravatin[3], Chatmongkon Suwannapoom[4], Metawee Srikummool[5], Suparat Srithawong[3], Sengvilay Lorphengsy[6] and Wibhu Kutanan[3]**

[1]Biological Science Program, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, [2]Department of Forensic Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand, [3]Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, [4]School of Agriculture and Natural Resources, University of Phayao, Muang Phayao, Thailand, [5]Department of Biochemistry, Faculty of Medical Science, Naresuan University, Phitsanulok, Thailand, [6]The Biotechnology and Ecology Institute Ministry of Science and Technology, Vientiane, Laos

The Lao Isan and Laotian are the major groups in the area of present-day northeastern Thailand and Laos, respectively. Several previous genetic and forensic studies indicated an admixed genetic structure of Lao Isan with the local Austroasiatic speaking groups, e.g. Khmer, whereas there is a paucity of reporting Laotian's forensic short tandem repeats (STRs). Here, we newly generated 451 genotypes of seven Lao Isan and three Laotian populations (two Lao Lum and one Lao Thoeng) using 23 autosomal STRs embedded in Verifiler™ plus PCR Amplification kit. We reported allelic frequency and forensic parameters in different dataset: combined ethnic Lao groups, combined Lao Isan populations and combined Laotians. Overall, the forensic parameter results indicate that this set of STRs is suitable for forensic investigation. The anthropological results revealed the genetic homogeneity of Tai-Kadai speaking Lao groups from Thailand and Laos, consistent with previous studies, while the Austroasiatic speaking groups from southern Laos showed genetic relatedness to both Lao Isan and Khmer. In sum, STRs allelic frequency results can provide the genetic backgrounds of populations which is useful for anthropological research and also strengthens the regional forensic database in both countries.

Keywords: laotian, Lao isan, strs, verifiler TM plus PCR amplification kit, Laos

OPEN ACCESS

Edited by:
Guanglin He,
Nanyang Technological University, Singapore

Reviewed by:
Daixin Huang,
Huangdaixin, China
Xiaoye Jin,
Xi'an Jiaotong University Health Science Center, China
Lingxiang Wang,
Fudan University, China

****Correspondence:***
Wibhu Kutanan
wibhu@kku.ac.th

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *27 May 2022*
Accepted: *20 June 2022*
Published: *07 July 2022*

Citation:
Than KZ, Muisuk K, Woravatin W, Suwannapoom C, Srikummool M, Srithawong S, Lorphengsy S and Kutanan W (2022) Genetic Structure and Forensic Utility of 23 Autosomal STRs of the Ethnic Lao Groups From Laos and Thailand. Front. Genet. 13:954586. doi: 10.3389/fgene.2022.954586

INTRODUCTION

As lands with much common history, the areas of present-day Laos and Thailand share multiple cultural and historical perspectives (Arnamwat, 1985; Teerawit, 2001) and a border delineated by high mountains and the Mekong River. With a geography that encompasses both upland and lowland areas located in the heart of Mainland Southeast Asia (MSEA), both countries are served with land suitable for human occupations and harbor multiple diverse ethnolinguistic groups. With population sizes of ~6.86 million in Laos and 68.62 million in Thailand (Eberhard et al., 2020), there are 85 and 70 different languages in Laos and Thailand, respectively. All of these languages are classified as belonging to five major language families: Tai-Kadai (TK), Austroasiatic (AA), Sino-Tibetan (ST), Hmong-Mien (HM) and Austronesian (AN). The most common language family is the TK language with ~4.29 million speakers in Laos and 46.69 million in Thailand, while AA is the second most commonly spoken language in Laos (~1.71 million) and Thailand (~2.07 million).

However, the number of AA languages is higher (49 in Laos and 26 in Thailand) than TK languages (20 in Laos and 16 in Thailand), reflecting greater diversification of AA than TK. The ST and HM languages are less spoken in both countries (ST: 11 in Laos and 19 in Thailand; HM: 4 in Laos and 3 in Thailand) while the AN family is restricted to southern Thailand (6 languages) (Eberhard et al., 2020).

In Laos, geographic criteria are generally used to grouping populations; the major TK speaking Laotians or Lao Lum (~4.3 million) inhabit in the lowland with various dialects spoken, e.g., Luang Prabang, Vientiane and Savannakhet (Eberhard et al., 2020). There are ~1.7 million AA speaking Laotians or Lao Thoeng which refers to mid-landers or uplanders, while ~0.2 million of the HM and ST speaking Laotians are highlanders known as Lao Sung (Schliesinger, 2003; Eberhard et al., 2020). In Thailand, the major TK speaking groups in four different regions are known as Khonmueang in the North, Lao Isan in the Northeast, Central Thai in the Central, and Southern Thai or Khon Tai in the South. Among those four major Thai groups, Lao Isan is ethnically closer to Laotian. Lao Isan are ethnically Lao but citizens of Thailand; they were historically relocated from Laos during the 14th to 18th century A.D. (Vallibhotama, 1989; Myers, 2005; Mishra, 2010). Both Laotian and Lao Isan are regarded as the ethnic Lao group which makes up ~62.53% of the total population in Laos and a major group, with a population of ~15 million, in the northeast of Thailand (Rakow, 1992; Schliesinger, 2003; Schlemmer, 2017; Eberhard et al., 2020).

Beside the Lao Isan, northeastern Thailand especially in the lower part is home to ~1.4 million AA-speaking Khmer people (Eberhard et al., 2020) and ~400,000 AA-speaking Kuy or Suay people who are presently trilingual, speaking both Khmer and Lao in addition to Kuy language (Premsrirat, 1997; Eberhard et al., 2020). Numerous archaeological sites since around 6th century A.D. evidently attest that the Khmer are native to present-day northeastern Thailand prior to the arrival of the Lao (Coedes, 1968) whereas the Kuy migrated from southern Laos to northeastern Thailand around the 17th to 18th century A.D. (Sa-ard, 1984). Generally speaking, the AA-speaking Kuy in Laos nowadays are also one of the Lao Thoeng groups.

Several previous genetic studies of Laotian and Lao Isan were based on mitochondrial (mt) DNA and Y chromosome (Bodner et al., 2011; Kutanan et al., 2014; Kutanan et al., 2017; Kutanan et al., 2019; Kutanan et al., 2021) and genome-wide data (McColl et al., 2018; Tätte et al., 2019; Kutanan et al., 2021). However, data from forensic microsatellites or short tandem repeats (STRs) has been much less published, especially with populations from Laos (Srithawong et al., 2015, 2020). Furthermore, previous STRs data were based on 15 autosomal loci. However, more STRs can increase resolution for complex paternity cases, e.g., determining the true relationship between parent-child, siblings or half sibling

TABLE 1 | General information of the studied populations, genetic diversity indices and forensic parameters.

Ethnicity	Code	Sample size	Location	Linguistic classification	Average H_E	Total allele	Gene Diversity (SD)	CMP	CPE	Loci Departed from HWE
Laotian	LA1	39	Savannakhet, southern Laos	Austroasiatic	0.7867	182	0.7811 (0.3890)	3.5854×10^{-24}	0.999999999902770	*PentaE*
Laotian	LA2	46	Luang Prabang, northern Laos	Tai-Kadai	0.7861	199	0.7853 (0.3895)	6.0225×10^{-25}	0.999999998620348	
Laotian	LA3	88	Vientiane, central Laos	Tai-Kadai	0.7962	213	0.7886 (0.3890)	1.2984×10^{-26}	0.999999998315516	
Lao Isan	LAO1	41	Roi-Et, northeastern Thailand	Tai-Kadai	0.7928	197	0.7878 (0.3912)	6.5469×10^{-25}	0.999999967018448	
Lao Isan	LAO2	45	Buriram, northeastern Thailand	Tai-Kadai	0.7901	198	0.7884 (0.3911)	5.2816×10^{-25}	0.999999969814718	
Lao Isan	LAO3	41	Ubon Ratchathani (1), northeastern Thailand	Tai-Kadai	0.7859	200	0.7763 (0.3858)	7.8522×10^{-25}	0.999999996073971	
Lao Isan	LAO4	48	Chaiyaphum, northeastern Thailand	Tai-Kadai	0.7946	212	0.7938 (0.3934)	1.7946×10^{-25}	0.999999999331283	
Lao Isan	LAO5	41	Loei, northeastern Thailand	Tai-Kadai	0.7812	182	0.7782 (0.3867)	3.1171×10^{-24}	0.999999999972446	*D5S818*
Lao Isan	LAO6	28	Ubon Ratchathani (2), northeastern Thailand	Tai-Kadai	0.7841	179	0.7792 (0.3894)	2.704×10^{-23}	0.999999999798795	
Lao Isan	LAO7	34	Nakhon Ratchasima, northeastern Thailand	Tai-Kadai	0.7912	188	0.7889 (0.3928)	2.0605×10^{-24}	0.999999999339757	

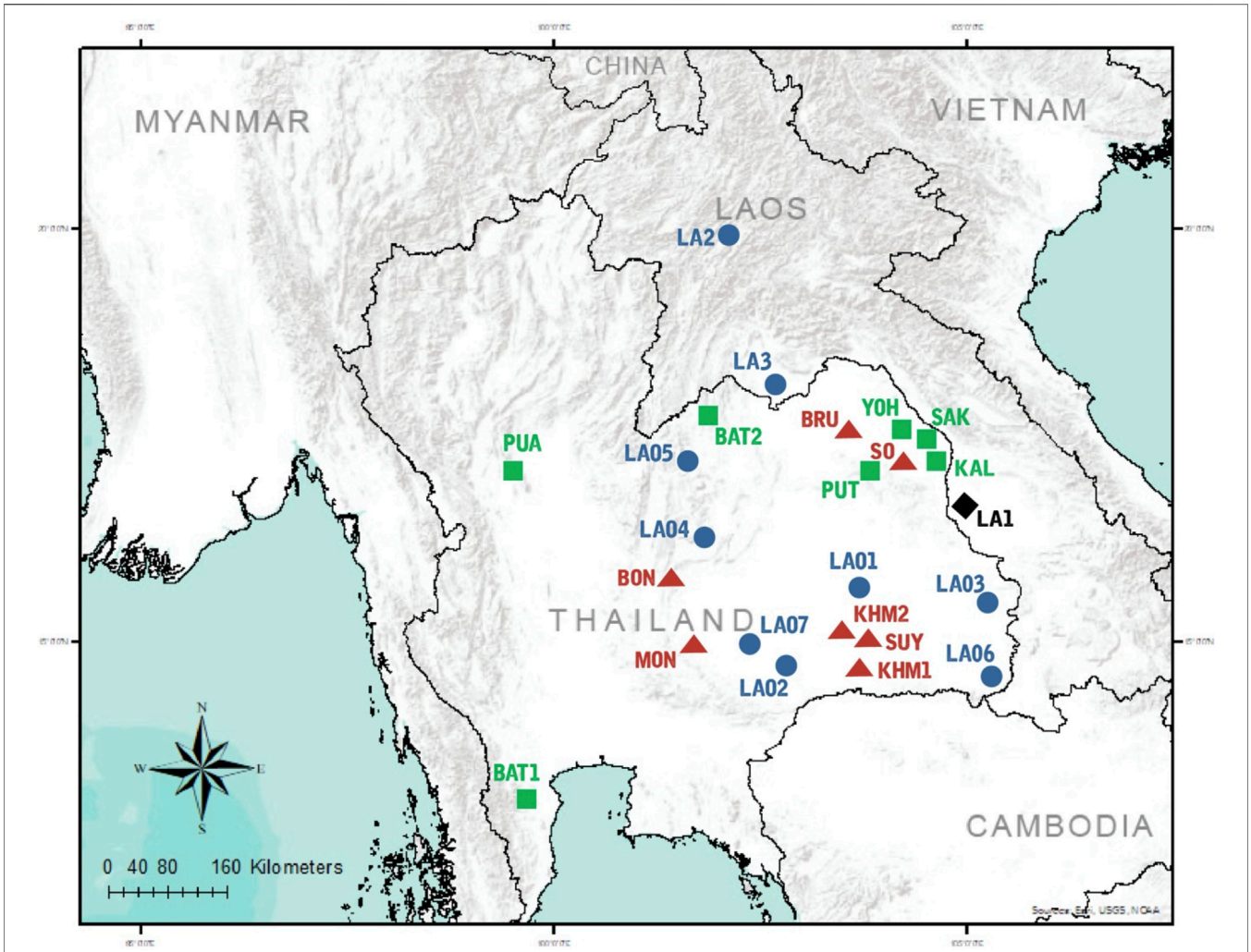


FIGURE 1 | Map of sampling locations. 10 populations were sampled in the present study, combined with data from 14 populations collected in previous studies. Population codes are given in .**Supplementary Table S16** Black diamond, blue circles, green squares, and red triangles represent AA-speaking Laotian, ethnic Lao group, compared TK-speaking and AA-speaking populations respectively.

(O'Connor et al., 2010; Alsafiah et al., 2019) and enhance the discrimination power in cases of partial DNA profiles and DNA mixtures (Haidar et al., 2021). Several forensic kits have been developed to expand the number of STR markers required for the Combined DNA Index System (CODIS) and for the European Standard Set (ESS) (Gill et al., 2006; Hares, 2015), e.g., VeriFiler Plus PCR Amplification Kit (Applied Biosystems, United States) which is a six-dye kit that can amplify 23 autosomal STR loci (*D3S1358, vWA, D16S539, CSF1PO, TPOX, D8S1179, D21S11, D18S51, D2S441, D19S433, TH01, FGA, D22S1045, D5S818, D13S317, D7S820, D10S1248, D1S1656, D12S391, D2S1338, D6S1043, Penta D, Penta E*), 1 insertion/deletion polymorphic marker on the Y chromosome (Y indel) and Amelogenin (sex-determining marker). This kit has already been validated following SWGDAM (Scientific Working Group on DNA Analysis Methods) guidelines (Al Janaahi, 2019; Alsafiah, 2019; Alsafiah et al., 2019; Green et al., 2021). This present study newly generated the genetic data of 451 Lao Isan from northeastern Thailand and Laotian from Laos, using the battery of VeriFiler Plus PCR Amplification Kit. We also reported population genetics results and established allelic frequency of 23 autosomal STRs which is beneficial for future forensic investigation in Thailand and Laos.

MATERIALS AND METHODS

Samples Collection and Extraction

A total 451 genotypes that were newly generated for 23 autosomal STR loci belonged to ten populations: seven Lao Isan and three Laotian (two Lao Lum and one Lao Thoeng). Buccal swab samples of 39 volunteers from one Laotian population were newly collected from Savannakhet Province, southern Laos; this Austroasiatic-speaking Lao population (LA1) is called Lao Thoeng. To recruit samples,

we interviewed volunteers to include subjects who were unrelated for at least two generations and then collected buccal samples with informed consent. Genomic DNA was extracted using the Gentra Puregene Buccal Cell Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. An additional 412 genomic DNA samples from other Laotian and Lao Isan populations were retrieved from previous studies (Kutanan et al., 2017; Srithawong et al., 2020; Kutanan et al., 2021) (**Table 1**). Ethical approval for this study was provided by Khon Kaen University for Lao Isan and Naresuan University for Laotian.

DNA Amplification and STR Genotyping

We amplified 23 autosomal STR loci of all genomic samples using Verifiler™ plus PCR amplification kit in a reaction volume of 25 µl with 5 µl of VeriFiler™ Plus PCR Master Mix (Applied Biosystems), 2.5 µl of the primer mix, and the

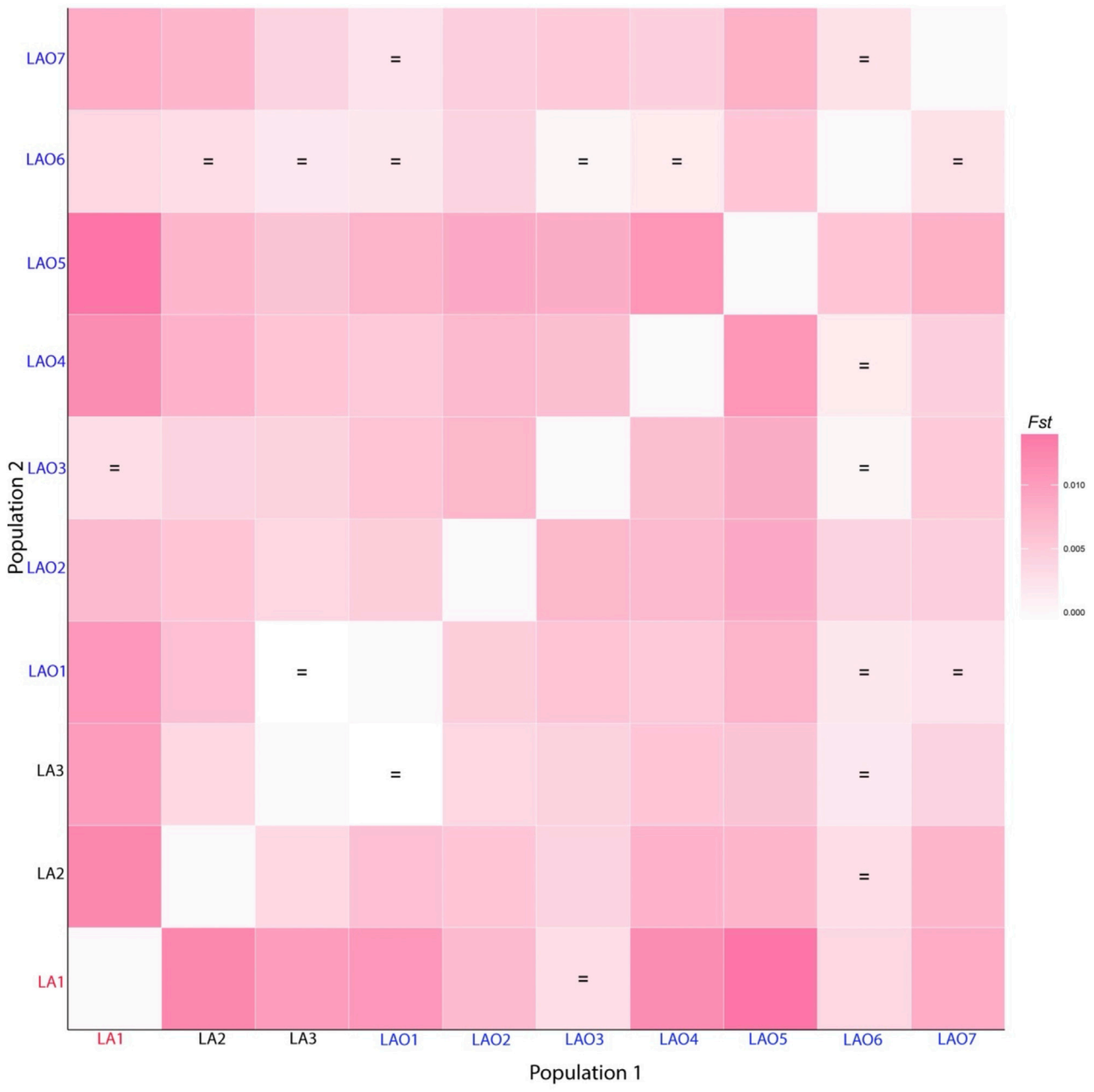


FIGURE 2 | Heat plot of F_{st} values between 10 populations from Laos and Thailand. Population names are color-coded as red, black and blue representing AA-speaking Laotian, TK-speaking Laotians and Lao Isan, respectively. The " = " symbol indicates F_{st} values that are not significantly different from zero ($p > 0.05$).

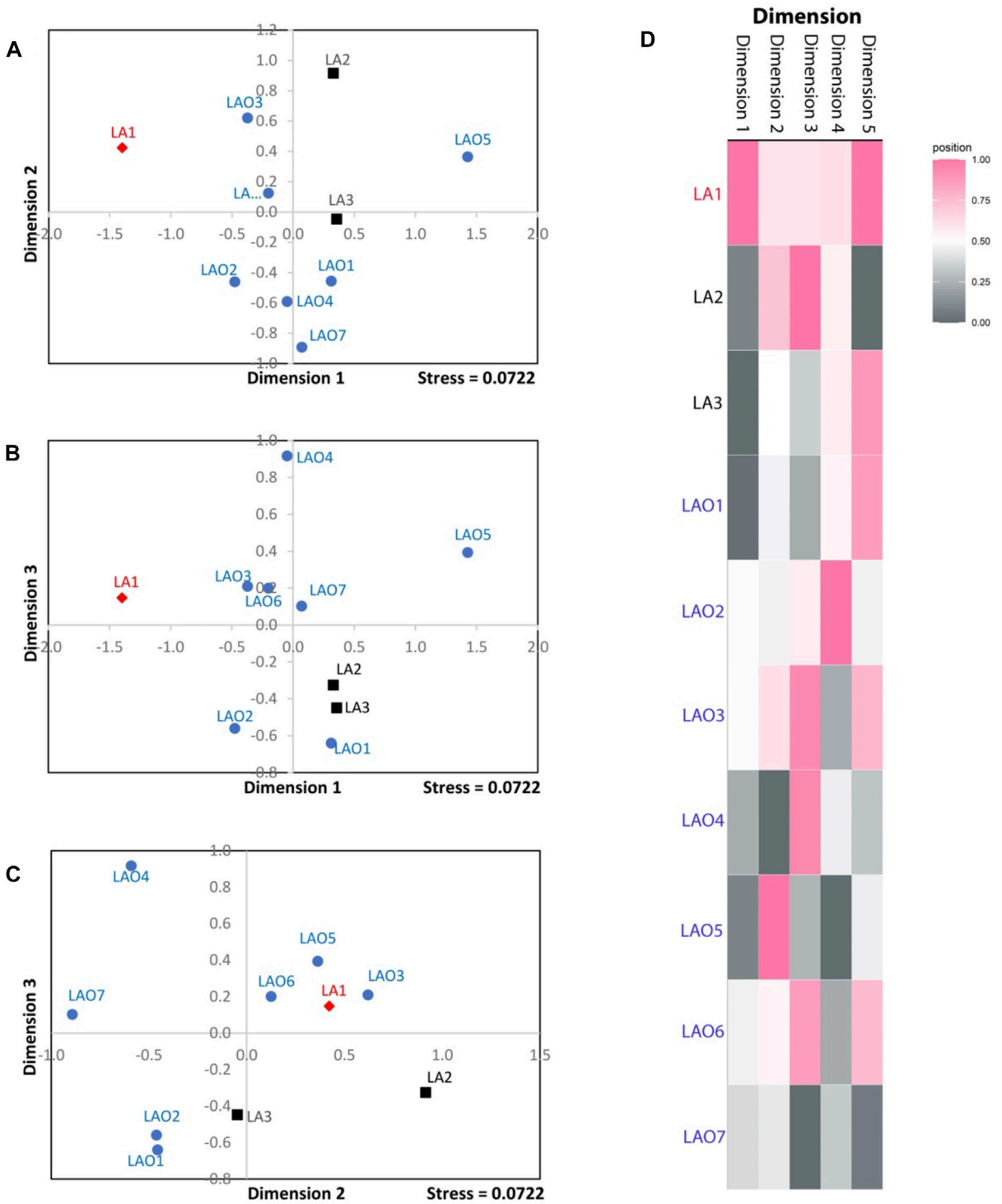


FIGURE 3 | The three-dimensional MDS plot of dimension 1 vs. 2 **(A)**, 1 vs. 3 **(B)** and 2 vs. 3 **(C)** of total 10 populations. The heat plot of standardized values of MDS with five dimensions **(D)**. Red diamond, black squares, and blue circles represent AA-speaking Laotian, TK-speaking Laotians and Lao Isan populations, respectively.

remaining 17.5 μl composed of DNA template and water to adjust the DNA input amount to reach 500 pg. Amplification process was carried out on GeneAmp™ PCR System 9700 in the following conditions; 95°C for 1 min; 2 cycles: 96°C for 10 s, 62°C for 90 s; 27 cycle: 96°C for 10 s, 59°C for 90 s; 60°C for 5 min; 4°C for ∞. The PCR products were genotyped by multi-capillary

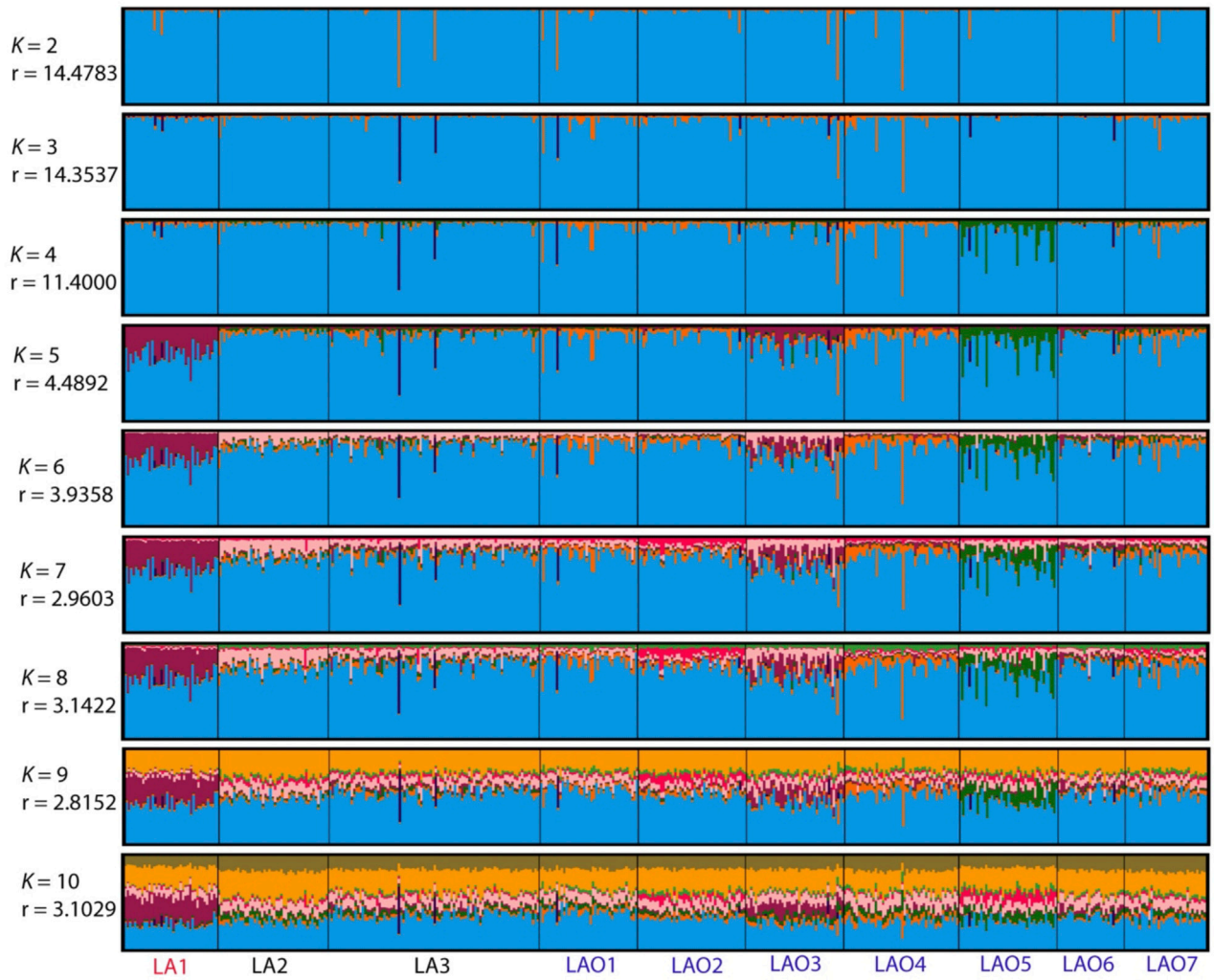


FIGURE 4 | The STRUCTURE result of 10 populations from number of cluster (*K*) = 2 to 10. Each individual population is represented by a single column divided into segments whose size and colour correspond to the relative proportion of a particular cluster. Populations are separated by black lines. Population names are color-coded; red, black and blue represent AA-speaking Laotian, TK-speaking Laotians and Lao Isan populations, respectively. The parameters "*r*" indicate the informativeness of sampling location data.

electrophoresis in an ABI 3500 genetic analyzer (Applied Biosystems). The genotyping data were analyzed by Gene Mapper software v.3.7 (Applied Biosystem).

Statistical Analyses

Arlequin v.3.5.2.2 (Excoffier and Lischer, 2010) was used to calculate allele frequency, Hardy–Weinberg equilibrium (HWE) *p* values, observed heterozygosity (H_O), expected heterozygosity (H_E), total alleles, and gene diversity (GD). Significant levels for the HWE were adjusted according to the sequential Bonferroni correction ($\alpha = 0.05/23$) (Rice, 1989). We used *STRAF* (http://cmpg.unibe.ch/shiny/STRAF/), an online tool for STR data analysis (Gouy and Zieger, 2017), to compute several forensic parameters, i.e., power of discrimination (PD), matching probability (MP), polymorphic information content (PIC), power of exclusion (PE), and typical paternity index (TPI).

Among all 10 Lao Isan and Laotian populations, we first computed a genetic distance matrix based on number of different alleles (F_{st}) using Arlequin, then plotting a matrix in three dimensions by means of multidimensional scaling (MDS) using Statistica v.10 demo (StatSoft, Inc, United States). The heatmap visualization of F_{st} and MDS values were obtained using R package (R Development Core Team). To delineate cryptic population structure, a model-based cluster analysis was investigated by STRUCTURE 2.3.4 under the following prior parameters: admixture, correlated allele frequencies, and assistance of sampling locations (LOCPRIOR model) (Pritchard et al., 2000; Falush et al., 2003; Hubisz et al., 2009). We ran ten replications for each cluster (K) from 2 to 10 with burn-in length of 100,000 iterations followed by 200,000 iterations. The STRUCTURE outputs were merged to compute a second-order rate of change

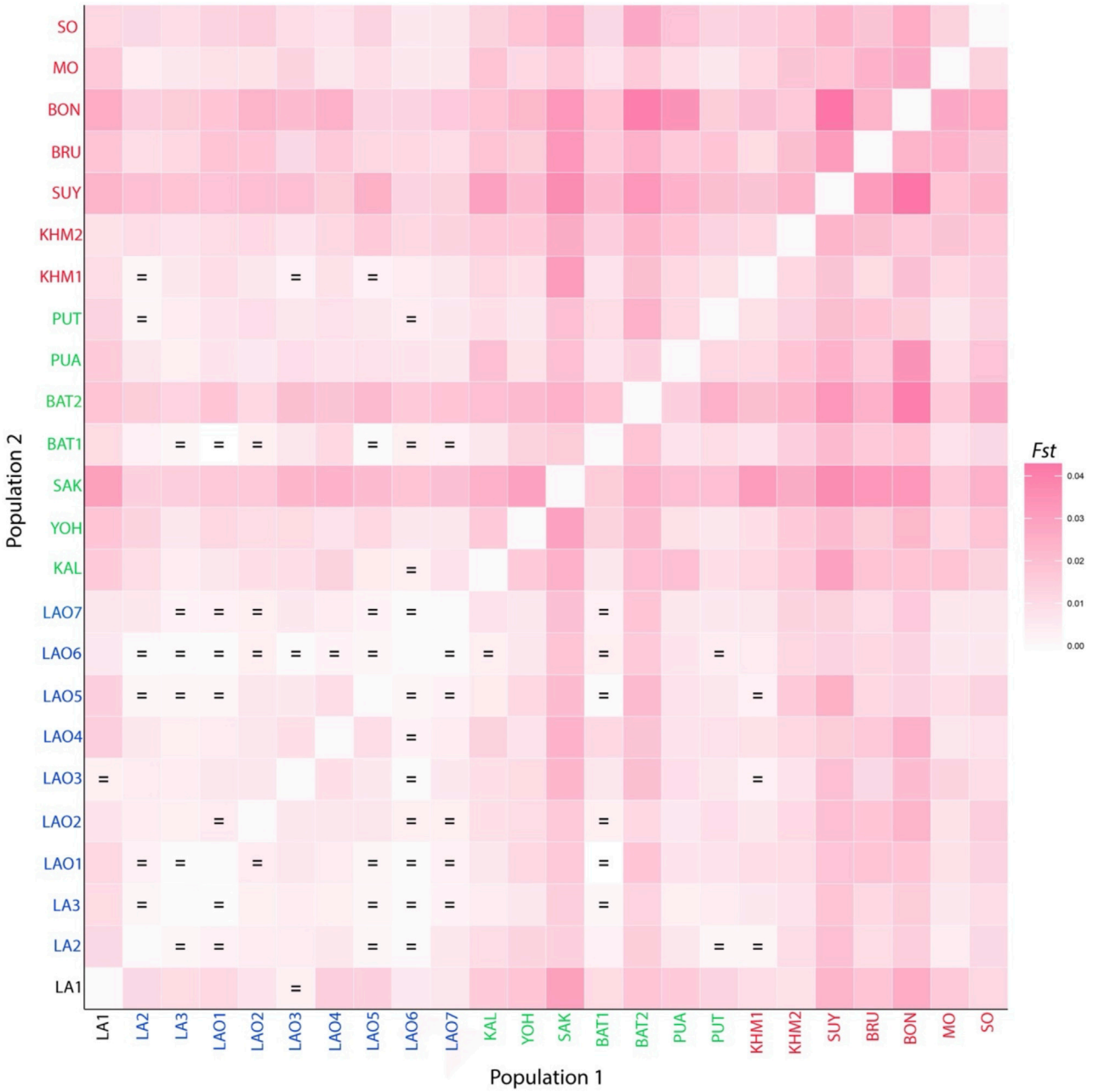


FIGURE 5 | Heat plot of F_{st} values between 24 populations from Laos and Thailand. Population names are color-coded as black, blue, green and red representing AA-speaking Laotian, ethnic Lao group, compared TK-speaking and AA-speaking populations, respectively. The " = " symbol indicates F_{st} values that are not significantly different from zero ($p > 0.05$).

logarithmic likelihood between subsequent K values ($\triangle K$) (Evanno et al., 2005) by STRUCTURE Harvester (Earl and vonHoldt, 2012) in order to discover the ideal K value in the data. CLUMPAK was used to construct a single-set result from 10 replications of STRUCTURE outputs to validate the dynamic approach determining the optimal similarity threshold for each value of K (Kopelman et al., 2015); CLUMPAK outputs were graphically modified by DISTRUCT (Rosenberg, 2003).

To reveal population relationships with other Thai populations, we collected previously published data of 15 STRs from 14 populations: BlackTai1, BlackTai2, Phutai, Phuan, Seak, Nyaw, Kaleang, Bru, Nyahkur, Mon, Soa, Khmer1, Khmer 2 and Suay (Srithawong et al., 2015; Chantakot et al., 2017; Srithawong et al., 2020). When the newly generated data of Lao Thoeng were combined with this published data, it provided a total raw 15 STRs genotype data of 1,039 samples belonging to 24 populations for subsequent analyses (**Figure 1**). As mentioned previously, the

same software and parameters were employed to estimate genetic distances and genetic structure.

Mantel test (Mantel, 1967) in Arlequin was used to estimate correlation between distance matrices of genetic vs. geography. Geographic distances (in form of great-circle distances) in Km between the approximate locations of each population were calculated from their latitudinal and longitudinal coordinates using an online tool (http://onlineconversion.com/map-greatcircle-distance.htm).

To get a more comprehensive picture of population linkages in Asia, we employed POPTREE v.2 (Takezaki et al., 2014) to build a neighbor-joining tree (NJ) based on F_{st} computation by allele frequency from 13 STRs with publicly accessible data from relevant populations (Brinkmann et al., 2002; Kim et al., 2003; Seah et al., 2003; De Ungria et al., 2005; Dobashi et al., 2005; Tie et al., 2006; Maruyama et al., 2008; Zhu et al., 2008; Untoro et al., 2009; Song et al., 2010; Kutanan et al., 2011; Yang et al., 2013; Zhai et al., 2014; Huang et al., 2015; Kutanan et al., 2015; Srithawong et al., 2015; Zhang, 2015a; Zhang, 2015b; Chantakot et al., 2017; Guo, 2017; Guo et al., 2017; Srithawong et al., 2020; Mawan et al., 2021).

RESULTS AND DISCUSSION

Genetic Diversities and Forensic Parameters

A total of 451 individual raw genotypes are provided in **Supplementary Table S1**. The allelic frequency table of 23 STR loci in 10 individual studied populations are reported in **Supplementary Table S2–S11**. Hardy–Weinberg equilibrium (HWE) tests showed no significant deviation from expected values for all 23 loci ($p > 0.05$) after Bonferroni adjustment (0.05/23 = 0.002) in exception with *PentaE* in LAO1 and *D5S818* in LAO5 (**Table 1**). The genetic diversity indices and forensic parameters, including average H_E, total alleles, gene diversity (GD), forensic parameters; combined matching probability (CMP), combined power of exclusion (CPE) are presented in **Table 1**. The average H_E is greater than 0.7 in all studied populations, ranging from 0.7962 (LA3) to 0.7812 (LAO5). Total number of alleles is highest in LA3 (213 alleles) and lowest in LAO6 (179 alleles). The GD ranges from 0.7763 ± 0.3858 in LAO3 to 0.7938 ± 0.3934 in LAO4.

The allelic frequency of 23 STR loci in the ethnic Lao groups ($n = 412$), comprising two Lao Lum (LA2-3) from Laos and seven Lao Isan (LAO1-7) from northeastern Thailand, is presented in **Supplementary Table S12**. Only two loci (*D10S1248, TPOX*) departed from the HWE ($p > 0.05$) after Bonferroni adjustment (0.05/23 = 0.002) (**Supplementary Table S12**). There are total 292 alleles, with the range of 8 alleles at *D16S539* and *D5S818* loci to 23 alleles at *FGA* locus and allelic frequency ranging from 0.0012 to 0.6163. Among the tested loci, *FGA* was the most polymorphic and discriminative locus with highest H_E (0.8821), H_O (0.9005), PIC (0.8699), TPI (5.0244), MP (0.0276), PD (0.9724) and PE (0.7964) with a combined power of discrimination (CPD) value of 0.999999999999999 and a CPE value of 0.999999995430163. The least polymorphic and discriminative locus was *TPOX* with the lowest H_E (0.5629), H_O (0.5122), PIC (0.5163), TPI (1.0250), MP (0.2391), PD (0.7609), PE (0.1984) (**Supplementary Table S12**).

We also reported the allele frequency based on the 23 STR loci of the combined Lao Isan (LAO1-7: n = 278) (**Supplementary Table S13**) and combined Laotian (LA1-3: n = 173) data (**Supplementary Table S14**). In Lao Isan dataset, there are 278 alleles varied from 8 alleles (*D16S539* and *D5S818*) to 23 alleles (*FGA*) (**Supplementary Table S13**) with the allelic frequency ranging from 0.0018 to 0.6025 and only one locus (*D10S1248*) departed from HWE after Bonferroni correction (0.05/23 = 0.002). Among the tested loci, *FGA* was the most polymorphic and discriminative locus with highest H_E (0.8837), H_O (0.8957), PIC (0.8712), TPI (4.7931), MP (0.0275), PD (0.9725) and PE (0.7866) with a CPD value of 0.999999999999999 and a CPE value of 0.999999994092542. The least polymorphic and discriminative locus was *TPOX* with the lowest H_E (0.5759), H_O (0.5324), PIC (0.5280), TPI (1.0692), MP (0.2275), PD (0.7725), PE (0.2174) (**Supplementary Table S13**).

In the combined Laotian (LA1-3) dataset, all of loci are in agreement with HWE even after Bonferroni correction. The total number of alleles is 233, varying from 6 alleles at *D3S1358, D16S539, TPOX* to 19 alleles at *FGA* and allelic frequency ranging from 0.0029 to 0.6170 (**Supplementary Table S14**). Among the tested loci, *FGA* was the most polymorphic and discriminative locus with highest H_E (0.8844), H_O (0.9017), PIC (0.8706), TPI (5.0882), MP (0.0314), PD (0.9686) and PE (0.7990) with a CPD value of 0.999999999999999 and a CPE value of 0.999999998686772. The least polymorphic and discriminative locus was *TPOX* with the lowest H_E (0.5633), H_O (0.5118), PIC (0.5143), TPI (1.0241), MP (0.2432), PD (0.7568), PE (0.1980) (**Supplementary Table S14**).

This study provides additional forensic STR loci of Lao Isan and Laotian populations. After applying Bonferroni's correction, there was absence of departure from Hardy–Weinberg equilibrium tests in several datasets (**Supplementary Table S2–S4, S6–S8, S10–S11, S14**), implying that the samples are representative and the data is credible. When the PD was larger than 0.80, a STR locus was considered highly polymorphic (Shriver et al., 1995). In the ethnic Lao dataset all loci except *TPOX* had a PD value of more than 0.80 (**Supplementary Table S12, S13**) while in the pooled Laotian populations, all other markers were found to be highly discriminative except for *D3S1358* and *TPOX,* which showed PD values lower than 0.80 (**Supplementary Table S14**). All populations had high heterozygosity (average H_E greater than 0.7) and high CPD values (more than 0.999999999999999) which reflects the high discriminatory power of these 23 loci. In addition, the CPE values were greater than 0.9999999 which indicate that these markers are undoubtedly acceptable for paternal and maternal identification in the studied populations (Al-Eitan and Tubaishat, 2016). Likewise, the CMP in individuals improved to values ranging from 2.704×10^{-23} (LAO6) to 1.2984×10^{-26} (LA3) in this study (**Table 1**), when compared with the result of CMP (ranged from 2.26×10^{-14} to 4.16×10^{-16}) in previous studies using AmpFLSTR Identifiler kit (Srithawong et al., 2020). As a result of using these 23 loci, this statistic implies that the

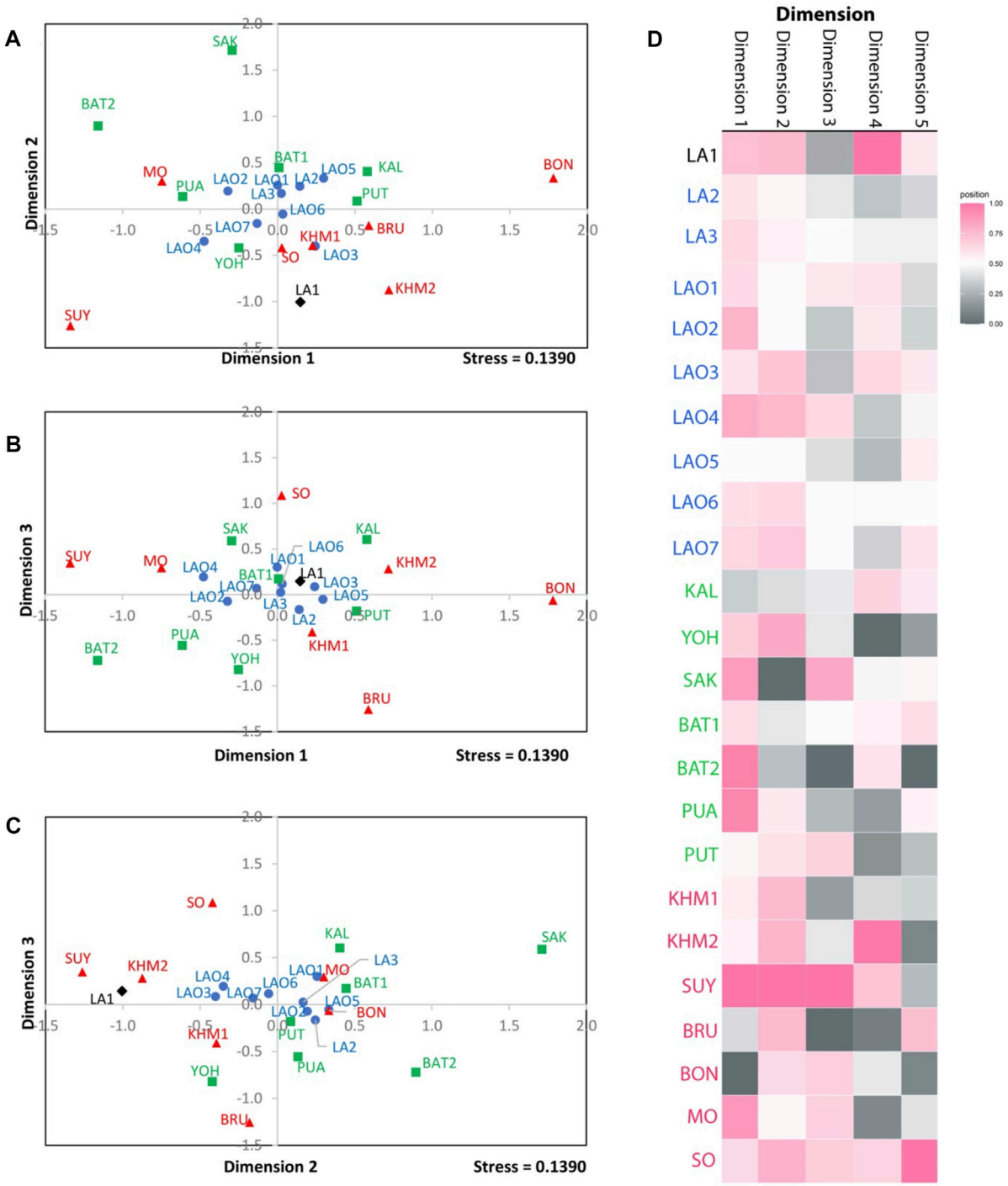


FIGURE 6 | The three-dimensional MDS plots of dimension 1 vs. 2 **(A)**, 1 vs. 3 **(B)** and 2 vs. 3 **(C)** of total 24 populations. The heat plot of standardized values of MDS with five dimensions **(D)**. Black diamond, blue circles, green squares, and red triangles represent AA-speaking Laotian, ethnic Lao group, compared TK-speaking and AA-speaking populations respectively.

chances of two people in the population having the same genetic profile are almost negligible and Verifiler^{TM} plus PCR amplification kit is an effective method for kinship analysis (Alsafiah, 2019; Alsafiah et al., 2019). The *FGA* and *TPOX* were respectively the loci showing highest and lowest TPI values in all combined datasets (**Supplementary Table S12, S13, S14**). Consistent with previous studies, *TPOX* was the least discriminative locus. While in this study *FGA* was the

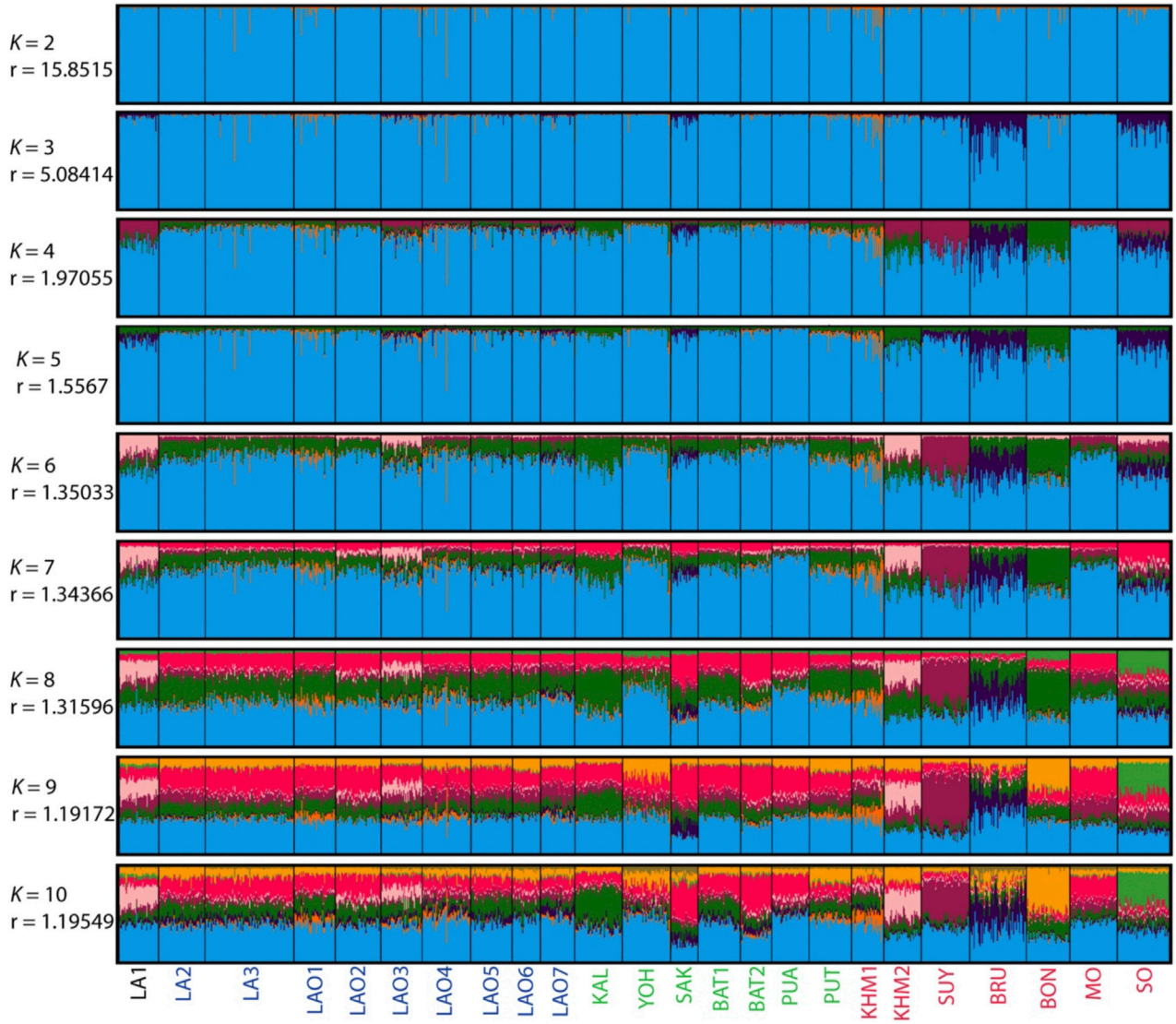


FIGURE 7 | The STRUCTURE result of 24 populations from number of cluster (*K*) = 2 to 10. Each individual population is represented by a single column divided into segments whose size and colour correspond to the relative proportion of a particular cluster. Populations are separated by black lines. Population names are color-coded; black, blue, green and red represent AA-speaking Laotian, ethnic Lao group, comparing TK-speaking and AA-speaking populations, respectively. The parameters "*r*" indicate the informativeness of sampling location data.

most informative locus, other studies have supported *PentaE* as the most powerful locus (Rodriguez et al., 2015; Sheng et al., 2018; Castillo et al., 2019; Chen et al., 2019; Naji et al., 2019; Shrivastava et al., 2019; Vu et al., 2021). Overall, the 23 STR loci in Verifiler™ plus PCR amplification kit could be valuable for forensic investigation in Thailand and Laos.

Genetic Affinity and Genetic Structure

To investigate genetic relationship between these 10 populations, three from Laos (LA1-3) and seven from northeastern Thailand (LAO1-7), we computed pairwise genetic distance (F_{st}) based on 23 STR loci and among 45 pairwise comparisons there are 36 pairs (80%) which showed statistical differences ($p < 0.05$) (**Supplementary Table S15**). In general, the results of distance-based clustering methods revealed that LA1 and LAO5 showed genetic divergence from other populations (**Figures 2**, **3**). For the model-based clustering results revealed by STRUCTURE, we ran 10 iterations of consecutive clusters (*K*) from 2 to 10. The most appropriate $\triangle K$ are K = 5 and 7 (**Supplementary Figure S1**). The LA1 showed genetic differentiation from other groups with the presence of a purple component that also saw minor emergence in LAO3 (**Figure 4**), indicating a certain relatedness between them. The AA-speaking LA1 and TK-speaking LAO3 speak different language but the geographic locations of these two populations are close (**Figure 1**); they may have contact and gene flow

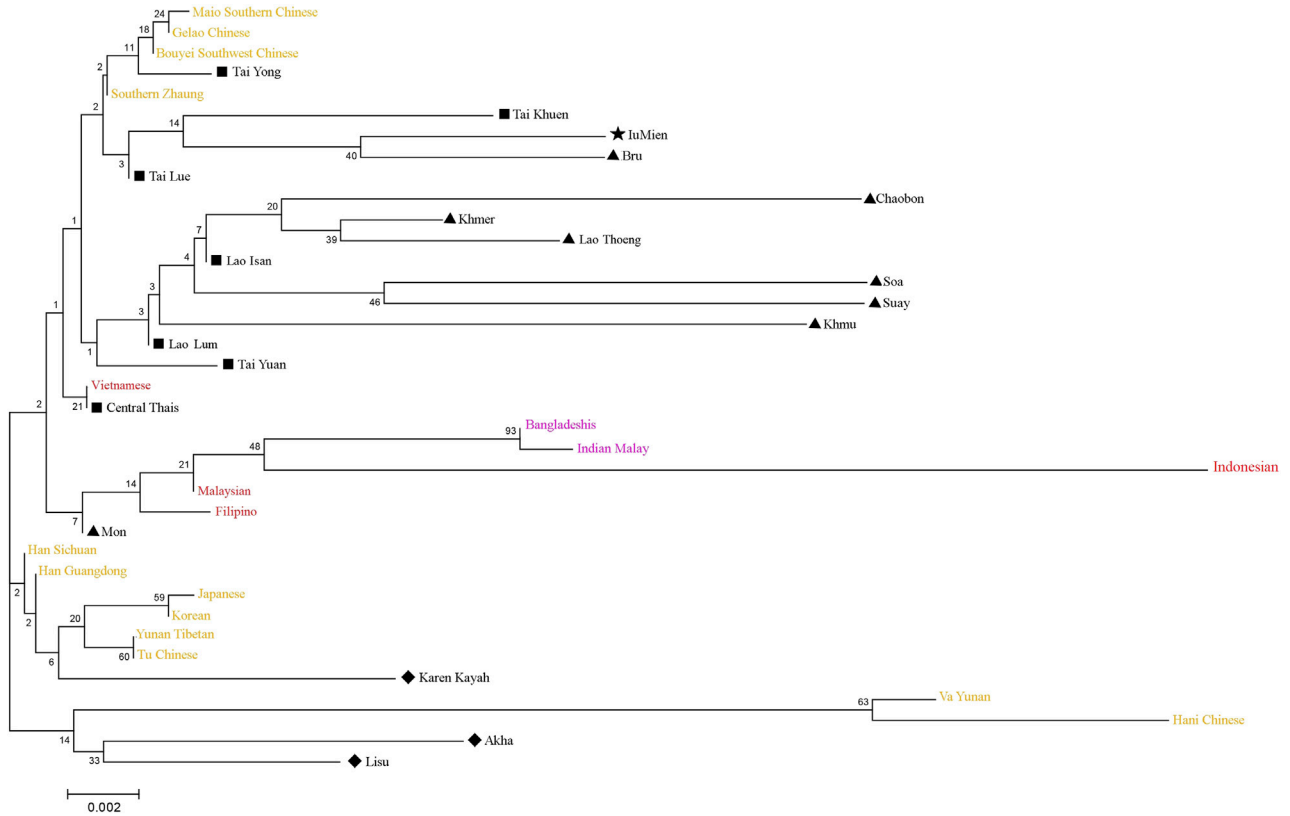


FIGURE 8 | Neighbor-joining tree constructed from F_{ST} genetic distance based on the allelic frequencies of 13 STR loci among total 37 populations. The comparative populations from China/East Asia, Southeast Asia and South Asia were represented by yellow, red and purple, respectively. Populations from Thailand were indicated with black and different symbols. Diamonds, triangles, star, and squares represent ST-, AA-, HM-, and TK-speaking populations, respectively.

between them. In addition, there is a small green component in LAO5 from Loei provinces (**Figure 4**), reflecting the differentiation of this group, consistent with previous study (Srithawong et al., 2020).

To understand comprehensive genetic relationships among populations in northeastern Thailand and Laos, we retrieved genotypic data on 15 STRs of 14 populations from previous studies for comparison with present data (**Supplementary Table S16**). Again, the pairwise F_{st} value, three dimensions MDS plots and the model-based clustering STRUCTURE based on the 15 STR loci were analyzed. Among 276 pairwise comparisons of F_{st} value, there are 263 pairs (95.29%) with statistical differences ($p < 0.05$) (**Supplementary Table S17**). The ethnic Lao groups (LA2-3 and LAO1-7) showed a narrow range of F_{st} value, reflecting close genetic relatedness. Interestingly, with a reduced number of STRs, LA1 from southern Laos still showed significant relatedness to LAO3 from Thailand (**Figure 5**; **Supplementary Table S17**) and LAO6 showed genetic similarity to all of ethnic Lao populations (**Figure 5**). The MDS plots based of F_{st} value (**Figures 6A–C**) revealed an overall pattern of genetic homogeneity of TK speaking groups and genetic heterogeneity of AA speaking populations. The positions on the margin of MDS plots of AA-speaking LA1, BON, SUY, KHM2, SO and BRU and TK-speaking SAK and BAT2 reflect the genetic divergence of the others (**Figures 6A,C,D**). The STRUCTRUE results indicated that $K = 6$ is the most appropriate $\triangle K$ for describing sub-structuring of populations (**Supplementary Figure S2**) and at $K = 6$ all populations shared a common blue component with different proportions. In general, the AA speaking groups have reduced blue component but show additional various minor components, indicating their genetic differentiation from each other and from the TK speaking groups, which is consistent with previous mtDNA, Y chromosome and genome-wide studies (Kutanan et al., 2014; Kutanan et al., 2017; Kutanan et al., 2019; Kutanan et al., 2021). Genetic drift, isolation and population interactions with other groups are factors promoting genetic differentiations of AA speaking groups in Thailand (Kutanan et al., 2018; Kutanan et al., 2021). Interestingly, the AA-speaking LA1 and KHM2 groups and TK-speaking LAO3 share minor pink components (**Figure 7**), reflecting interactions among these groups. Although there is limited historical evidence supporting interactions among LA1 and KHM2, there were several reports about genetic connections between some AA-peaking populations in northeastern Thailand (Khmer, Kuy, Soa and Bru) and Lao Isan groups (Chantakot et al., 2017;

Kutanan et al., 2019; 2021). Population admixture could explain a shared genetic component among these three groups.

Although some populations that lived in close geographic locations showed a certain similarity in genetic structures, e.g. LA1 and LAO3, Mantel testing showed that there were no correlations between genetic vs. geographic distances (r = 0.0689, $p > 0.05$ for dataset of 10 populations and r = -0.1570, $p > 0.05$ for dataset of 24 populations). Therefore, the overall genetic variation did not correlate with geography. In contrast to the previous study that reported correlation between mtDNA variations and geography in northeastern Thailand (Kutanan et al., 2014), this result indicates that genetic divergences between populations do not primarily influence by geography but other driving forces, e.g. genetic drift might be the probable driven factors, particularly in the AA speaking groups.

In sum, the present results indicate a genetic homogeneity of TK speaking groups in northeastern Thailand but some populations within the ethnic Lao groups exhibited their unique genetic characters, e.g. genetic distinction of LAO3 and LAO4 from the others and genetic similarity of LAO6 to other groups (**Figure 5**). This within-group heterogeneity was arisen from various sources as mentioned previously. We emphasize the important to study multiple samples from the same ethnic group that can provide more insights into genetic history of population.

Asian Phylogenetic Tree

A neighbor-joining (NJ) tree based on F_{st} computation by allele frequency of 13 STR loci was constructed to evaluate Asian population relationships. We pooled seven Lao Isan to one group due to their close genetic relationship and likewise, the two Laotians (LA2-3) were combined to one group (Lao Lum), while the LA1 was represented by Lao Thoeng (**Figure 8**). The Lao Isan, Lao Thoeng and Lao Lum were clustered on the same clade with AA-speaking Chaobon, Khmer, Soa, Suay and Khmu, reflecting genetic interaction between northeastern Thai and Lao populations and AA speaking groups. Previous Y chromosomal study indicated AA ancestry in Lao Isan groups (Kutanan et al., 2019) and recent genome-wide study also supports genetic relatedness among TK-speaking groups in northeastern Thailand and Laos with AA-speaking populations, especially the Khmuic-Katuic groups (Kutanan et al., 2021).

CONCLUSION

This study genotyped 23 autosomal forensic STRs using Verifiler^{TM} plus PCR amplification kit of the ethnic Lao groups from northeastern Thailand and Laos and AA-speaking Laotian from southern Laos. Although there have been some previous investigations on STRs in the region, only 15 loci were previously published. Here, we expanded the study, genotyping 23 STR loci, of which 8: *D2S441, D22S1045, D10S1248, D1S1656, D12S391, D6S1043, Penta D*, and *Penta E* were firstly reported. We generated allelic frequency table, calculated forensic parameters and investigated genetic relationships among populations. Previously there were much less forensic STRs data produced in Laos, compared to Thailand; this present study established the complete allelic frequency of STRs. Although no STRs in Laotians showed significant departure from HWE and few loci departed from HWE in other datasets, all forensic parameters indicate that this kit is suitable for forensic investigation. Genetic characterization showed that AA-speaking Laotian from southern Laos was genetically different from northern and central Lao groups who speak TK languages. In fact, the southern Laotian was more related to Lao Isan who live in vicinity. Although the Lao Isan migrated from Laos ~200 years ago, the ethnic Lao groups (Lao Isan and TK-speaking Laotian) still showed closer genetic relatedness than the other ethnolinguistic groups, reflecting a common ancestry. In sum, the STRs allelic frequency results strengthen the regional forensic database in both countries and provide the genetic backgrounds of populations that are useful for anthropological 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 studies involving human participants were reviewed and approved by Khon Kaen University and Naresuan University ethic committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WK conceived and designed the project; CS, SL, SS, and MS collected samples; SS, WW KT, and KM generated data; KT and SS carried out the data analyses; KT and WK prepared the original draft. All authors reviewed and edited the final draft. WK and KT revised the manuscript.

ACKNOWLEDGMENTS

We would like to thank village chief and participants who donated their biological samples. KT was supported by KKU Scholarship for ASEAN and GMS Countries' Personal, Academic Year 2020. CS was supported by the Thailand science research and innovation fund and the University of Phayao (Grant Nos. FF65-UoE003). W.W. was supported by National Research Council of Thailand (NRCT) (Grant Nos. N41A640275). SS was funded by the Post-Doctoral Training Program from Khon Kaen University, Thailand (PD-2564–10). WK and KM was funded by Khon Kaen University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.954586/full#supplementary-material

REFERENCES

Al Janaahi, N. S. (2019). *Forensic Evaluation of 6-Dye Chemistry Kit Composed of 23 Loci with Casework Samples*. Biology Theses, 20. Available at: https://scholarworks.uaeu.ac.ae/bio_theses/20.

Al-Eitan, L. N., and Tubaishat, R. R. (2016). Evaluation of Forensic Genetic Efficiency Parameters of 22 Autosomal STR Markers (PowerPlex Fusion System) in a Population Sample of Arab Descent from Jordan. *Aust. J. Forensic Sci.* 50, 97–109. doi:10.1080/00450618.2016.1212401

Alsafiah, H. M., Aljanabi, A. A., Hadi, S., Alturayeif, S. S., and Goodwin, W. (2019). An Evaluation of the SureID 23comp Human Identification Kit for Kinship Testing. *Sci. Rep.* 9, 16859. doi:10.1038/s41598-019-52838-7

Alsafiah, H. M. (2019). *Evaluation of DNA Polymorphisms for Kinship Testing in the Population of Saudi Arabia*. PhD thesis. The University of Central Lancashire.

Arnamwat, T. (1985). *Thai-history: From the Past to the End of Aryutthaya Era*. Bangkok: Amorn Karn Pim, 166.

Bodner, M., Zimmermann, B., Röck, A., Kloss-Brandstätter, A., Horst, D., Horst, B., et al. (2011). Southeast Asian Diversity: First Insights into the Complex mtDNA Structure of Laos. *BMC Evol. Biol.* 11 (1), 49. doi:10.1186/1471-2148-11-49

Brinkmann, B., Shimada, I., Tuyen, N., and Hohoff, C. (2002). Allele Frequency Data for 16 STR Loci in the Vietnamese Population. *Int. J. Leg. Med.* 116 (4), 246–248. doi:10.1007/s00414-002-0313-z

Castillo, A., Pico, A., Gil, A., Gusmão, L., and Vargas, C. (2019). Genetic Variation of 23 STR Loci in a Northeast Colombian Population (Department of Santander). *Forensic Sci. Int. Genet. Suppl. Ser.* 7, 33–35. doi:10.1016/j.fsigss.2019.09.015

Chantakot, P., Srithongdang, K., Srithawong, S., Boonsoda, P., Pittayaporn, P., and Kutanan, W. (2017). Genetic Divergence of Austroasiatic Speaking Groups in the Northeast of Thailand: A Case Study on Northern Khmer and Kuy. *Chiang Mai J. Sci.* 44, 1279–1294.

Chen, P., Adnan, A., Rakha, A., Wang, M., Zou, X., Mo, X., et al. (2019). Population Background Exploration and Genetic Distribution Analysis of Pakistan Hazara via 23 Autosomal STRs. *Ann. Hum. Biol.* 46, 514–518. doi:10.1080/03014460.2019.1673483

Coedes, G. (1968). *TheIndianized States of Southeast Asia*. Honolulu: East-West Center.

De Ungria, M. C. A., Roby, R. K., Tabbada, K. A., Rao-Coticone, S., Tan, M. M. M., and Hernandez, K. N. (2005). Allele Frequencies of 19 STR Loci in a Philippine Population Generated Using AmpFlSTR Multiplex and ALF Singleplex Systems. *Forensic Sci. Int.* 152 (2-3), 281–284. doi:10.1016/j.forsciint.2004.09.125

Dobashi, Y., Kido, A., Fujitani, N., Hara, M., Susukida, R., and Oya, M. (2005). STR Data for the AmpFLSTR Identifiler Loci in Bangladeshi and Indonesian Populations. *Leg. Med.* 7 (4), 222–226. doi:10.1016/j.legalmed.2005.04.001

Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a Website and Program for Visualizing STRUCTURE Output and Implementing the Evanno Method. *Conserv. Genet. Resour.* 4, 359–361. doi:10.1007/s12686-011-9548-7

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2020). *Ethnologue: Languages of the World*. 23rd ed. Dallas, Texas, USA: SIL International.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the Number of Clusters of Individuals Using the Software STRUCTURE: a Simulation Study. *Mol. Ecol.* 14, 2611–2620. doi:10.1111/j.1365-294x.2005.02553.x

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin Suite Ver 3.5: a New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Falush, D., Stepheas, M., and Pritchard, J. K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* 164, 156–187. doi:10.1093/genetics/164.4.1567

Gill, P., Fereday, L., Morling, N., and Schneider, P. M. (2006). New Multiplexes for Europe-Amendments and Clarification of Strategic Development. *Forensic Sci. Int.* 163, 155–157. doi:10.1016/j.forsciint.2005.11.025

Gouy, A., and Zieger, M. (2017). STRAF-A Convenient Online Tool for STR Data Evaluation in Forensic Genetics. *Forensic Sci. Int. Genet.* 30, 148–151. doi:10.1016/j.fsigen.2017.07.007

Green, R., Elliott, J. L., Norona, W., Go, F., Nguyen, V. T., Ge, J., et al. (2021). Developmental Validation of VeriFiler Plus PCR Amplification Kit: A 6-dye Multiplex Assay Designed for Casework Samples. *Forensic Sci. Int. Genet.* 53, 102494. doi:10.1016/j.fsigen.2021.102494

Guo, F. (2017). Allele Frequencies of 17 Autosomal STR Loci in the Va Ethnic Minority from Yunnan Province, Southwest China. *Int. J. Leg. Med.* 131, 1251–1252. doi:10.1007/s00414-017-1620-8

Guo, F., Li, J., Wei, T., Ye, Q., and Chen, Z. (2017). Genetic Variation of 17 Autosomal STR Loci in the Zhuang Ethnic Minority from Guangxi Zhuang Autonomous Region in the South of China. *Forensic Sci. Int. Genet.* 28, e51–e52. doi:10.1016/j.fsigen.2017.03.015

Haidar, M., Abbas, F. A., Alsaleh, H., and Haddrill, P. R. (2021). Population Genetics and Forensic Utility of 23 Autosomal PowerPlex Fusion 6C STR Loci in the Kuwaiti Population. *Sci. Rep.* 11, 1865. doi:10.1038/s41598-021-81425-y

Hares, D. R. (2015). Selection and Implementation of Expanded CODIS Core Loci in the United States. *Forensic Sci. Int. Genet.* 17, 33–34. doi:10.1016/j.fsigen.2015.03.006

Huang, Y., Yao, J., Li, J., Wen, J., Yuan, X., and Xu, B. (2015). Population Genetic Data for 17 Autosomal STR Markers in the Hani Population from China. *Int. J. Leg. Med.* 129, 995–996. doi:10.1007/s00414-015-1176-4

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring Weak Population Structure with the Assistance of Sample Group Information. *Mol. Ecol. Resour.* 9, 1322–1332. doi:10.1111/j.1755-0998.2009.02591.x

Ireson, C. J., and Ireson, W. R. (1991). Ethnicity and Development in Laos. *Asian Surv.* 31 (10), 920–937. doi:10.2307/2645064

Keyes, C. (1967). *Isan: Regionalism in Northeastern Thailand*. New York: Department of Asian Studies, Southeast Asia Program, Cornell University.

Kim, Y. L., Hwang, J. Y., Kim, Y. J., Lee, S., Chung, N. G., Goh, H. G., et al. (2003). Allele Frequencies of 15 STR Loci Using AmpF/STR Identifiler Kit in a Korean Population. *Forensic Sci. Int.* 136 (1-3), 92–95. doi:10.1016/s0379-0738(03)00255-x

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). Clumpak: a Program for Identifying Clustering Modes and Packaging Population Structure Inferences acrossK. *Mol. Ecol. Resour.* 15, 1179–1191. doi:10.1111/1755-0998.12387

Kutanan, W., Kampuansai, J., Srikummool, M., Kangwanpong, D., Ghirotto, S., Brunelli, A., et al. (2017). Complete Mitochondrial Genomes of Thai and Lao Populations Indicate an Ancient Origin of Austroasiatic Groups and Demic Diffusion in the Spread of Tai-Kadai Languages. *Hum. Genet.* 136 (1), 85–98. doi:10.1007/s00439-016-1742-y

Kutanan, W., Ghirotto, S., Bertorelle, G., Srithawong, S., Srithongdaeng, K., Pontham, N., et al. (2014). Geography Has More Influence Than Language on Maternal Genetic Structure of Various Northeastern Thai Ethnicities. *J. Hum. Genet.* 59 (9), 512–520. doi:10.1038/jhg.2014.64

Kutanan, W., Kampuansai, J., Brunelli, A., Ghirotto, S., Pittayaporn, P., Ruangchai, S., et al. (2018). New Insights from Thailand into the Maternal Genetic History of Mainland Southeast Asia. *Eur. J. Hum. Genet.* 26 (6), 898–911. doi:10.1038/s41431-018-0113-7

Kutanan, W., Kampuansai, J., Colonna, V., Nakbunlung, S., Lertvicha, P., Seielstad, M., et al. (2011). Genetic Affinity and Admixture of Northern Thai People along Their Migration Route in Northern Thailand: Evidence from Autosomal STR Loci. *J. Hum. Genet.* 56 (2), 130–137. doi:10.1038/jhg.2010.135

Kutanan, W., Kampuansai, J., Srikummool, M., Brunelli, A., Ghirotto, S., Arias, L., et al. (2019). Contrasting Paternal and Maternal Genetic Histories of Thai and Lao Populations. *Mol. Biol. Evol.* 36, 1490–1506. doi:10.1039/molbev/msz08310.1093/molbev/msz083

Kutanan, W., Liu, D., Kampuansai, J., Srikummool, M., Srithawong, S., Shoocongdej, R., et al. (2021). Reconstructing the Human Genetic History of Mainland Southeast Asia: Insights from Genome-wide Data from Thailand and Laos. *Mol. Biol. Evol.* 38 (8), 3459–3477. doi:10.1101/2020.12.24.42429410.1093/molbev/msab124

Kutanan, W., Srikummool, M., Pittayaporn, P., Seielstad, M., Kangwanpong, D., Kumar, V., et al. (2015). Admixed Origin of the Kayah (Red Karen) in Northern Thailand Revealed by Biparental and Paternal Markers. *Ann. Hum. Genet.* 79 (2), 108–121. doi:10.1111/ahg.12100

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Res.* 27, 209–220.

Maruyama, S., Minaguchi, K., Takezaki, N., and Nambiar, P. (2008). Population Data on 15 STR Loci Using AmpF/STR Identifiler Kit in a Malay Population Living in and Around Kuala Lumpur, Malaysia. *Leg. Med.* 10, 160–162. doi:10.1016/j.legalmed.200711.00210.1016/j.legalmed.2007.11.002

Mawan, A., Prakhun, N., Muisuk, K., Srithawong, S., Srikummool, M., Kampuansai, J., et al. (2021). Autosomal Microsatellite Investigation Reveals Multiple Genetic Components of the Highlanders from Thailand. *Genes* 12, 383. doi:10.3390/genes12030383

McColl, H., Racimo, F., Vinner, L., Demeter, F., Gakuhari, T., Moreno-Mayar, J. V., et al. (2018). The Prehistoric Peopling of Southeast Asia. *Science* 361, 88–92. doi:10.1126/science.aat3628

Mishra, P. P. (2010). *The History of Thailand*. Santa Barbara: ABC-CLIO/ Greenwood, 209.

Myers, R. L. (2005). *The Isan Saga: The Inhabitants of Rural Northeast Thailand and Their Struggle for Identity, Equality and Acceptance (1964-2004) (Master's Thesis)*. San Diego, CA: San Diego State University.

Naji, M., Damji, R., Adan, A., Abu Qamar, S., and Alghafri, R. (2019). Population Genetics Data of 23 Autosomal STR Loci for Three Populations in United Arab Emirates. *Forensic Sci. Int. Genet. Suppl. Ser.* 7, 187–188. doi:10.1016/j.fsigss.2019.09.073

O'Connor, K. L., Butts, E., Hill, C. R., Butler, J. M., and Vallone, P. M. (2010). *Evaluating the Effect of Additional Forensic Loci on Likelihood Ratio Values for Complex Kinship Analysis (The 21st International Symposium on Human Identification)*.

Premsrirat, S. (1997). Linguistic Contributions to the Study of the Northern Khmer Language of Thailand in the Last Two Decades. *Mon-Khmer Stud.* 27, 129–136.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155 (2), 945–959. doi:10.1093/genetics/155.2.945

Promega (1999). Powerstats Version 1.2, Tools for Analysis of Population Statistics. Available at: https://www.promega.com.cn/products/genetic-identity.

Rakow, M. R. (1992181p). *Laos and Laotians*. Honolulu: Center for Southeast Asian Studies, School of Hawaiian, Asian and Pacific Studies, University of Hawaii at Manoa.

Rice, W. R. (1989). Analyzing Tables of Statistical Tests. *Evolution* 43, 223–225. doi:10.1111/j.1558-5646.1989.tb04220.x

Rodriguez, J. J. R. B., Salvador, J. M., Calacal, G. C., Laude, R. P., and De Ungria, M. C. A. (2015). Allele Frequencies of 23 Autosomal Short Tandem Repeat Loci in the Philippine Population. *Leg. Med.* 17, 295–297. doi:10.1016/j.legalmed.2015.02.005

Rosenberg, N. A. (2003). DISTRUCT: a Program for the Graphical Display of Population Structure. *Mole Ecol. Notes* 4, 137–138. doi:10.1046/j.1471-8286.2003.00566.x

Sa-ard, O. (1984). *Phrase to Sentence in Kuay (Surin), Master Thesis*. Thailand: Mahidol University.

Schlemmer, G. (2017). Ethnic Belonging in Laos: A Politico-Historical Perspective. Changing Lives. New Perspectives on Society, Politics and Culture in Laos. Available at: https://hal.archives-ouvertes,fr/hal-01853834.

Schliesinger, J. (2003). *Ethnic Groups of Laos: Profiles of Austro-Thai-speaking Peoples*. Bangkok, Thailand: White Lotus Press.

Seah, L. H., Jeevan, N. H., Othman, M. I., Jaya, P., Ooi, Y. S., Wong, P. C., et al. (2003). STR Data for the AmpFlSTR Identifiler Loci in Three Ethnic Groups (Malay, Chinese, Indian) of the Malaysian Population. *Forensic Sci. Int.* 138, 134–137. doi:10.1016/j.forsciint.2003.09.005

Sheng, X., Wang, Y., Zhang, J., Chen, L., Lin, Y., Zhao, Z., et al. (2018). Forensic Investigation of 23 Autosomal STRs and Application in Han and Mongolia Ethnic Groups. *Forensic Sci. Res.* 3, 138–144. doi:10.1080/20961790.2018.1428782

Shrivastava, P., Kaitholia, K., Kumawat, R. K., Dixit, S., Dash, H. R., Srivastava, A., et al. (2019). Forensic Effectiveness and Genetic Distribution of 23 Autosomal STRs Included in Verifiler PlusTM Multiplex in a Population Sample from Madhya Pradesh, India. *Int. J. Leg. Med.* 134, 1327–1328. doi:10.1007/s00414-019-02172-4

Shriver, M. D., Jin, L., Boerwinkle, E., Deka, R., Ferrell, R. E., and Chakraborty, R. (1995). A Novel Measure of Genetic Distance for Highly Polymorphic Tandem Repeat Loci. *Mol. Biol. Evol.* 12 (5), 914–920. doi:10.1093/oxfordjournals.molbev.a040268

Song, X.-b., Zhou, Y., Ying, B.-w., Wang, L.-l., Li, Y.-s., Liu, J.-f., et al. (2010). Short-tandem Repeat Analysis in Seven Chinese Regional Populations. *Genet. Mol. Biol.* 33, 605–609. doi:10.1590/s1415-47572010000400002

Srithawong, S., Muisuk, K., Srikummool, M., Mahasirikul, N., Triyarach, S., Sriprasert, K., et al. (2020). Genetic Structure of the Ethnic Lao Groups from Mainland Southeast Asia Revealed by Forensic Microsatellites. *Ann. Hum. Genet.* 84, 357–369. doi:10.1111/ahg.12379

Srithawong, S., Srikummool, M., Pittayaporn, P., Ghirotto, S., Chantawannakul, P., Sun, J., et al. (2015). Genetic and Linguistic Correlation of the Kra-Dai-Speaking Groups in Thailand. *J. Hum. Genet.* 60, 371–380. doi:10.1038/jhg.2015.32

Takezaki, N., Nei, M., and Tamura, K. (2014). POPTREEW: Web Version of POPTREE for Constructing Population Trees from Allele Frequency Data and Computing Some Other Quantities. *Mol. Biol. Evol.* 31, 1622–1624. doi:10.1093/molbev/msu093

Tätte, K., Pagani, L., Pathak, A. K., Kõks, S., Ho Duy, B., Ho, X. D., et al. (2019). The Genetic Legacy of Continental Scale Admixture in Indian Austroasiatic Speakers. *Sci. Rep.* 9, 3818. doi:10.1038/s41598-019-40399-8

Teerawit, K. (2001). *Thai-Lao Relations in Laotian Perception*. Bangkok: Chulalongkorn Printing house.

Tie, J., Wang, X., and Oxida, S. (2006). Genetic Polymorphisms of 15 STR Loci in a Japanese Population. *J. Forensic Sci.* 51, 188–189. doi:10.1111/j.1556-4029.2005.00037.x

Untoro, E., Atmadja, D. S., Pu, C.-E., and Wu, F.-C. (2009). Allele Frequency of CODIS 13 in Indonesian Population. *Leg. Med.* 11 (Suppl. 1), S203–S205. doi:10.1016/j.legalmed.2009.01.007

Vallibhotama, S. (1989). *A Northeastern Site of Civilization: New Archeolgical Evidence to Change the Face of Thai History*. Bangkok, Thailand: Matichon.

Vu, T. T. H., Do, T. T. M., Nguyen, T. H., and Luyen, Q. H. (2021). Allele Frequencies of 23 Short Tandem Repeat Loci in the Vietnamese Kinh Population. *Forensic Sci. Int. Rep.* 3, 100210. doi:10.1016/j.fsir.2021.100210

Yang, L., Zhao, Y., Liu, C., Chan, D. W. T., Chan, M., and He, M. (2013). Allele Frequencies of 15 STRs in Five Ethnic Groups (Han, Gelao, Jing, Shui and Zhuang) in South China. *Forensic Sci. Int. Genet.* 7, e9–e14. doi:10.1016/j.fsigen.2012.10.009

Zhai, D., Yang, J., Huang, Y., Chen, L., Wu, D., Wu, J., et al. (2014). The Allele Frequency of 15 STRs Among Three Tibeto-Burman-Speaking Populations from the Southwest Region of Mainland China. *Forensic Sci. Int. Genet.* 13, e22–e24. doi:10.1016/j.fsigen.2014.09.001

Zhang, L. (2015b). Population Data for 15 Autosomal STR Loci in the Bouyei Ethnic Minority from Guizhou Province, Southwest China. *Forensic Sci. Int. Genet.* 17, 108–109. doi:10.1016/j.fsigen.2015.04.006

Zhang, L. (2015a). Population Data for 15 Autosomal STR Loci in the Dong Ethnic Minority from Guizhou Province, Southwest China. *Forensic Sci. Int. Genet.* 16, 237–238. doi:10.1016/j.fsigen.2015.02.005

Zhu, B., Yan, J., Shen, C., Li, T., Li, Y., Yu, X., et al. (2008). Population Genetic Analysis of 15 STR Loci of Chinese Tu Ethnic Minority Group. *Forensic Sci. Int.* 174, 255–258. doi:10.1016/j.forsciint.2007.06.013

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

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

Copyright © 2022 Than, Muisuk, Woravatin, Suwannapoom, Srikummool, Srithawong, Lorphengsy and Kutanan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 12 August 2022
DOI 10.3389/fgene.2022.967684

Check for updates

OPEN ACCESS

EDITED BY
Guanglin He,
Nanyang Technological University, Singapore

REVIEWED BY
Lan Zhang,
Southwest Jiaotong University, China
Haiyang Yu,
Tianjin University of Traditional Chinese Medicine, China

*CORRESPONDENCE
Ding Bai,
baiding@scu.edu.cn
Yan Zhang,
zhang.yan@scu.edu.cn

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

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 13 June 2022
ACCEPTED 27 June 2022
PUBLISHED 12 August 2022

CITATION
Wang P, Sun X, Miao Q, Mi H, Cao M, Zhao S, Wang Y, Shu Y, Li W, Xu H, Bai D and Zhang Y (2022), Novel genetic associations with five aesthetic facial traits: A genome-wide association study in the Chinese population.
Front. Genet. 13:967684.
doi: 10.3389/fgene.2022.967684

COPYRIGHT
© 2022 Wang, Sun, Miao, Mi, Cao, Zhao, Wang, Shu, Li, Xu, Bai and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Novel genetic associations with five aesthetic facial traits: A genome-wide association study in the Chinese population

Peiqi Wang[1,2†], Xinghan Sun[3,4†], Qiang Miao[5†], Hao Mi[4], Minyuan Cao[2], Shan Zhao[2], Yiyi Wang[6], Yang Shu[2], Wei Li[6], Heng Xu[2,5], Ding Bai[1*] and Yan Zhang[7,8*]

[1]State Key Laboratory of Oral Diseases & National Clinical Research Center for Oral Diseases, West China Hospital of Stomatology, Sichuan University, Chengdu, China, [2]State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu, China, [3]Genomic & Phenomic Data Center, Chengdu 23Mofang Biotechnology Co., Ltd, Chengdu, China, [4]Department of Biobank, Chengdu 23Mofang Biotechnology Co., Ltd, Chengdu, China, [5]Department of Laboratory Medicine/Research Center of Clinical Laboratory Medicine, West China Hospital, Sichuan University, Chengdu, China, [6]Department of Dermatology, Rare Disease Center, West China Hospital, Sichuan University, Chengdu, China, [7]Lung Cancer Center, West China Hospital, Sichuan University, Chengdu, China, [8]State Key Laboratory of Biotherapy, Department of Thoracic Oncology, Cancer Center, West China Hospital, Sichuan University, Chengdu, China

Background: The aesthetic facial traits are closely related to life quality and strongly influenced by genetic factors, but the genetic predispositions in the Chinese population remain poorly understood.

Methods: A genome-wide association studies (GWAS) and subsequent validations were performed in 26,806 Chinese on five facial traits: widow's peak, unibrow, double eyelid, earlobe attachment, and freckles. Functional annotation was performed based on the expression quantitative trait loci (eQTL) variants, genome-wide polygenic scores (GPSs) were developed to represent the combined polygenic effects, and single nucleotide polymorphism (SNP) heritability was presented to evaluate the contributions of the variants.

Results: In total, 21 genetic associations were identified, of which ten were novel: *GMDS-AS1* (rs4959669, $p = 1.29 \times 10^{-49}$) and *SPRED2* (rs13423753, $p = 2.99 \times 10^{-14}$) for widow's peak, a previously unreported trait; *FARSB* (rs36015125, $p = 1.96 \times 10^{-21}$) for unibrow; *KIF26B* (rs7549180, $p = 2.41 \times 10^{-15}$), *CASC2* (rs79852633, $p = 4.78 \times 10^{-11}$), *RPGRIP1L* (rs6499632, $p = 9.15 \times 10^{-11}$), and *PAX1* (rs147581439, $p = 3.07 \times 10^{-8}$) for double eyelid; *ZFHX3* (rs74030209, $p = 9.77 \times 10^{-14}$) and *LINC01107* (rs10211400, $p = 6.25 \times 10^{-10}$) for earlobe attachment; and *SPATA33* (rs35415928, $p = 1.08 \times 10^{-8}$) for freckles. Functionally, seven identified SNPs tag the missense variants and six may function as eQTLs. The combined polygenic effect of the associations was represented by GPSs and contributions of the variants were evaluated using SNP heritability.

Conclusion: These identifications may facilitate a better understanding of the genetic basis of features in the Chinese population and hopefully inspire further genetic research on facial development.

KEYWORDS

facial trait, aesthetics, genome-wide association study, genome-wide polygenic score, widow's peak

1 Introduction

Facial features exhibit a higher degree of variability than other physical features, thus making human faces unique and recognizable. Appearance variations impact quality of life, in most cases, from the perspective of aesthetics. Although sometimes acquired over the lifespan due to external factors, the variation is closely connected with the inherited complexity of facial morphogenesis (Weinberg et al., 2013; Cole et al., 2017). The correlation has been extensively researched in genetic studies and experimental animal models (Weinberg et al., 2019), and a thorough understanding of the genetic basis of specific facial traits provides insights into, for instance, the mechanisms of facial morphogenesis as well as biometrics and forensic science (Kayser and De Knijff, 2011; Claes, 2014; Sturm and Duffy, 2018).

Despite evidence accumulated to illustrate the association between facial traits and genetic variants (Liu et al., 2012; Huang et al., 2021), a considerable fraction remains to be discovered. Our study aimed at five aesthetic facial features: widow's peak, unibrow, double eyelid, earlobe attachment, and freckles. To date, the correlated genetic factors involved in some of these traits have been studied. For instance, unibrow has been reported with associations in 2q36 near the *PAX3* gene (Adhikari et al., 2016). Regarding eyelid trait that has a pronounced level of variation in East Asians, genome-wide association studies (GWASs) have revealed *HOXD-MTX2* to be relevant to eyelid curvature in Koreans (Seongwon et al., 2018) and *EMX2* associated with eyelid folding in Japanese (Chihiro et al., 2018). Meanwhile, a large-scale multiethnic GWAS revealed multiple loci associated with earlobe attachment harboring several candidate genes (e.g. *MRPS22*, *EDAR*, and *PAX9*) (Shaffer et al., 2017). Moreover, several variants of pigmentary genes, such as *BNC2*, *IRF4*, and *MC1R*, have been identified by recent studies, especially in Caucasians (Maarten et al., 2001; Eriksson et al., 2010; Jacobs et al., 2015; Kim et al., 2022), while only a few studies have been performed in Asians, mainly in Japanese and Korean population (Chihiro et al., 2018; Shido et al., 2019; Shin et al., 2021). In spite of the previous findings, the genetic background of these facial traits remains far from fully understood, especially in the Chinese population. Besides, although long been understood to have a genetic basis, genetic predispositions to widow's peak, an important aesthetic trait, have not been reported.

Therefore, we performed a large-scale GWAS on the Chinese population to gain insights into genetic variants contributing to the five aesthetic facial traits (widow's peak, unibrow, double eyelid, earlobe attachment, and freckles). Functional annotation of the genome-wide significant single nucleotide polymorphisms (SNPs) was performed based on the expression quantitative trait loci (eQTL) variants and genome-wide polygenic scores (GPSs) were subsequently developed to represent the combined polygenic effects in these five traits (Supplementary Figure S1). In total, 21 associations were identified and ten of them were novel (Figure 1). Specifically, to our knowledge, this is the first genetic report of widow's peak, identifying *GMDS-AS1* (rs4959669, $p = 1.29 \times 10^{-49}$) and *SPRED2* (rs13423753, $p = 2.99 \times 10^{-14}$) as genome-wide significant associations in the Chinese population. The other novel associations included *FARSB* (rs36015125, $p = 1.96 \times 10^{-21}$) for unibrow; *KIF26B* (rs7549180, $p = 2.41 \times 10^{-15}$), *CASC2* (rs79852633, $p = 4.78 \times 10^{-11}$), *RPGRIP1L* (rs6499632, $p = 9.15 \times 10^{-11}$), and *PAX1* (rs147581439, $p = 3.07 \times 10^{-8}$) for double eyelid; *ZFHX3* (rs74030209, $p = 9.77 \times 10^{-14}$) and *LINC01107* (rs10211400, $p = 6.25 \times 10^{-10}$) for earlobe attachment; and *SPATA33* (rs35415928, $p = 1.08 \times 10^{-8}$) for freckles. This study was expected to facilitate a better understanding of the genetic basis of the facial features and inspire further research on the biological functions of the relevant genes.

2 Materials and methods

2.1 Subject, sample, and phenotypes

Subjects were voluntarily enrolled in the study and filled out the questionnaires designed by West China Hospital, Sichuan University. Questionnaires soliciting trait information including "widow's peak," "unibrow," "double eyelid," "earlobe attachment," and "freckles," were collected (Supplementary Methods). Subsequently, phenotypical data were filtered and merged on the grounds of the verification questions to ensure the authenticity and accuracy of the preprocessed data. The study was approved by the local ethics committee [West China Hospital, Sichuan University, approval no. 2017(241)] and all participants signed an electronic informed consent form. Methods were performed following the relevant guidelines and regulations.

2.2 DNA extraction and genotyping

Each participant donated 2 ml of their saliva into a sample tube which was later sent to the laboratory to extract DNA, and

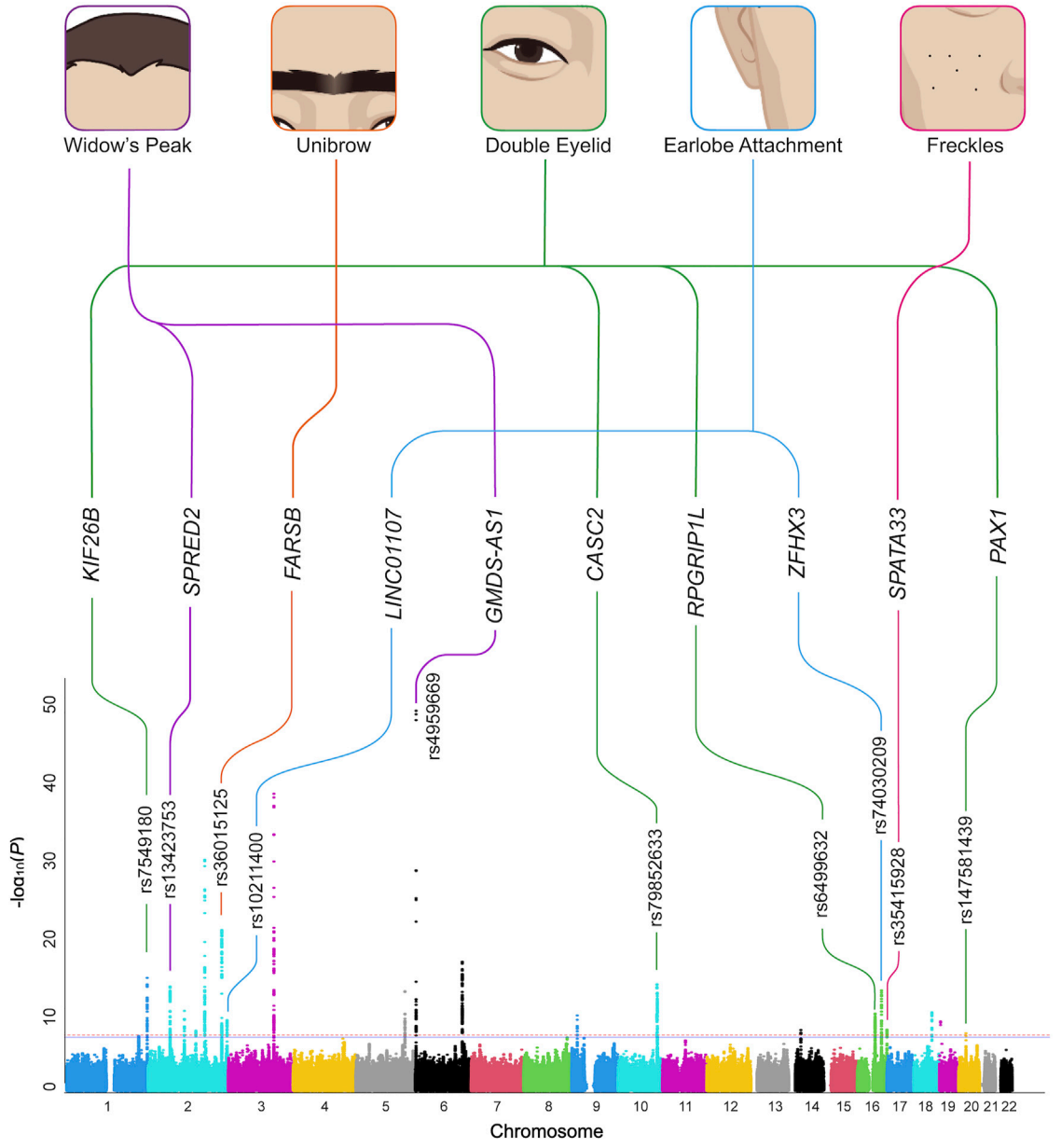


FIGURE 1
Overview of the GWAS results. The five aesthetic facial traits studied in the Chinese study sample (top) are connected with the candidate genes identified in regions with novel genome-wide significant associations. The GWAS results of the five traits were summarized on a single composite Manhattan plot (bottom). The rs ID of the SNP with the smallest *p*-value at the top of each association peak is given. GWAS, genome-wide association study. SNP, single nucleotide polymorphism.

DNA quality was determined by examining the OD260/OD280 ratio and integrity in agarose gels. Due to the long time span of the project, samples were randomly genotyped with one of the three highly correlated versions of chip arrays – Mofang v1.0, Mofang v2.0, and Mofang v2.1, which were all Affymetrix Axiom Precision Medicine Research Array (PMRA)-based high-throughput SNP chip arrays (Affymetrix, Santa Clara, CA, United States).

2.3 Quality controls

To control the genotyping quality, QCs were performed at both the individual and SNP levels: 1) SNP with genotype call rate (CR) below 0.98, 2) individual CR below 0.98, 3) gender inconsistencies, 4) number of alleles >2, 5) minor allele frequency (MAF) below 0.01 (Supplementary Table S1), 6) deviation from Hardy–Weinberg equilibrium (p-value $< 1 \times 10^{-6}$), 7)

outliers ±3 SD from the samples' heterozygosity rate, 8) individuals with cryptic relatedness, 9) outliers from multidimensional scaling (MDS) analysis (Xu et al., 2015a; Xu et al., 2015b; Qian et al., 2019; Zhang et al., 2019; Hao et al., 2021; Hertz et al., 2021) (Supplementary Methods; Supplementary Figure S1).

2.4 Genome-wide association study

For each of the five traits, 80% of the samples were randomly selected to perform GWASs as the discovery set, and the rest 20% were used for validation (Supplementary Table S2). Additional QC was further performed before the association analyses: inclusion of SNPs with CR 0.98 and MAF ≥0.01, removal of heterozygosity outliers, removal of individuals with cryptic relatedness and population structure outliers. The genotype frequency between cases and controls was compared with sex, age, and five top principal components (PCs) as covariates, by the logistic regression model using PLINK v1.90b5.4 (Supplementary Methods) (Chang et al., 2015).

2.5 Functional annotation

The genome-wide SNPs were subjected to HaploReg database (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php) (Ward and Kellis, 2012), WashU EpiGenome Browser (http://epigenomegateway.wustl.edu/browser/) (Zhou et al., 2015) and Genotype-Tissue Expression (GTEx) dataset (https://gtexportal.org/home/) (GTEx Consortium, 2015) for functional annotation.

2.6 Construction of the GPSs

28 candidate GPSs based on a pruning and thresholding (P-T) method were derived for each trait using the GWAS discovery set and discovery GWAS summary statistics from the previous step. The best scores, defined by the maximal area under curve (AUC), were applied to the validation set with 20% samples to generate a polygenic score for each individual. The individuals were binned into 20 groups according to the GPS quantile and the prevalence of each trait (Supplementary Methods) (Khera et al., 2018).

2.7 Estimation of SNP heritability

To evaluate the contribution of the variants to heritability, SNP heritability of the five traits was estimated based on the GWAS summary statistics using linkage disequilibrium score regression analysis (LDSC) (Bulik-Sullivan et al., 2015).

3 Results

In this study, a total of 26,806 Chinese volunteers who passed QC were enrolled to investigate the associations between genetic variants and the five facial traits: widow's peak (N = 11,946), unibrow (N = 7,254), double eyelid (N = 7,473), earlobe attachment (N = 9,977), and freckles (N = 8,251). All volunteers had phenotype information of one or more traits. The discovery and validation sets for each trait were randomly drawn respectively (Supplementary Table S2; Supplementary Figure S1). GWAS was performed for each trait, showing no obvious inflation (Supplementary Figure S2). Associations of genome-wide significant signals in the discovery sets with a predisposition to each trait were estimated in the respective validation sets to verify their reliability (Table 1; Supplementary Table S3).

3.1 Association analyses of the five facial traits

3.1.1 Widow's peak

Although widow's peak is regarded as a genetic heritable phenotypic pattern (Rassman et al., 2013; Kyriakou et al., 2021), genetic study of this trait is still lacking. In the present study, three loci reached genome-wide significance (p-value < 5×10^{-8}) in the discovery cohort, including the strongest signals at 6p25.2 downstream *GMDS-AS1* (top SNP: rs4959669, p = 1.29×10^{-49}), followed by 2p14 downstream *SPRED2* (top SNP: rs13423753, p = 3.0×10^{-14}) and 2q22.3 downstream *ARHGAP15* (top SNP: rs4662351, p = 1.42×10^{-8}) (Table 1; Figure 2A, Figures 3A,B). However, associations could only be reproduced for rs4959669 and rs13423753, but not for rs4662351 (Supplementary Table S3).

3.1.2 Unibrow

As for unibrow, the only previously reported significant association signal, rs2218065 at 2q36.1 (Adhikari et al., 2016), could not be validated in our cohort ($p = 3.11 \times 10^{-3}$). Instead, we identified a locus 300–500 kb downstream of rs2218065 with the top signal at *FARSB* rs36015125 (p = 1.96×10^{-21}) (Figure 2B, Figure 3C) and rs36015125 was not in linkage disequilibrium (LD) with rs2218065 (r^2<0.2).

3.1.3 Double eyelid

Concerning double eyelid, five genetic loci were genome-wide significant and validated in the validation set, four of which have not been reported to our knowledge (Figure 2C). The loci at 10q26.11 replicated signals reported in Japanese women (Chihiro et al., 2018) with the top hit at rs10749244 near *EMX2* and *RAB11FIP2* (p = 1.96×10^{-13}), in high LD with the reported variant rs1415425 (r^2 = 0.97) (Supplementary Figure S3D). A novel locus ~500 kb downstream of the reported one,

TABLE 1 Summary of previously unreported genome-wide significant loci.

Region	SNP ID	Gene(s)	Alleles	OR [95% CI] (discovery)	*p*-value (discovery)	OR (validation)	*p*-value (validation)
Widow's peak							
6p25.2	rs4959669	*GMDS-AS1*, *LINC01600*	T > C	0.49 [0.44–0.54]	1.29×10^{-49}	0.50	3.87×10^{-14}
2p14	rs13423753	*SPRED2*, *MIR4778*	G > A	0.78 [0.73–0.83]	2.99×10^{-14}	0.80	5.65×10^{-4}
Unibrow							
2q36.1	rs36015125	*FARSB*	C > G	0.69 [0.63–0.74]	1.96×10^{-21}	0.73	6.76×10^{-5}
Double Eyelid							
1q44	rs7549180	*KIF26B*	C > A	1.59 [1.42–1.78]	2.41×10^{-15}	1.31	2.12×10^{-2}
10q26.11	rs79852633	*CASC2*	G > A	1.35 [1.30–1.63]	4.78×10^{-11}	1.59	5.82×10^{-5}
16q12.2	rs6499632	*RPGRIP1L*	T > C	1.30 [1.20–1.40]	9.15×10^{-11}	0.80	5.16×10^{-3}
20p11.22	rs147581439	*PAX1*, *LINC01432*	G > C	2.06 [1.60–2.66]	3.07×10^{-8}	2.00	6.81×10^{-3}
Earlobe Attachment							
16q22.3	rs74030209	*ZFHX3*	C > T	0.77 [0.72–0.82]	9.77×10^{-14}	0.85	1.83×10^{-2}
2q37.3	rs10211400	*LINC01107*	G > T	0.75 [0.69–0.82]	6.25×10^{-10}	0.73	5.75×10^{-4}
Freckles							
16q24.3	rs35415928	*SPATA33*	C > T	1.43 [1.26–1.61]	1.08×10^{-8}	1.34	1.57×10^{-2}

overlapping the long noncoding RNA (lncRNA) gene *CASC2* (top SNP: rs79852633, $p = 4.78 \times 10^{-11}$) exhibited an independent association (Supplementary Figure S3A). The other three unreported genome-wide significant loci overlapped *KIF26B* (1q44, top SNP: rs7549180, $p = 5.75 \times 10^{-39}$), *RPGRIP1L* (16q12.2, top SNP: rs6499632, $p = 9.2 \times 10^{-11}$), and *PAX1/LINC01432* (20p11.22, top SNP: rs147581439, $p = 3.07 \times 10^{-8}$), respectively (Figure 3D; Supplementary Figure S3).

3.1.4 Earlobe attachment

For earlobe attachment, eight loci reached genome-wide significance in the discovery stage and were validated in a validation set (Figure 2D; Supplementary Table S3). Among them, two loci were novel to our knowledge, including a series of variants at 16q22.3 (top SNP: rs74030209, $p = 9.8 \times 10^{-14}$) and 2q37.3 (top SNP: rs10211400, $p = 6.3 \times 10^{-10}$) (Table 1; Figure 3E; Supplementary Figure S4A). The other six have been either previously reported or in LD with the reported SNPs in other ethnic populations (Supplementary Table S3; Supplementary Figure S4).

3.1.5 Freckles

In pursuit of genetic associations with freckles, seven genome-wide significant loci were identified in the discovery stage, while only five passed validation (Figure 2E; Supplementary Table S3), among which only one has not been previously reported (16q24.3, *SPATA33*, top SNP: rs35415928, $p = 1.1 \times 10^{-8}$) (Table 1; Figure 3F). The other significant loci that passed validation overlapped *HSPA12A* (10q25.3), *PPARGC1B* (5q32), *BNC2* (9p22), and *EMX2/RAB11FIP2* (10q26.11) (Supplementary Figure S5).

3.2 The possible impact of the genome-wide significant associations

Functionally, the variants identified by GWAS may impact the corresponding phenotype by either altering the amino acids or regulating the expression of their nearby genes (Moriyama et al., 2016; Zhu et al., 2018; Tam et al., 2019; Moriyama et al., 2021). In this study, almost all the genome-wide significant variants are located in noncoding regions, except for rs3827760, a missense mutation point of the *EDAR* gene (Supplementary Table S3).

Meanwhile, seven variants are in LD ($r^2 \geq 0.2$) with the coding variants based on LD calculations using 1,000 Genome Project data according to the HaploReg database (Supplementary Table S4) (Ward and Kellis, 2012). Therefore, we consider that the significant variants might mainly function by affecting gene expressions. Altogether, six of the genome-wide significant SNPs have GTEx eQTL associations ($p < 1 \times 10^{-4}$) in skin tissue and cultured fibroblasts that are of potential relevance to the five facial traits (Supplementary Table S5) (GTEx Consortium, 2015). Some of the SNPs show associations with only one gene. For instance, the previously unreported double eyelid-associated SNP rs6499632 is in LD with a missense variant of *RPGRIP1L* ($r^2 = 0.26$) and is a strong eQTL for *RP11-36I17.2* expression in cultured fibroblasts (Supplementary Table S4; Supplementary Figure S6). Meanwhile, some SNPs may have a tissue-specific

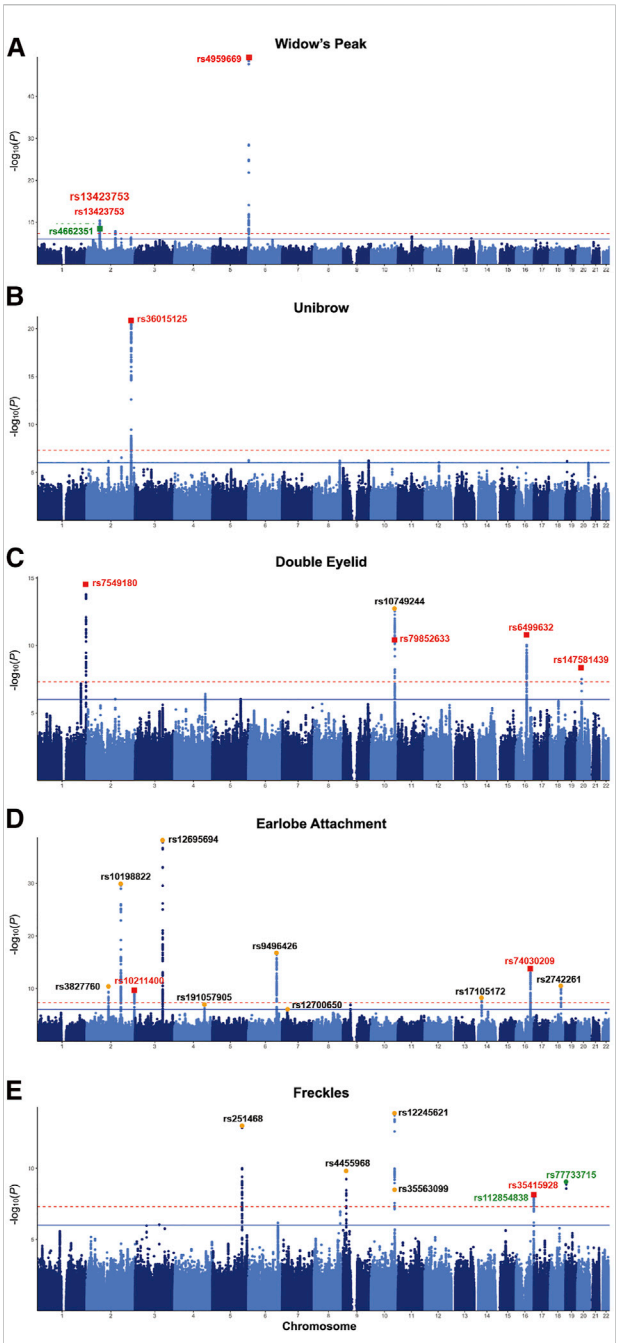


FIGURE 2
Manhattan plots of the discovery GWAS. Manhattan plot for **(A)** widow's peak; **(B)** unibrow; **(C)** double eyelid; **(D)** earlobe attachment; **(E)** freckles. Bonferroni corrected threshold and candidate threshold correspond to 7.30 and 5.30, respectively, with regard to $-\log_{10}(P)$. Previously unreported SNPs are marked RED, previously reported SNPs are dotted ORANGE, and the SNPs failing validation are marked GREEN. GWAS, genome-wide association study. SNP, single nucleotide polymorphism.

eQTL association with multiple genes. The novel freckles-associated variant rs35415928 in *SPATA33* serves as a strong eQTL for several genes in both skin tissue (sun-exposed and no-sun-exposed) and fibroblasts, showing the strongest association with *DBNDD1* (Supplementary Figure S7), a gene involved in tanning ability (Nan et al., 2009) and squamous cell carcinoma (Asgari et al., 2016). Unibrow-associated *FARSB* rs36015125 is an eQTL for *RP11-16P6.1*, *SGPP2*, and *FARSB* expression in skin tissues (sun-exposed and no-sun-exposed) and cultured fibroblasts (Supplementary Figure S6), but information regarding these genes' function in unibrow or hair appearance is unavailable.

3.3 Genome-wide polygenic score analysis

GPS was constructed to manifest the genomic polygenic effect. For each trait, we derived 28 GPS predictors based on a P-T method from the discovery GWAS summary statistics and selected one best predictor defined by the maximal AUC in the discovery set (Supplementary Table S6). Taking widow's peak as an example, the AUCs of the predictors ranged from 0.563 to 0.598 and reached the maximum when $p = 1 \times 10^{-5}$ and $r^2 = 0.6$ (Supplementary Table S6). Afterward, polygenic scores were generated in the validation set. Across the population, GPS was distributed with the empirical risk of the traits, showing a generally rising trend from 0.322 in the lowest quantile to 0.566 in the highest quantile (Figure 4A). Odds ratios (ORs) based on the quantile were given (Supplementary Table S7). Likewise, GPSs with the best performance were selected in the other four traits, and phenotype prevalence according to GPS was generated (Figures 4B–E). AUC reached maximum values of 0.594, 0.665, 0.657, and 0.625 in unibrow ($p = 1 \times 10^{-5}$, $r^2 = 0.4$), double eyelid ($p = 1 \times 10^{-4}$, $r^2 = 0.2$ or 0.8), earlobe attachment ($p = 1 \times 10^{-5}$, $r^2 = 0.4$), and freckles ($p = 1 \times 10^{-7}$, $r^2 = 0.4$), respectively (Supplementary Table S6).

3.4 SNP heritability of each trait

SNP heritability (h^2) using LDSC for the five traits was presented. The highest h^2 was seen for double eyelid (h^2 = 0.4487, standard error [SE] = 0.0765), and the h^2 for widow's peak, unibrow, earlobe attachment, and freckles were estimated to be 0.3046 (SE = 0.0591), 0.433 (SE = 0.0881), 0.2443 (SE = 0.0765), and 0.1431 (SE = 0.0733), respectively.

4 Discussion

It is important to understand the complicated genetic background of the facial traits since it may facilitate further understanding of the basic mechanism of facial development. It becomes even more significant upon the notion that, while some

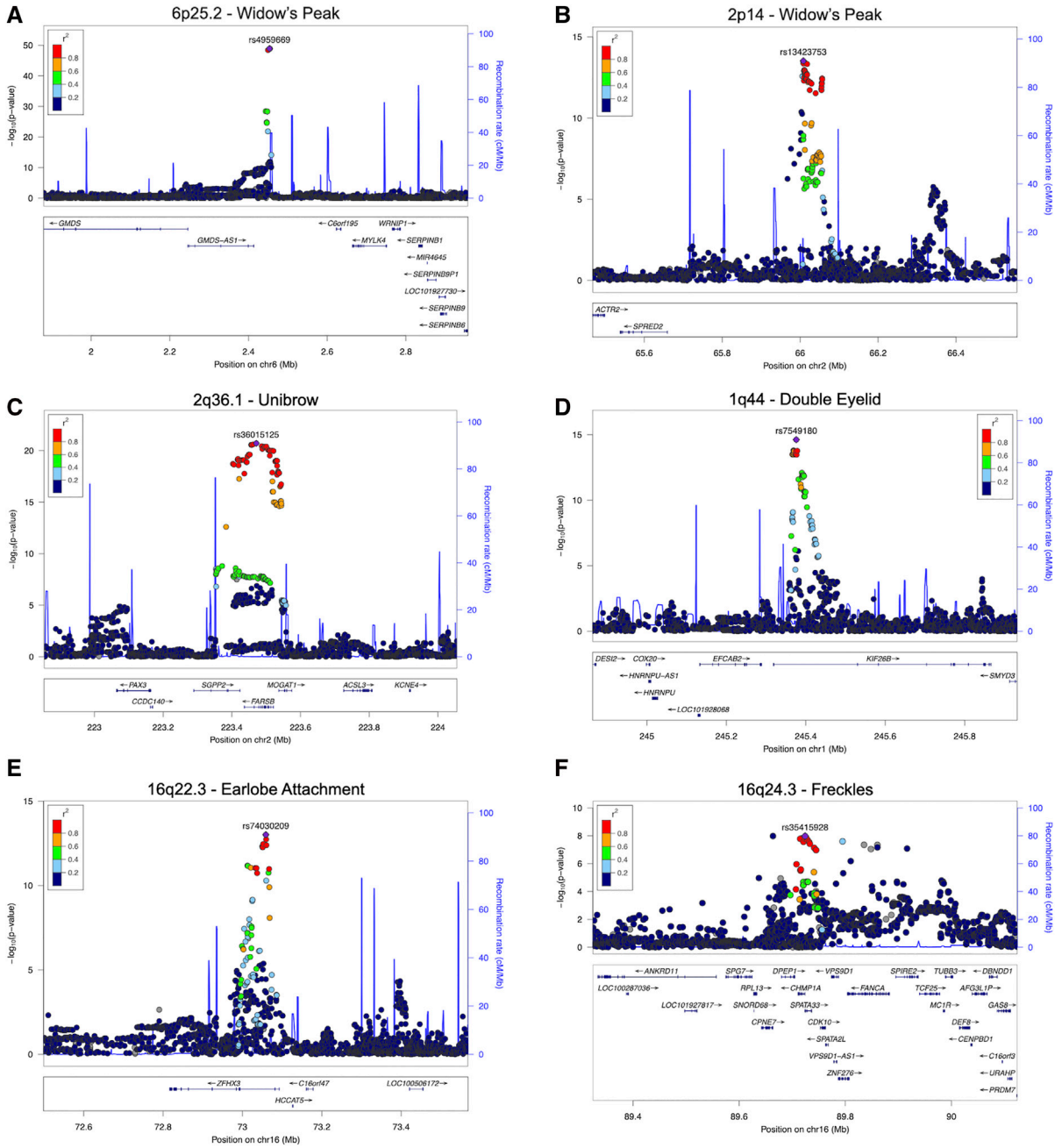


FIGURE 3
Regional association plots for eight regions with novel SNPs showing genome-wide significant associations with the five facial traits. Two novel associations for widow's peak and novel associations with the smallest *p*-values for unibow, double, eyelid, earlobe attachment, and freckles are shown **(A)** and **(B)**, Regional association plot for **(A)** 6p25.2 and **(B)** 2p14 with novel SNP showing genome-wide significant association with widow's peak; **(C)** regional association plot for 2q36.1 showing association with unibrow; **(D)** regional association plot for 1q44 showing association with double eyelid; **(E)** regional association plot for 16q22.3 showing association with earlobe attachment; **(F)** regional association plot for 16q24.3 showing association with freckles. SNP, single nucleotide polymorphism.

facial traits only represent nonsyndromic conditions, some can be clinical manifestations of certain syndromes.

To our knowledge, we are the first to provide public GWAS results on widow's peak, presenting two associated loci. The strongest signal was at 6p25.2 (rs4959669), near RNA genes *GMDS-AS1* and *LINC01600*, but further studies are needed to verify how the genes and their variants contribute to hairline morphology. The other association

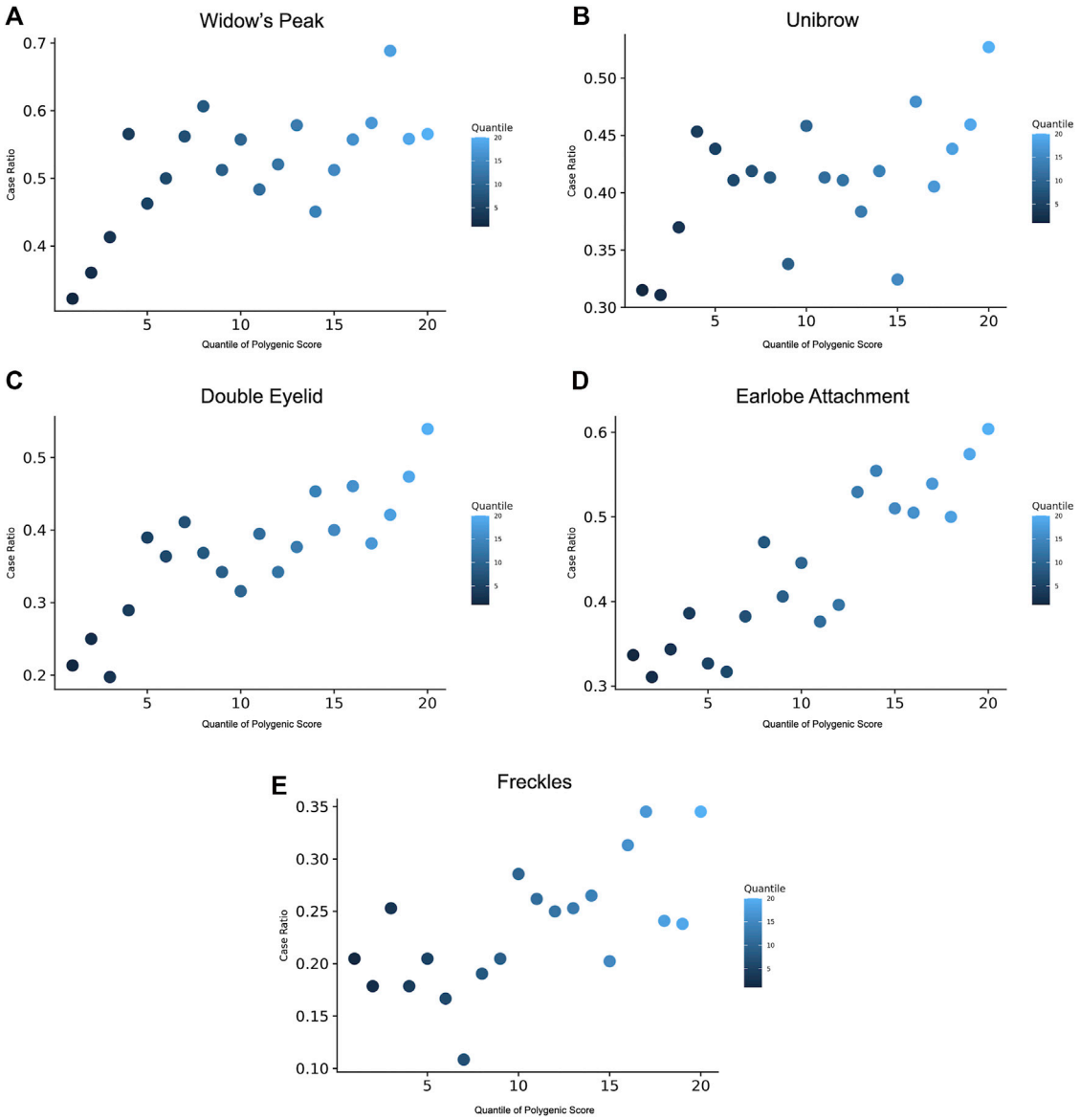


FIGURE 4
Prevalence of the traits according to the GPS quantile. 20 groups of the validation were derived based on the percentile of the GPS. Prevalence of phenotype displayed for the risk of **(A)** widow's peak, **(B)** unibrow, **(C)** double eyelid, **(D)** earlobe attachment, and **(E)** freckles, within each quantile. GPS, genome-wide polygenic score.

(rs13423753) occurred at 2p14 near *SPRED2*. Of potential relevance, some other members of the SPRED family have been reported to activate MAPK cascade (Nonami et al., 2004) which is implicated in hair follicle cell development (Yoon et al., 2011). Moreover, widow's peak sometimes manifests as a symptom of certain syndromes, such as Donnai-Barrow syndrome (Longoni et al., 1993; Khalifa et al., 2015), Waardenburg syndrome type 1 (WS1), and Aarskog syndrome (Pingault et al., 2010), but variants related to these syndromes did not reach genome-wide significance in our study, probably due to the rare syndromic incidence and the difference among populations. It is important to take into consideration rare genetic-based disorders and diseases when discussing variant and phenotype association. Noteworthily, 23andMe Co. attempted to identify significant variants for unibrow and widow's peak. According to the regional plot released on their website (https://medical.23andme.com/), associated loci for unibrow existed on several chromosomes, while significant variants for widow's peak were located at 2q and 6p. Since detailed information from

23andMe Co.'s research is restrained, associations in our present GWAS provide novel insights for the public.

Unibrow is also related to attractiveness in many cultures. As far as we know, the only published associations for unibrow were in 2q36 with the lead SNP of rs2218065 near the *PAX3* gene, which was not validated in the present study (Adhikari et al., 2016). *PAX3* is a key transcription factor that guides the normal development of neural crest derivatives (Sang et al., 2012), and its mutations have been shown to cause WS1, 85% of which has manifestations including unibrow (Pingault et al., 2010). Of potential relevance, the PAX3 locus has previously been shown to control the location of "nasion," the point at the middle of two eyebrows (Liu et al., 2012; Paternoster et al., 2012). In the present study, unlinked significant signals occurred near *PAX3* at 2q36.1, overlapping with the *FARSB* gene, a member of the *ARS* class IIc subfamily (Rodova et al., 1999). Other *ARS* members such as *KARS*, *CARS*, and *TARS* have been associated with hair phenotype (Santos-Cortez et al., 2013; Theil et al., 2019), suggesting a potential connection between *FARSB* and hair/brow development.

Interestingly, although East Asians are genetically closely-related, the present-day populations from different countries may have distinct genetic makeup (Wang et al., 2018), as seen in the pursuit of eyelid-associated variants. Among the five double eyelid-associated variants identified in our research, only rs10749244 at 10q26.11 is in high LD with previously reported rs1415425 (r^2 = 0.97) in Japanese (Chihiro et al., 2018). Among the four novel signals, the strongest association was observed for *KIF26B* rs7549180. In light of findings of *Kif26b* in the development of face (Marikawa et al., 2004), our results may suggest a regulatory role of *KIF26B* in the development of facial structure and concomitant upper eyelid differences. Functionally, *RPGRIP1L* rs6499632 serves as a strong GTEx eQTL for *RPGRIP1L* in cultured fibroblasts. The gene has been suggested to be involved in mechanisms such as craniofacial development, patterning of the limbs, and formation of the left-right axis (Delous et al., 2007). Another novel association (top SNP: rs147581439) overlapped with *PAX1*, a member of the *PAX* transcription factor family that plays a critical role during fetal development. Specifically, *PAX1* functions in pattern formation during embryogenesis (Wallin et al., 1994), and a missense mutation in *PAX1* has been shown to cause autosomal recessive Oto-Facio-Cervical syndrome, a disorder characterized by markedly skeletal and facial abnormalities (Pohl et al., 2013).

Regarding earlobe attachment, we identified two novel associations. One (rs74030209) is an intron point of *ZFHX3*, in LD with mutation points of the gene (Supplementary Table S4). *ZFHX3* is of potential relevance to ear development since it is involved in myogenic control by modulating myoblast differentiation (Berry et al., 2001), lack of which has been found to influence organogenesis in the inner ear phenotype (Rot et al., 2017). The other (rs10211400) is at the noncoding RNA (ncRNA) *LINC01107*. As some ncRNAs are correlated with nearby gene expression (Cabili et al., 2011; Guil and Esteller, 2012), the variant has a chance to be related to the regulation of *TWIST2* 314 kb downstream. Mutations of *TWIST2* have been associated with ectodermal dysplasia, such as Ablepharon-Macrostomia syndrome and Barber-Say syndrome (Marchegiani et al., 2015) whose manifestations include dysmorphic ears. Further studies are still needed to verify the conjecture. Besides these two unreported variants, our results of the attached earlobe mostly replicated the previous findings in diverse cohorts (Dutta and Ganguly, 1965; Adhikari et al., 2015; Shaffer et al., 2017). For instance, the strongest association was seen for the intergenic SNP rs12695694 near *MRPS22*, showing a strong GTEx eQTL association with *MRPS22* expression in cultured fibroblasts. Mutations of *MRPS22* have been previously implicated in earlobe size in Latin Americans and in lobe attachment in multiple cohorts (Adhikari et al., 2015; Shaffer et al., 2017), and relatively, a homozygous mutation in *MRPS22* has been reported to lead to oxidative phosphorylation system deficiency, which may manifest as dysmorphic features including low implanted posteriorly rotated ears (Smits et al., 2011). Meanwhile, the previously reported *EDAR* exonic variant rs3827760 (Bryk et al., 2008; Mou et al., 2008) also reached genome-wide significance in the present study. *EDAR* is involved in the prenatal development of ectoderm (Mikkola, 2009), and its deficiency has been suggested to result in abnormally shaped ears in mice (Adhikari et al., 2015).

Despite that pigmentary traits could be induced by extrinsic factors such as sun exposure, genetic predisposition has been suggested among different populations (Jacobs et al., 2015; Crawford et al., 2017; Chihiro et al., 2018; Adhikari et al., 2019; Shin et al., 2021). This may be because pigmentation is mainly contributed by a complicated process of melanin synthesis which is tightly associated with multiple genetic variants, and response after sun exposure is also genetically controlled (Nan et al., 2009; Shido et al., 2019). The freckles-associated variants identified in the present study were highly consistent with findings from Japanese and Korean cohorts (Chihiro et al., 2018; Shin et al., 2021), presumably due to the shared genetic backgrounds of East Asian populations. The only novel variant was *SPATA33* rs35415928. *SPATA33* has long been associated with facial pigmentation (Jacobs et al., 2015), cutaneous squamous cell carcinoma (Asgari et al., 2016), and melanoma (Fang et al., 2019). Closely downstream of the associations also lies the well-defined freckles-associated gene *MC1R*. (Maarten et al., 2001; Sulem et al., 2008; Eriksson et al., 2010). Among the identified associations, *BNC2* has been identified in Europeans (Jacobs et al., 2015). The top signal within *BNC2* (rs4455968) is in high LD with rs16935073 (r^2 = 0.94) and rs10816035 (r^2 = 0.83) that have been associated with pigmentary traits or tanning ability in Koreans (Shin et al., 2021) and Japanese (Chihiro et al., 2018; Shido et al., 2019). Another significant association existed in 10q25.3, led by *HSPA12A* rs12245621 which is in LD with the reported variant rs12259842 (r^2 = 0.76) (Chihiro et al., 2018). *HSPA12A* is affiliated to *HSP70* family whose members (*HSP70* and *HSP47*) are expressed in the dermis and epidermis following laser irradiation, which has been related to pigmentation (Sajjadi et al., 2013). Interestingly, the nearby *RAB11FIP2* has been proved to facilitate melanin exocytosis from melanocytes and filopodia-mediated melanin transfer (Beaumont et al., 2011;

Tarafder et al., 2014), and SNP rs35563099 192 kb upstream of *RAB11FIP2* also reached a genome-wide significance. To be noted, rs77733715 that has been associated with ease of tanning and darker skin color in UK BioBank samples (Sturm and Duffy, 2018) reached genome-wide significance in the discovery set but failed validation in our cohort. rs77733715 lies near the pigmentary gene *MFSD12* (Crawford et al., 2017; Adhikari et al., 2019; Hédan et al., 2019; Tanaka et al., 2019) and is in LD with the well-documented pigmentation-associated missense variant *MFSD12* rs2240751 (r^2 = 0.56) in the Korean and Latin American populations (Adhikari et al., 2019; Shin et al., 2021).

The study also has room for improvement. First, the individuals were recruited based on their self-reported traits instead of professional assessment. As the traits included are easily distinguished aesthetic traits and illustrations were added for each question, we consider the reports reliable. Second, functional annotation suggested that some of the variants may be associated with phenotype by impacting the expression level or coding sequence of the nearby genes, but the functions of these variants should be further determined and experimentally evaluated. Moreover, SNP heritability was presented for each trait, but since it can only include contributions from causal variants tagged by the measured SNPs, it is lower than total narrow-sense heritability, such as estimation from twin or family studies. For instance, an adult twin study on the relative contribution of genetic and environmental effects on the expression of nevi and freckles suggested that additive genetic effects explained 91% of the variance in freckle counts (Bataille et al., 2000). Future studies on narrow-sense heritability could be adopted to understand the genetic contribution of the five aesthetic facial traits.

5 Conclusion

This GWAS of five aesthetic facial traits in a large Chinese cohort of 26,806 uncovered ten novel genetic associations. Specifically, this is the first study, to our knowledge, to report genetic predispositions to widow's peak. The identified variants indicated both important similarities and differences among different ethnic groups. Hopefully, the findings would facilitate an understanding of the genetic basis of facial traits and, more importantly, facial development.

Data availability statement

The GWAS summary statistics can be found at the publicly available at http://www.biosino.org/node/project/detail/OEP002975.

Ethics statement

The studies involving human participants were reviewed and approved by West China Hospital, Sichuan University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

PW, XS, QM, HX, DB, and YZ contributed to the conceptualization; XS and QM contributed to the methodology; XS, HM, MC, and YZ performed the formal analysis; PW, QM, and MC performed the investigation; XS, HM, MC took charge in the resources; MC, SZ, and YW contributed to the data curation; PW prepared the original draft; PW, XS, and MC prepared the figures; DB and YZ supervised the study; All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

The study was supported by the National Key Research and Development Program of China (2021YFA1301203), grants from National Natural Science Foundation of China (No. 81973408, 81903735, 82002569, and 82071146), and 1.3.5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZYYC20003 and ZYJC18004).

Conflict of interest

Xinghan Sun and Hao Mi are employees of Chengdu 23Mofang Biotechnology. The remaining authors declare no competing interests.

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.967684/full#supplementary-material

References

Adhikari, K., Reales, G., Smith, A. J. P., Konka, E., Palmen, J., Quinto-Sanchez, M., et al. (2015). A genome-wide association study identifies multiple loci for variation in human ear morphology. *Nat. Commun.* 6, 7500. doi:10.1038/ncomms8500

Adhikari, K., Fontanil, T., Cal, S., Mendoza-Revilla, J., Fuentes-Guajardo, M., Chacón-Duque, J.-C., et al. (2016). A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat. Commun.* 7 (1), 10815. doi:10.1038/ncomms10815

Adhikari, K., Mendoza-Revilla, J., Sohail, A., Fuentes-Guajardo, M., Lampert, J., Chacón-Duque, J. C., et al. (2019). A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. *Nat. Commun.* 10 (1), 358. doi:10.1038/s41467-018-08147-0

Asgari, M. M., Wang, W., Ioannidis, N. M., Itnyre, J., Hoffmann, T., Jorgenson, E., et al. (2016). Identification of susceptibility loci for cutaneous squamous cell carcinoma. *J. Investigative Dermatol.* 136 (5), 930–937. doi:10.1016/j.jid.2016.01.013

Bataille, V., Snieder, H., MacGregor, A. J., Sasieni, P., and Spector, T. D. (2000). Genetics of risk factors for melanoma: an adult twin study of nevi and freckles. *J. Natl. Cancer Inst.* 92 (6), 457–463. doi:10.1093/jnci/92.6.457

Beaumont, K. A., Hamilton, N. A., Moores, M. T., Brown, D. L., Ohbayashi, N., Cairncross, O., et al. (2011). The recycling endosome protein Rab17 regulates melanocytic filopodia formation and melanosome trafficking. *Traffic* 12 (5), 627–643. doi:10.1111/j.1600-0854.2011.01172.x

Berry, F. B., Miura, Y., Mihara, K., Kaspar, P., Sakata, N., Hashimoto-Tamaoki, T., et al. (2001). Positive and negative regulation of myogenic differentiation of C2C12 cells by isoforms of the multiple homeodomain zinc finger transcription factor ATBF1. *J. Biol. Chem.* 276 (27), 25057–25065. doi:10.1074/jbc.M010378200

Bryk, J., Hardouin, E., Pugach, I., Hughes, D., Strotmann, R., Stoneking, M., et al. (2008). Positive selection in East Asians for an EDAR allele that enhances NF-κB activation. *PloS one* 3 (5), e2209. doi:10.1371/journal.pone.0002209

Bulik-Sullivan, B. K., Loh, P. R., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47 (3), 291–295. doi:10.1038/ng.3211

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25 (18), 1915–1927. doi:10.1101/gad.17446611

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci* 4, 7. doi:10.1186/s13742-015-0047-8

Chihiro, E., Johnson, T. A., Ryoko, M., Kazuyuki, N., Shigeo, K., Masanori, A., et al. (2018). Genome-wide association study in Japanese females identifies fifteen novel skin-related trait associations. *Sci. Rep.* 8 (1), 8974. doi:10.1038/s41598-018-27145-2

Cole, J. B., Manyama, M., Larson, J. R., Liberton, D. K., Ferrara, T. M., Riccardi, S. L., et al. (2017). Human facial shape and size heritability and genetic correlations. *Genetics* 205 (2), 967–978. doi:10.1534/genetics.116.193185

Crawford, N. G., Kelly, D. E., Hansen, M. E. B., Beltrame, M. H., Fan, S., Bowman, S. L., et al. (2017). Loci associated with skin pigmentation identified in African populations. *Science* 358 (6365), eaan8433. doi:10.1126/science.aan8433

Claes, P. (2014). Modeling 3D facial shape from DNA. *PLoS Genet.* 10 (3), e1004224. doi:10.1371/journal.pgen.1004224

Delous, M., Baala, L., Salomon, R., Laclef, C., Vierkotten, J., Tory, K., et al. (2007). The ciliary gene RPGRIP1L is mutated in cerebello-oculo-renal syndrome (Joubert syndrome type B) and Meckel syndrome. *Nat. Genet.* 39 (7), 875–881. doi:10.1038/ng2039

Dutta, P., and Ganguly, P. (1965). Further observations on ear lobe attachment. *Hum. Hered.* 15, 77–86. doi:10.1159/000151894

Eriksson, N., Macpherson, J. M., Tung, J. Y., Hon, L. S., Naughton, B., Saxonov, S., et al. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* 6 (6), e1000993. doi:10.1371/journal.pgen.1000993

Fang, S., Lu, J., Zhou, X., Wang, Y., Ross, M. I., Gershenwald, J. E., et al. (2019). Functional annotation of melanoma risk loci identifies novel susceptibility genes. *Carcinogenesis* 41 (4), 452–457. doi:10.1093/carcin/bgz173

GTEx Consortium (2015). Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348 (6235), 648–660. doi:10.1126/science.1262110

Guil, S., and Esteller, M. (2012). Cis-acting noncoding RNAs: friends and foes. *Nat. Struct. Mol. Biol.* 19 (11), 1068–1075. doi:10.1038/nsmb.2428

Hao, Q., Cao, M., Zhang, C., Yin, D., Wang, Y., Ye, Y., et al. (2021). Age-related differences of genetic susceptibility to patients with acute lymphoblastic leukemia. *Aging* 13 (9), 12456–12465. doi:10.18632/aging.202903

Hédan, B., Cadieu, E., Botherel, N., Dufaure de Citres, C., Letko, A., Rimbault, M., et al. (2019). Identification of a missense variant in MFSD12 involved in dilution of phaeomelanin leading to white or cream coat color in dogs. *Genes* 10 (5), 386. doi:10.3390/genes10050386

Hertz, D. L., Douglas, J. A., Kidwell, K. M., Gersch, C. L., Desta, Z., Storniolo, A.-M., et al. (2021). Genome-wide association study of letrozole plasma concentrations identifies non-exonic variants that may affect CYP2A6 metabolic activity. *Pharmacogenet Genomics* 31 (5), 116–123. doi:10.1097/fpc.0000000000000429

Huang, Y., Li, D., Qiao, L., Liu, Y., Peng, Q., Wu, S., et al. (2021). A genome-wide association study of facial morphology identifies novel genetic loci in Han Chinese. *J. Genet. Genomics* 48 (3), 198–207. doi:10.1016/j.jgg.2020.10.004

Jacobs, L. C., Hamer, M. A., Gunn, D. A., Deelen, J., Lall, J. S., van Heemst, D., et al. (2015). A genome-wide association study identifies the skin color genes IRF4, MC1R, ASIP, and BNC2 influencing facial pigmented spots. *J. Investigative Dermatol.* 135 (7), 1735–1742. doi:10.1038/jid.2015.62

Kayser, M., and De Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nat. Rev. Genet.* 12 (3), 179–192. doi:10.1038/nrg2952

Khalifa, O., Al-Sahlawi, Z., Imtiaz, F., Ramzan, K., Allam, R., Al-Mostafa, A., et al. (2015). Variable expression pattern in Donnai-Barrow syndrome: report of two novel LRP2 mutations and review of the literature. *Eur. J. Med. Genet.* 58 (5), 293–299. doi:10.1016/j.ejmg.2014.12.008

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50 (9), 1219–1224. doi:10.1038/s41588-018-0183-z

Kim, Y., Yin, J., Huang, H., Jorgenson, E., Choquet, H., and Asgari, M. M. (2022). Genome-wide association study of actinic keratosis identifies new susceptibility loci implicated in pigmentation and immune regulation pathways. *Commun. Biol.* 5 (1), 386. doi:10.1038/s42003-022-03301-3

Kyriakou, G., Glentis, A., and Papanikolaou, S. (2021). Widow's peak: a usually overlooked, yet significant morphogenetic trait. *J. Dtsch. Dermatol. Ges.* 19 (9), 1271–1275. doi:10.1111/ddg.14502

Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., et al. (2012). A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genet.* 8 (9), e1002932. doi:10.1371/journal.pgen.1002932

Longoni, M., Kantarci, S., Donnai, D., and Pober, B. R. (1993). "Donnai-barrow syndrome," in *GeneReviews(®)*. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. H. Bean, K. Stephens, et al. (Seattle, WA: University of Washington). Copyright © 1993-2020, GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved.

Maarten, B., ter Huurne, J., Nelleke, G., Wilma, B., Rudi, W., Vermeer, B.-J., et al. (2001). The melanocortin-1-receptor gene is the major freckle gene. *Hum. Mol. Genet.* 10 (16), 1701. doi:10.1093/hmg/10.16.1701

Marchegiani, S., Davis, T., Tessadori, F., van Haaften, G., Brancati, F., Hoischen, A., et al. (2015). Recurrent mutations in the basic domain of TWIST2 cause Ablepharon macrostomia and barber-say syndromes. *Am. J. Hum. Genet.* 97 (1), 99–110. doi:10.1016/j.ajhg.2015.05.017

Marikawa, Y., Fujita, T. C., and Alarcón, V. B. (2004). An enhancer-trap LacZ transgene reveals a distinct expression pattern of Kinesin family 26B in mouse embryos. *Dev. Genes Evol.* 214 (2), 64–71. doi:10.1007/s00427-003-0377-x

Mikkola, M. L. (2009). Molecular aspects of hypohidrotic ectodermal dysplasia. *Am. J. Med. Genet.* 149a (9), 2031–2036. doi:10.1002/ajmg.a.32855

Moriyama, T., Nishii, R., Perez-Andreu, V., Yang, W., Klussmann, F. A., Zhao, X., et al. (2016). NUDT15 polymorphisms alter thiopurine metabolism and hematopoietic toxicity. *Nat. Genet.* 48 (4), 367–373. doi:10.1038/ng.3508

Moriyama, T., Yang, W., Smith, C., Pui, C.-H., Evans, W. E., Relling, M. V., et al. (2021). Comprehensive characterization of pharmacogenetic variants in TPMT and NUDT15 in children with acute lymphoblastic leukemia. *Pharmacogenet Genomics* 32, 60–66. doi:10.1097/fpc.0000000000000453

Mou, C., Thomason, H. A., Willan, P. M., Clowes, C., Harris, W. E., Drew, C. F., et al. (2008). Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form. *Hum. Mutat.* 29 (12), 1405–1411. doi:10.1002/humu.20795

Nan, H., Kraft, P., Qureshi, A. A., Guo, Q., Chen, C., Hankinson, S. E., et al. (2009). Genome-wide association study of tanning phenotype in a population of European ancestry. *J. Investigative Dermatol.* 129 (9), 2250–2257. doi:10.1038/jid.2009.62

Nonami, A., Kato, R., Taniguchi, K., Yoshiga, D., Taketomi, T., Fukuyama, S., et al. (2004). Spred-1 negatively regulates interleukin-3-mediated ERK/mitogen-

activated protein (MAP) kinase activation in hematopoietic cells. *J. Biol. Chem.* 279 (50), 52543–52551. doi:10.1074/jbc.M405189200

Paternoster, L., Zhurov, A. I., Toma, A. M., Kemp, J. P., Pourcain, B. S., Timpson, N. J., et al. (2012). Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *Am. J. Hum. Genet.* 90, 478. doi:10.1016/j.ajhg.2011.12.021

Pingault, V., Ente, D., Dastot-Le Moal, F., Goossens, M., Marlin, S., and Bondurand, N. (2010). Review and update of mutations causing Waardenburg syndrome. *Hum. Mutat.* 31 (4), 391–406. doi:10.1002/humu.21211

Pohl, E., Aykut, A., Beleggia, F., Karaca, E., Durmaz, B., Keupp, K., et al. (2013). A hypofunctional PAX1 mutation causes autosomal recessively inherited otofaciocervical syndrome. *Hum. Genet.* 132 (11), 1311–1320. doi:10.1007/s00439-013-1337-9

Qian, M., Xu, H., Perez-Andreu, V., Roberts, K. G., Zhang, H., Yang, W., et al. (2019). Novel susceptibility variants at the ERG locus for childhood acute lymphoblastic leukemia in Hispanics. *Blood* 133 (7), 724–729. doi:10.1182/blood-2018-07-862946

Rassman, W. R., Pak, J. P., and Kim, J. (2013). Phenotype of normal hairline maturation. *Facial Plastic Surg. Clin. N. Am.* 21 (3), 317–324. doi:10.1016/j.fsc.2013.04.001

Rodova, M., Ankilova, V., and Safro, M. G. (1999). Human phenylalanyl-tRNA synthetase: cloning, characterization of the deduced amino acid sequences in terms of the structural domains and coordinately regulated expression of the α and β subunits in chronic myeloid leukemia cells. *Biochem. Biophys. Res. Commun.* 255 (3), 765–773. doi:10.1006/bbrc.1999.0141

Rot, I., Baguma-Nibasheka, M., Costain, W. J., Hong, P., Tafra, R., Mardesic-Brakus, S., et al. (2017). Role of skeletal muscle in ear development. *Histol. Histopathol.* 32 (10), 987–1000. doi:10.14670/hh-11-886

Sajjadi, A. Y., Mitra, K., and Grace, M. (2013). Expression of heat shock proteins 70 and 47 in tissues following short-pulse laser irradiation: assessment of thermal damage and healing. *Med. Eng. Phys.* 35 (10), 1406–1414. doi:10.1016/j.medengphy.2013.03.011

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2012). A better coefficient of determination for genetic profile Analysis. *Genet. Epidemiol.* 36 (3), 214–224. doi:10.1002/gepi.21614

Santos-Cortez, R. L. P., Lee, K., Azeem, Z., Antonellis, P. J., Pollock, L. M., Khan, S., et al. (2013). Mutations in KARS, encoding lysyl-tRNA synthetase, cause autosomal-recessive nonsyndromic hearing impairment DFNB89. *Am. J. Hum. Genet.* 93 (1), 132–140. doi:10.1016/j.ajhg.2013.05.018

Seongwon, C., Eun, L. J., Yeon, P. A., Jun-Hyeong, D., Woo, L. S., Chol, S., et al. (2018). Identification of five novel genetic loci related to facial morphology by genome-wide association studies. *Bmc Genomics* 19 (1), 481. doi:10.1186/s12864-018-4865-9

Shaffer, L. J., Lee, M. K., Roosenboom, J., Orlova, E., Adhikari, K., Gallo, C., et al. (2017). Multiethnic GWAS reveals polygenic architecture of earlobe attachment. *Am. J. Hum. Genet.* 101, 913. doi:10.1016/j.ajhg.2017.10.001

Shido, K., Kojima, K., Yamasaki, K., Hozawa, A., Tamiya, G., Ogishima, S., et al. (2019). Susceptibility loci for tanning ability in the Japanese population identified by a genome-wide association study from the tohoku medical megabank project cohort study. *J. Investigative Dermatol.* 139 (7), 1605–1608. doi:10.1016/j.jid.2019.01.015

Shin, J.-G., Leem, S., Kim, B., Kim, Y., Lee, S.-G., Song, H. J., et al. (2021). GWAS analysis of 17,019 Korean women identifies the variants associated with facial pigmented spots. *J. Investigative Dermatol.* 141 (3), 555–562. doi:10.1016/j.jid.2020.08.007

Smits, P., Saada, A., Wortmann, S. B., Heister, A. J., Brink, M., Pfundt, R., et al. (2011). Mutation in mitochondrial ribosomal protein MRPS22 leads to Cornelia de Lange-like phenotype, brain abnormalities and hypertrophic cardiomyopathy. *Eur. J. Hum. Genet.* 19 (4), 394–399. doi:10.1038/ejhg.2010.214

Sturm, R. A., and Duffy, D. L. (2018). Toward the full spectrum of genes for human skin colour. *Pigment. Cell Melanoma Res.* 31, 457. doi:10.1111/pcmr.12691

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Jakobsdottir, M., et al. (2008). Two newly identified genetic determinants of pigmentation in Europeans. *Nat. Genet.* 40 (7), 835–837. doi:10.1038/ng.160

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20 (8), 467–484. doi:10.1038/s41576-019-0127-1

Tanaka, J., Leeb, T., Rushton, J., Famula, T. R., Mack, M., Jagannathan, V., et al. (2019). Frameshift variant in MFSD12 explains the mushroom coat color dilution in shetland ponies. *Genes* 10 (10), 826. doi:10.3390/genes10100826

Tarafder, A. K., Bolasco, G., Correia, M. S., Pereira, F. J. C., Iannone, L., Hume, A. N., et al. (2014). Rab11b mediates melanin transfer between donor melanocytes and acceptor keratinocytes via coupled exo/endocytosis. *J. Investigative Dermatology* 134 (4), 1056–1066. doi:10.1038/jid.2013.432

Theil, A. F., Botta, E., Raams, A., Smith, D. E. C., Mendes, M. I., Caligiuri, G., et al. (2019). Bi-Allelic TARS mutations are associated with brittle hair phenotype. *Am. J. Hum. Genet.* 105 (2), 434–440. doi:10.1016/j.ajhg.2019.06.017

Wallin, J., Wilting, J., Koseki, H., Fritsch, R., Christ, B., and Balling, R. (1994). The role of Pax-1 in axial skeleton development. *Development* 120 (5), 1109–1121. doi:10.1242/dev.120.5.1109

Wang, Y., Lu, D., Chung, Y.-J., and Xu, S. (2018). Genetic structure, divergence and admixture of Han Chinese, Japanese and Korean populations. *Hereditas* 155 (1), 19. doi:10.1186/s41065-018-0057-5

Ward, L. D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40 (Database issue), D930–D934. doi:10.1093/nar/gkr917

Weinberg, S. M., Parsons, T. E., Marazita, M. L., and Maher, B. S. (2013). Heritability of face shape in twins: A preliminary study using 3D stereophotogrammetry and geometric morphometrics. *Dent. 3000* 1, 14. doi:10.5195/d3000.2013.14

Weinberg, S. M., Roosenboom, J., Shaffer, J. R., Shriver, M. D., Wysocka, J., and Claes, P. (2019). Hunting for genes that shape human faces: Initial successes and challenges for the future. *Orthod. Craniofac Res.* 22 Suppl 1(Suppl 1), 207–212. doi:10.1111/ocr.12268

Xu, H., Robinson, G. W., Huang, J., Lim, J. Y.-S., Zhang, H., Bass, J. K., et al. (2015a). Common variants in ACYP2 influence susceptibility to cisplatin-induced hearing loss. *Nat. Genet.* 47 (3), 263–266. doi:10.1038/ng.3217

Xu, H., Zhang, H., Yang, W., Yadav, R., Morrison, A. C., Qian, M., et al. (2015b). Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukaemia in children. *Nat. Commun.* 6, 7553. doi:10.1038/ncomms8553

Yoon, S.-Y., Kim, K.-T., Jo, S. J., Cho, A.-R., Jeon, S.-I., Choi, H.-D., et al. (2011). Induction of hair growth by insulin-like growth factor-1 in 1,763 MHz radiofrequency-irradiated hair follicle cells. *PLOS ONE* 6 (12), e28474. doi:10.1371/journal.pone.0028474

Zhang, S. Y., Zhou, X. Y., Zhou, X. L., Zhang, Y., Deng, Y., Liao, F., et al. (2019). Subtype specific inherited predisposition to pemphigus in the Chinese population. *Br. J. Dermatol.* 180 (4), 828–835. doi:10.1111/bjd.17191

Zhou, X., Li, D., Zhang, B., Lowdon, R. F., Rockweiler, N. B., Sears, R. L., et al. (2015). Epigenomic annotation of genetic variants using the roadmap epigenome browser. *Nat. Biotechnol.* 33 (4), 345–346. doi:10.1038/nbt.3158

Zhu, Y., Yin, D., Su, Y., Xia, X., Moriyama, T., Nishii, R., et al. (2018). Combination of common and novel rare NUDT15 variants improves predictive sensitivity of thiopurine-induced leukopenia in children with acute lymphoblastic leukemia. *Haematologica* 103 (7), e293–e295. doi:10.3324/haematol.2018.187658

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 03 October 2022
DOI 10.3389/fgene.2022.971242

Check for updates

OPEN ACCESS

EDITED BY
Guanglin He,
Sichuan University, China

REVIEWED BY
Dennis McNevin,
University of Technology Sydney, Australia
Kelly M. Burkett,
University of Ottawa, Canada

*CORRESPONDENCE
Jianye Ge,
Jianye.Ge@unthsc.edu

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 16 June 2022
ACCEPTED 16 September 2022
PUBLISHED 03 October 2022

CITATION
Huang M, Liu M, Li H, King J, Smuts A, Budowle B and Ge J (2022), A machine learning approach for missing persons cases with high genotyping errors. *Front. Genet.* 13:971242. doi: 10.3389/fgene.2022.971242

COPYRIGHT
© 2022 Huang, Liu, Li, King, Smuts, Budowle and Ge. 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.

A machine learning approach for missing persons cases with high genotyping errors

Meng Huang[1], Muyi Liu[1], Hongmin Li[2], Jonathan King[1], Amy Smuts[1], Bruce Budowle[1,3] and Jianye Ge[1,3]*

[1]Center for Human Identification, University of North Texas Health Science Center, Fort Worth, TX, United States, [2]Department of Computer Science, College of Science, California State University, East Bay, Hayward, CA, United States, [3]Department of Microbiology, Immunology and Genetics, University of North Texas Health Science Center, Fort Worth, TX, United States

Estimating the relationships between individuals is one of the fundamental challenges in many fields. In particular, relationship.ip estimation could provide valuable information for missing persons cases. The recently developed investigative genetic genealogy approach uses high-density single nucleotide polymorphisms (SNPs) to determine close and more distant relationships, in which hundreds of thousands to tens of millions of SNPs are generated either by microarray genotyping or whole-genome sequencing. The current studies usually assume the SNP profiles were generated with minimum errors. However, in the missing person cases, the DNA samples can be highly degraded, and the SNP profiles generated from these samples usually contain lots of errors. In this study, a machine learning approach was developed for estimating the relationships with high error SNP profiles. In this approach, a hierarchical classification strategy was employed first to classify the relationships by degree and then the relationship types within each degree separately. As for each classification, feature selection was implemented to gain better performance. Both simulated and real data sets with various genotyping error rates were utilized in evaluating this approach, and the accuracies of this approach were higher than individual measures; namely, this approach was more accurate and robust than the individual measures for SNP profiles with genotyping errors. In addition, the highest accuracy could be obtained by providing the same genotyping error rates in train and test sets, and thus estimating genotyping errors of the SNP profiles is critical to obtaining high accuracy of relationship estimation.

KEYWORDS

genetic genealogy, machine learning, genotyping error, feature selection, missing person, hierarchical classification, single nucleotide polymorphisms, kinship estimation

Introduction

DNA-based relatedness estimation is essential for identifying missing persons and human remains. The current standard genotyping technology used in missing person cases (i.e., capillary electrophoresis) measures the lengths of a set of pre-selected short tandem repeat (STR) markers. The major forensic commercial STR kits (e.g., GlobalFiler™ PCR Amplification Kit) usually contain 20 to 25 STR markers, including the core markers defined by the FBI's National DNA Index System (CODIS) (Hares, 2015). Close relationships (e.g., parent-child and full-sibling) can be determined with high accuracy with this limited number of markers, but not for more distant relationships (i.e., 2nd or higher degrees) (Ge et al., 2011; Ge et al., 2011; Ge and Budowle, 2020). The recently developed investigative genetic genealogy (IGG) approach uses high-density single nucleotide polymorphisms (SNPs) to determine more distant relationships. In this approach, hundreds of thousands to tens of millions of SNPs are generated either by microarray genotyping or whole-genome sequencing (WGS). With the massive number of variants, the distant relationships can be determined with much higher accuracies (Li et al., 2014). Hundreds of missing persons and cold cases have been solved with IGG (Greytak et al., 2019; Tillmar et al., 2021).

The methods to determine relationships with SNPs can be generally classified into three main categories: Hidden Markov Model (HMM) based likelihood ratio methods (LRs) (Boehnke and Cox, 1997; Epstein et al., 2000; Heinrich et al., 2017; Kling, 2019; Galván-Femenía et al., 2021), genome-wide relatedness methods (namely, statistics based on individual SNPs) (Purcell et al., 2007; Manichaikul et al., 2010), and identity-by-descent (IBD) segment detection methods (Gusev et al., 2009; Browning and Browning, 2011; Qiao et al., 2021). The LR methods calculate the likelihoods of the given hypotheses, and the relationship is determined by the maximum likelihood. The LR methods require all loci to be in linkage equilibrium (namely, only a few thousand SNPs may be used) and allele frequencies of each locus are known. The genome-wide relatedness methods summarize the statistic measures from individual markers, and the calculations of these measures are very fast. With a sufficient number of markers, the accuracies are high enough to estimate close relationships (i.e., up to 3rd degree relationships). The IBD segment detection methods use the positions and/or linkage disequilibrium (LD) between markers that the genome-wide relatedness methods ignore and detect the identical haplotype segments shared between profiles, which provide the highest accuracies in estimating relationships, particularly distant relationships. All the methods presume that the SNP calling is accurate with negligible errors (Conomos et al., 2015; Korneliussen and Moltke, 2015; Nøhr et al., 2021; Pew et al., 2015; Shcherbina et al., 2016a; Sherry et al., 2017; Staples et al., 2014; Stevens et al., 2011; Waples et al., 2019). In addition, many of these methods also require allele frequencies or even population admixture ratios in their calculations (Alexander et al., 2009; Thornton et al., 2012; Morrison, 2013; Moltke and Albrechtsen, 2014; Conomos et al., 2015; Conomos et al., 2016).

TABLE 1 The list of 10 relationship types in Supplementary Figure S1.

Relationship type	Relationship degree
Unrelated	N/A
Parent-child	1
Full-sibling	1
Grandparent	2
Half-sibling	2
Uncle-nephew	2
First-cousin	3
Grand-uncle	3
Half-uncle	3
Great-grandparent	3

In missing persons cases, the samples (e.g., bones) can be highly degraded. Thus, the genotyping error rates (GERs) of these samples could be high (e.g., the GER could be 5–10% or higher depending on the quality filtering of the data), and precise allele frequency data are not available. The genome-wide relatedness methods may be more robust to genotyping errors, as the errors at individual markers may not impact the measures of the other markers. However, genotyping errors at one or a few loci can easily interrupt the IBD segments. Thus, the IBD segment detection methods are more sensitive to errors, and their performance may substantially decay as genotyping errors increase. A recent study (Turner et al., 2022) evaluated the impact of GERs on genome-wide relatedness methods and IBD segment methods. The results showed that the overall relationship classification accuracies of different methods were similar if GER is of a low level (GER = 0); however, the accuracies of the IBD segment methods drop quickly when GER is higher than 1% (e.g., the accuracy of hap-IBD (Zhou et al., 2020) approaches to random guessing when GER ≥1%), which means the IBD segments methods are very sensitive to high GERs and require high-quality genotype data. The genome-wide relatedness method (KING) (Manichaikul et al., 2010) had slightly lower accuracies than those of the IBD segment method (IBIS) (Seidman et al., 2020) when GERs were at low-level (i.e., 0 and 0.01). The accuracies of both genome-wide relatedness method and the IBD segment methods, such as IBIS and hap-IBD (Zhou et al., 2020), decreased with higher GERs (i.e., 0.05 and 0.1). However, the accuracies of KING were less impacted by GERs. Thus, more robust methods are needed for missing person samples with high genotyping error. In this study, a supervised machine learning approach was developed for classifying different degrees of relationships and relationship types within the same degrees based on SNP profiles with high genotyping errors. This approach combined 17 genome-wide relatedness measures to train classifiers aiming to reduce the effect of genotyping error and improve the accuracy of relationship estimation.

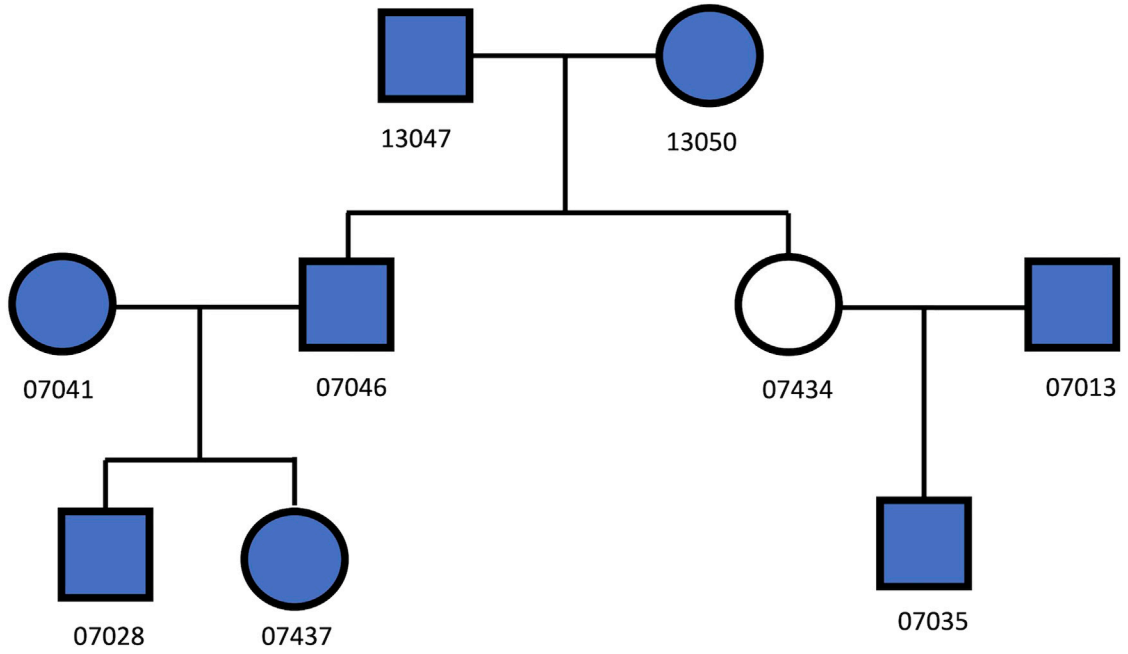


FIGURE 1
The relationships among the 8 UTAH/CEPH cell line samples (filled in with blue color).

Methods

Simulation data

Family-based genotype data were simulated to train and test the machine learning classifiers. A large pedigree was designed for simulation, which includes ten various relationships from 1st to 3rd degrees and unrelated individuals (Supplementary Figure S1 and Table 1). This designed pedigree was simulated using Ped-sim (Caballero et al., 2019) with the default setting, except that the GERs were specified (i.e., 0, 0.01, 0.03, 0.05, 0.07, and 0.1). GERs larger than 0.1 were not included in the simulation, as the GERs usually could be reduced to below 0.1 with proper quality control and data cleaning, although fewer SNPs would survive (Wall et al., 2014). We first randomly sampled founders and simulated offspring's genotype data using these founders' genotype data. In total, 10,000 families were simulated using the pedigree defined in Supplementary Figure S1 (supplementary material). Next, in each simulated family, one pair of each relationship type were sampled (Table 1), and thus the numbers of each relationship in the final dataset were balanced (i.e., each relationship type has the same number of pairs in the training set). The final simulated dataset included 100,000 pairs of individuals with 4 different degrees and 10 different relationship types (including unrelated relationships). Each relationship type has 10,000 instances in the dataset. The simulation adopted 503 unrelated European ancestry (EUR) samples from the 1,000 genomes project sequencing data (30X coverage) (Auton et al., 2015) as founders. For each founder, 582K autosomal biallelic SNPs from Illumina GSA (Global Screening Array) panel were extracted from the 1,000 genomes project and used in the simulation. The GSA panel was selected because most profiles in the genealogy databases (e.g., GEDmatch; https://www.gedmatch.com) were generated by microarray, and GSA is one of the most widely used panels.

Real data

Eight Utah European descendant samples (Dausset et al., 1990) (Figure 1) were selected. Among these samples, there were 28 pairs of relationships, including 11 unrelated, 7 1st degree, 8 2nd degree, and 2 3rd degree relationships. These samples were genotyped using Illumina Infinium Omni5-4 Kit, containing 4.3 million autosomal biallelic SNPs. Each sample was genotyped three times with various input DNA (i.e., 50 ng, 500 pg, and 100 pg). In total, 4,198,873 SNPs were called by Genome Studio, in which 418,513 SNPs were overlapped with the SNPs in the GSA panel. These ~419 K autosomal biallelic SNPs were extracted and used to test the performance of the trained classifiers.

For the classification with these real data, the simulated datasets with different genotyping errors were used as the training datasets to test the effect of the consistency of the GERs between the reference datasets (assumed to be minimum errors) and the test datasets. For each pair of individuals in the test

dataset of classification, a sample containing 50 ng DNA was considered as the reference sample, and another sample containing 500 pg or 100 pg was used as a case sample.

Feature extraction

For each pair of profiles, either simulated or real data, 17 measures were extracted as features to describe the relationships between individuals (Supplementary Table S1). These measures included KING-homo (K0) (Manichaikul et al., 2010), KING-robust (K1) (Manichaikul et al., 2010), IBS = 0, IBS = 1, IBS = 2, the union of IBS = 0 and IBS = 1 (Stevens et al., 2011), the union of IBS = 1 and IBS = 2, the union of IBS = 0 and IBS = 2, and nine allele combinations of a pair of individuals (j1—j9) (Waples et al., 2019). These features were solely based on the genotypes. Measures with allele frequencies and IBD segments were not included, considering that the allele frequencies of certain populations may not be accurate or even available. In addition, the IBD segment estimation is inaccurate with high error sequence or genotype data generated from degraded samples (e.g., DNA extracted from human bones).

Classification algorithms selection and hierarchical classification strategy

First, two strong classification algorithms, Random Forest (RF) and Support Vector Machine (SVM), which have different underline learning mechanisms, were compared using 10-fold cross-validation with simulation data to select an algorithm for high classification accuracy and high robustness with noisy data. The higher-performing algorithm would be used for all feature selections and classification. The classification accuracy was determined by 10-fold cross-validations.

The accuracy of determining the relationship degree is usually much higher than those of determining the relationships within the 2nd or 3rd degrees. The best features to differentiate relationship degrees and relationship types within various degrees may also be different. Thus, a hierarchical classification strategy was implemented to first determine the relationship degree (one classifier) and then determine the relationship type within the same degrees (three classifiers for 1st, 2nd, and 3rd, respectively; Figure 2). In total, there are four classifiers.

For each classifier, forward feature selection was implemented to seek the best performing sets of features for classifications, in which the top-performing (i.e., the highest classification accuracy with 10-fold cross-validation) features among all available features (that have not been selected) were iteratively added to the best performing set using a greedy algorithm. The selected features for a particular classifier may vary with different genotyping errors, as the features may have different degrees of robustness to genotyping errors. The most commonly selected features across all genotyping errors (i.e., the features with the highest robustness and/or the highest classification accuracies) were decided as the final set of selected features. In both real data (i.e., dilution series) and simulation data classifications, the train datasets were the simulated datasets with GER = 0, and the test datasets had various error rates.

Results

Supervised classification algorithms comparison

Based on the results of feature selection and classification (Figure 3), in general, the classification accuracies with RF were higher than those with SVM, except for the classifications of relationship types within 2nd and 3rd degrees with very high GER (i.e., 0.1). For the classification of degrees, both RF and SVM could achieve close to 100% accuracy with the top-performing features, but RF was much more robust. For the classification within the 1st degree relationship, 100% or close to 100% accuracies were obtained with top-performing features across all GERs using SVM (Figure 3). If more than 10 features were selected, the accuracy of RF drops quickly when GER was higher than 0.03, implying that SVM was more robust than RF within the 1st degree. For 2nd and 3rd degree relationships, when the GERs were low (e.g., 0.01), RF could achieve higher accuracies for all three classifications. For example, for relationship type within the 2nd degree, ~77% accuracy with RF was obtained by the best 8 features, while ~50% accuracy with SVM was obtained by the best 7 features. In the subsequent analysis, RF with 10-fold cross-validation was employed to conduct forward feature selections and classifications.

Feature selection

Different top-performing features might be selected with data simulated using different genotyping errors. For the classification of relationship degrees (Figure 4A), the top 7 features of each given GER would obtain 100% accuracy across all genotyping errors (Figure 4A), and adding additional features would lead to lower accuracies. In total, 42 (= 6 × 7) features were selected across 6 different errors. Figure 4B summarizes the counts of these 42 selected features (14 unique features). K1, j4 and j7 were selected across five genotyping errors, some features were commonly selected (e.g., K0, IBS0, j6, and j8 etc.), and some features were never selected (e.g., j2, j5, and IBS12).

Since the actual genotyping errors of the forensic samples are unknown or hard to estimate, it might not be appropriate in practice to select different features for different genotyping errors. Thus, feature selection was further conducted by the order of the counts in Figure 4B (i.e., similar fashion as the forward feature selection, but by the ranking of the order of counts) to decide the top-performing set of features that are accurate and improve the robustness across all GERs (Figure 5). For relationship degree classification, the top 7 features (i.e., red dash line in Figure 4B decided by

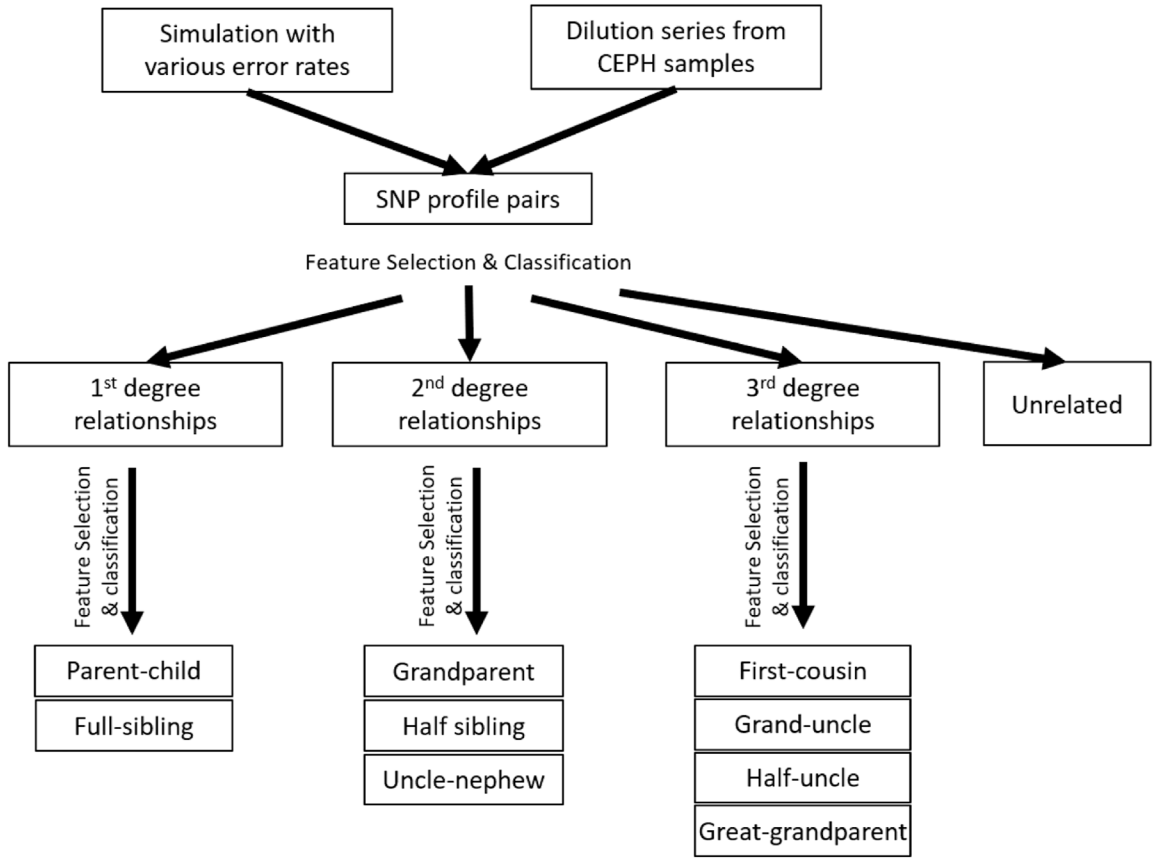


FIGURE 2
Experimental design and workflow of the whole study. The hierarchical classification was implemented with the simulation data, but not the real data, as the sample size of the real data was too small.

Figure 5A), including K1, j4, j7, K0, IBS0, j6 and j8, were selected as the final feature set because these 7 features had the highest accuracy among all the GERs.

Similar feature selections were conducted for classifying the relationship types within the 1st degree, the 2nd degree, and the 3rd degree (i.e., the ranking and counts in Figures 6D–F decided by Figures 6A–C). Figures 6D–F summarized the top features across different GERs (the red dash line in Figures 6D–F decided by Figures 5B–D). To balance the classification accuracy and robustness to genotyping errors, we selected the top 3 features for relationship types in the 1st degree, top 13 features for relationship types in the 2nd degree, and top 10 features for relationship types in the 3rd degree (Figures 5, 6).

Plain classification and hierarchical classification

For relationship degree classification, with the selected 7 top-performing features (i.e., K1, j4, j7, K0, IBS0, j6 and j8), close to 100% classification accuracies were obtained with various GERs (e.g., 99.02% with GER = 0 and 95.24% with GER = 0.1), which was much higher and robust compared with using a single feature K1 (e.g., 95.74% with GER = 0 and 75.45% with GER = 0.1; Figure 4). Apparently, in addition to K1, the other four features in the final feature set substantially improved the accuracy and robustness of the relationship degree classification.

The classification was employed for further classifying the relationship types within each classified relationship degree. The accuracies of classifying the relationship types within the 1st degree were almost 100% with the selected 3 features across all genotyping errors (Figure 6). In contrast, the method suggested in (Manichaikul et al., 2010) to differentiate parent-child and full-sibling (e.g., K1+IBS0) performed very well when genotyping errors were low (e.g., GER ≤0.03), but not for higher GERs (e.g., only 59.9% accuracy with GER = 0.05).

As expected, the classification accuracies within the 2nd degree and 3rd degree were much lower and were substantially affected by the GERs (Figures 5, 6). For the 2nd degree relationship types, the final 13 features together can reach

10.3389/fgene.2022.971242

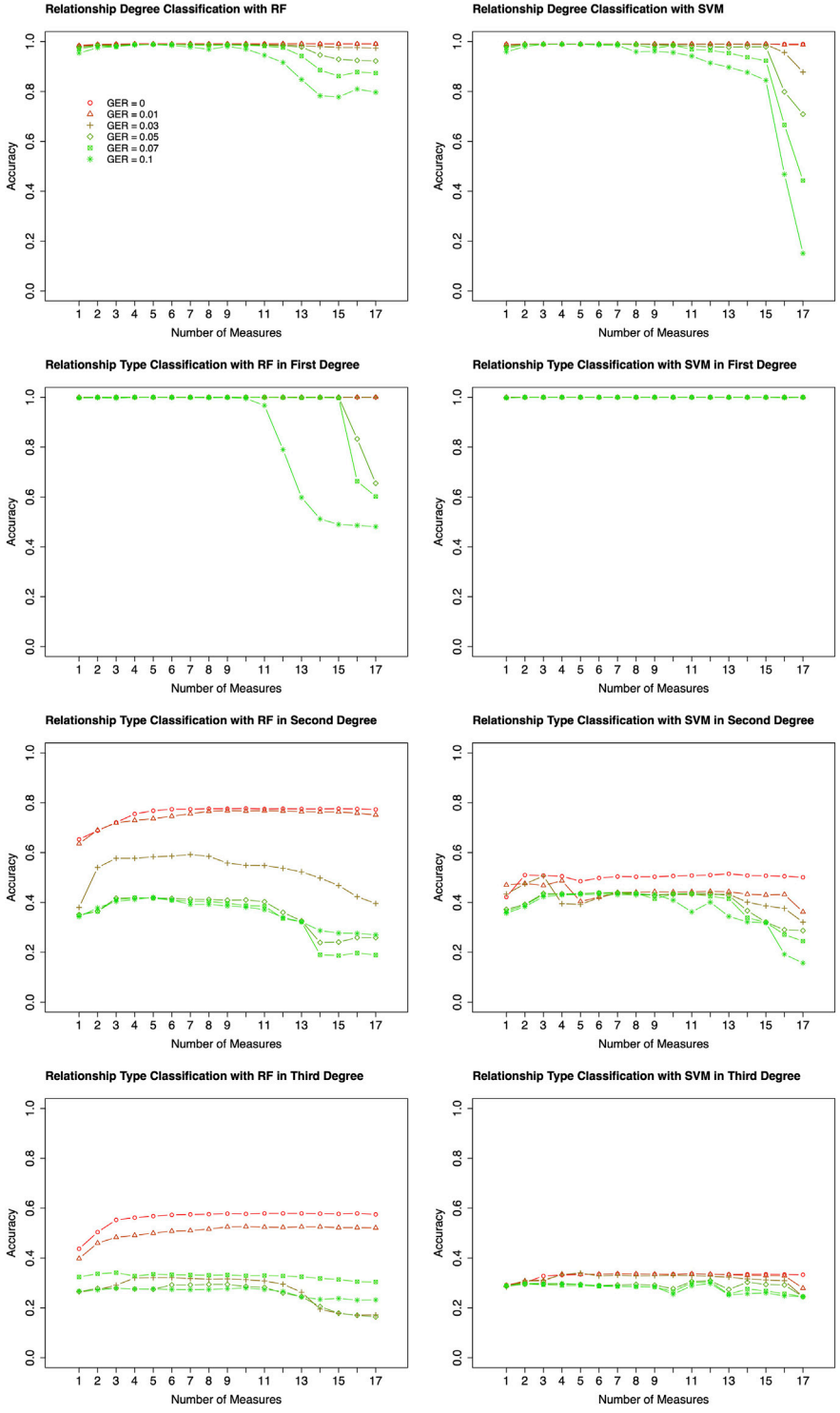


FIGURE 3
Classification algorithm comparisons between Random Forest (RF) and Support Vector Machine (SVM). Two algorithms (left four plots for RF; right four plots for SVM) were employed to conduct forward feature selection with 10-fold cross-validation for relationship degree, relationship types within the 1st degree, relationship types within the 2nd degree, and relationship types within the 3rd degree. Different genotyping error rates were presented with different colors. The *x*-axis is the number of selected measures (or features) in each step of the forward selection. GER = genotyping error rate of the test dataset.

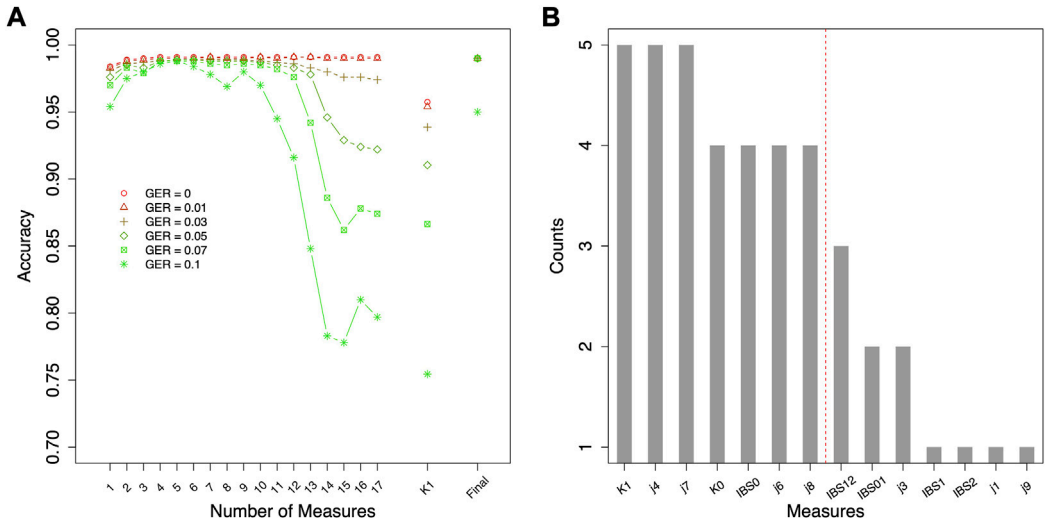


FIGURE 4
Feature selection for classifying relationship degrees using data simulated with various genotyping errors. **(A)** the accuracies by the forward feature selection with various genotyping errors, in which the ranking of the features for different genotyping error rates (GERs) was different (e.g., the first features were K1 and j7 with GER = 0 and 0.1, respectively), and **(B)** the counts of the commonly selected features across all genotyping errors (e.g., K1 was selected in the feature selections with all six genotyping errors). The features on the left of the red dash line were selected as final features. Final = classification with the selected top-performing features; GER = genotyping error rate of the test dataset.

77.5% accuracy with GER = 0, which was much higher than K1 alone (i.e., 33.3%). The accuracy differences between the final 13 feature set and K1 decreases with increasing GERs (e.g., 32.1% with the final feature set vs. 31.0% with K1 alone, when GER = 0.1). Thus, selecting and combining multiple features could improve the accuracy and robustness, particularly with lower GERs, for classifying relationship types within the 2nd degree.

For the 3rd degree relationships (Figures 5, 6), the accuracies with the final 10 features were higher with different GERs than those with K1 alone. When the GER was low (e.g., GER≤0.05), the final 10 features performed better than K1 alone (e.g., 33.3% with the final feature set vs. 25.2% with K1 alone, when GER = 0). With the GER increasing, the accuracy of the final 10 features could decrease quicker than that with K1 alone (e.g., 24.6% with the final feature set vs. 25.5% with K1 alone, when GER = 0.1). With 4 relationships within the 3rd degree and equal sample size for all relationships, the accuracy with random guessing among the 3rd degree relationships is 0.25. When the GER is high, the selected 10 features perform similar to random guessing.

Overall, the accuracy of classifying relationships within each degree and the robustness to genotyping errors were improved by selecting appropriate features, although the improvement decreases with more distant relationship degrees. As expected, the classifications within the 2nd and the 3rd degree relationships were still challenging. Better features, which contain more than only genotypes, may be needed to improve the accuracy, such as features with allele frequencies in populations.

Effect of missing data

To evaluate the impact of marker density on the classification accuracy, 50, 75, and 90% SNPs were randomly deleted from the 419 K SNPs, and the same analysis as above was conducted on these reduced datasets. Figure 7 compares the accuracies of using 7 top-performing features (Figure 4B) or using K1 only. The accuracies were the average of 10 different runs. Similar to the above analysis, the top-performing features substantially outperformed K1 alone for high genotyping errors. However, the missing data only changed the accuracies when GER is extremely high (e.g., accuracy = 83.7% with ~42 K SNPs and GER = 0.1), which was consistent with the results in (Manichaikul et al., 2010) and implied that our method is sensitive to SNP density only when GER is very high.

Effect of various genotyping error rates in the training dataset

In the above studies, the true GER was assumed unknown, and the simulated data with no genotyping error was used as the training dataset. However, if the GERs are known or may be precisely estimated, more appropriate training datasets, which have the same or similar GERs as the test sets, may be used to train the classifiers, and higher accuracies may be obtained.

Figure 8 shows the accuracies of classifying relationships within the same degrees with various genotyping errors in the

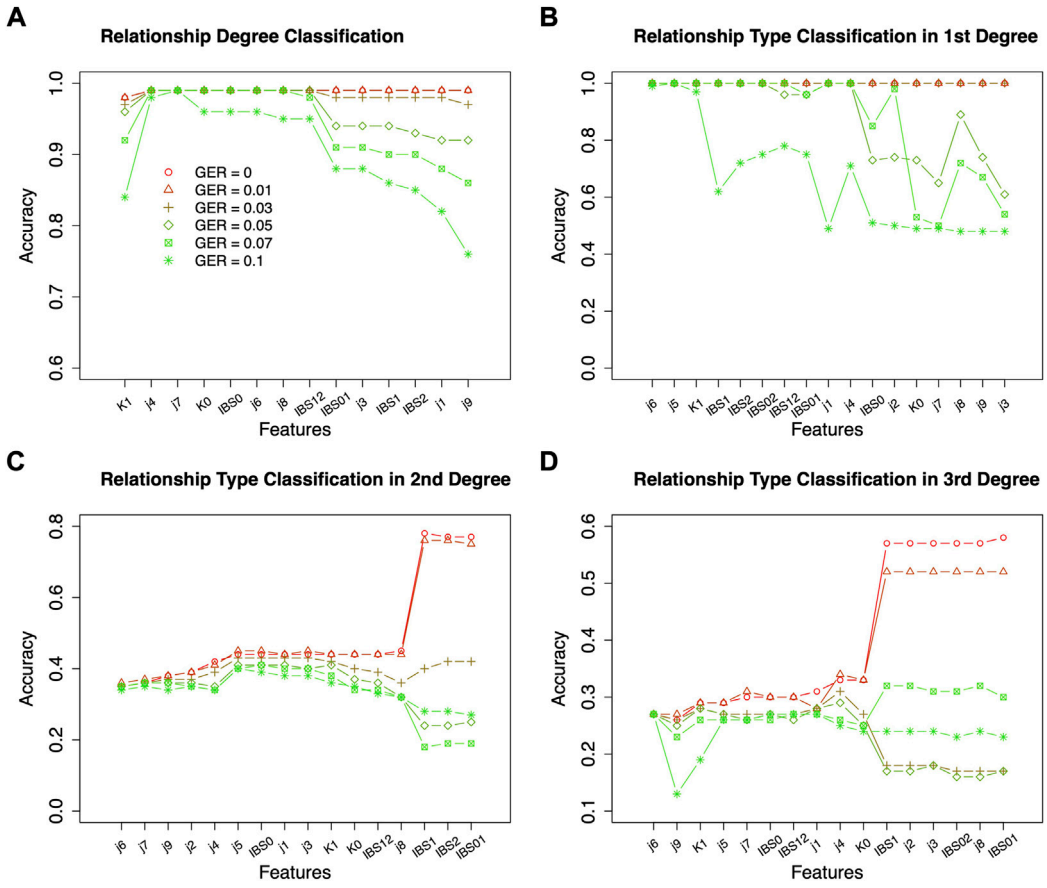


FIGURE 5
Classification accuracies with the selected features ranked as in Figures 3, 5 for relationship degree and types using data simulated with various genotyping errors. GER = genotyping error rate of the test dataset. **(A)** the accuracies by the forward features selection for the relationship degrees, **(B)** the accuracies by the forward features selection for the 1st degree relationships, **(C)** the accuracies by the forward features selection for the 2nd degree relationships, **(D)** the accuracies by the forward features selection for the 3rd degree relationships.

training dataset and test set. For the 1st degree relationships, all accuracies were close to 100%, and no impact could be observed. For the 2nd and 3rd degree relationships, the highest classification accuracy always is presented when the train datasets and test datasets have the same GER. Therefore, correctly estimating the GER of a sample could substantially increase the relationship estimation accuracy.

Relationship degree classification with real data

The classifiers, trained by the simulation data with various GERs, were used to classify the pairs in the real samples into degrees, with 50 ng, 500 pg or 100 pg DNA, and ~4M or 419 K SNPs (Figure 1). K1 with the cutoff thresholds defined in (Manichaikul et al., 2010) was used as the baseline to evaluate performance improvement. In general, the highest accuracies were achieved with close to the true GERs of the test sets (Figure 9). With 50 vs. 50 ng (i.e., both reference and test samples were genotyped with 50 ng DNA), the highest accuracy was 100% (= 28/28) with test sets' GERs ranging from 0.05 to 0.07 with 419 K SNPs, or accuracy was 96.4% (= 27/28) with GERs from 0 to 0.01 with 4M SNPs. K1 alone achieved 100% accuracy with both 419 K SNPs and 4M SNPs (Figure 9).

With 50 ng vs. 500pg, which may reflect more realistic scenarios in many missing persons cases, the highest accuracies were 92.9% (GER = 0.3) and 89.3% (GER = 0.07) with 419 K or 4M SNPs, respectively. In contrast, the K1 alone only achieved 67.9 and 57.1% with 419 K or 4M SNPs, respectively. Our approach outperformed K1 for high GER profiles generated from low-quality samples. Similar patterns were observed with 50 ng vs. 100pg, in which the highest accuracies were achieved with GER of 0.5 for 419 K SNPs (i.e., 71.4%) and with GER of 0.3 for 4M SNPs (i.e., 60.7%), respectively. The accuracy with K1 alone was only 39.3%.

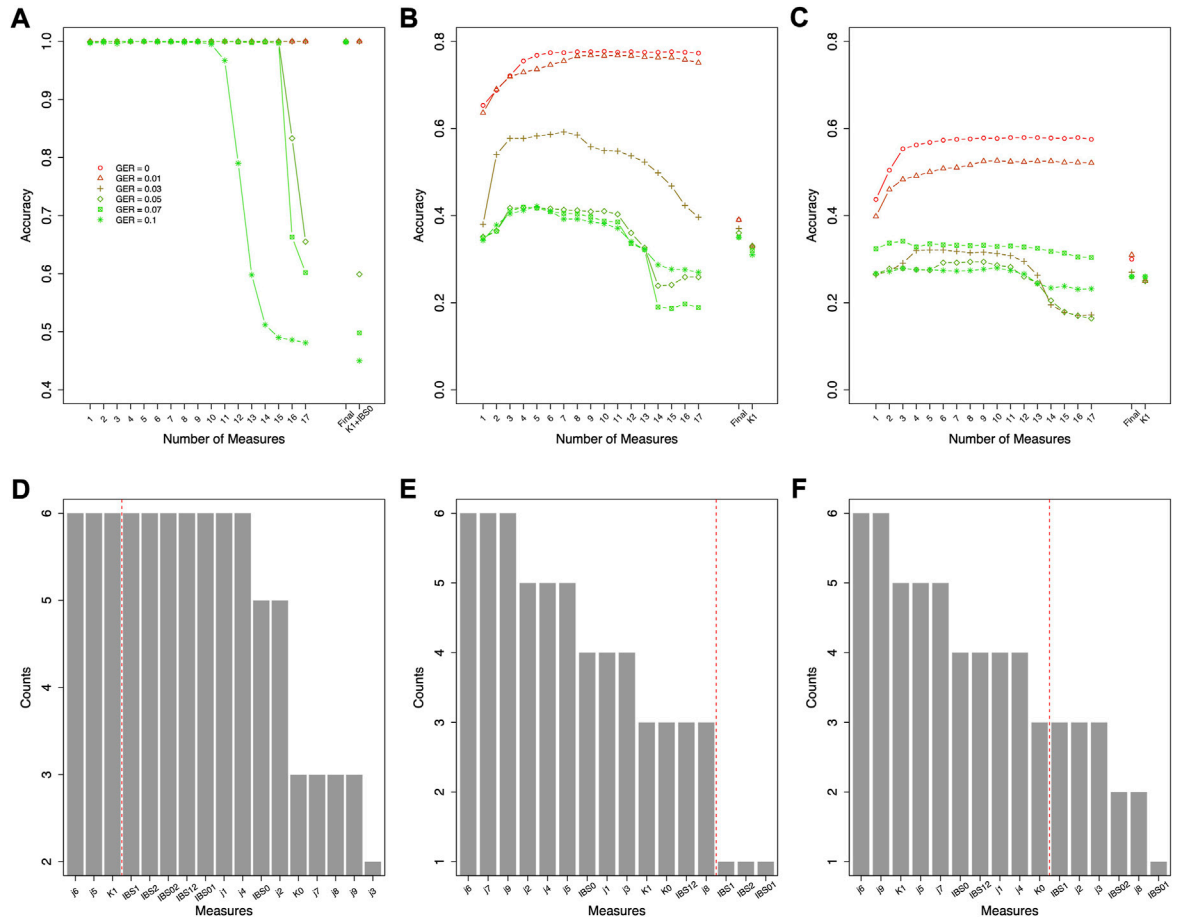


FIGURE 6
Forward feature selections and classification accuracies for relationship types using data simulated with various genotyping error rates. **(A)** the accuracies for the 1st degree relationships, **(B)** the accuracies for the 2nd degree relationships, **(C)** the accuracies for the 3rd degree relationships, **(D)** the counts of the selected features for the 1st degree relationships across all genotyping errors (e.g., K1 and K0 were selected in the feature selections with all six genotyping errors), **(E)** the counts of the selected features for the 2nd degree relationships across all genotyping errors, and **(F)** the counts of the selected features for the 3rd degree relationships across all genotyping errors. The features on the left of the red dash line were selected as final features. Final = classification with the final features; GER = genotyping error rate of the test dataset.

Therefore, higher kinship estimation accuracies could be achieved with our machine learning approach.

In addition, the GERs of these diluted samples might be roughly estimated, as the highest accuracies are likely obtained when the train and test sets share the same genotyping errors (Figure 8). The samples with 50 ng might have a GER lower than 0.03, the GERs of the samples with 500 pg might range from 0.05 to 0.07 if the test with 4M SNPs was considered to be more reliable, and the GERs of the samples with 100 pg might range from 0.2 to 0.3.

To further evaluate the performance of the IBD segment method in real data, we uploaded these 28 SNPs profiles (with ~4 million SNPs) to GEDmatch. GEDmatch only takes SNPs in the GSA panel in calculation, thus only ~419 K SNPs were used in calculation. As expected, the accuracy of GEDmatch for profiles generated with 50 ng DNA was 100% (Figure 9 & Supplementary Table S2), as these profiles had minimum genotyping errors. However, the accuracies of GEDmatch were much lower for profiles with 500 and 500 pg DNA, compared with the machine learning approach. For example, the total IBD segments between samples 13,047 and 7,046 dramatically decreased with the reduction of DNA concentration (i.e., 3,571.1 cM with 50ng, but 0 cM with 500 and 100 pg). With 500pg, 15 related pairs (out of 17) were determined as unrelated (i.e., 0 cM), one 1st degree pair was determined as 3rd degree, and 1st degree pair was determined as 8th degree. With 100 pg, 16 related pairs were determined as unrelated, and one 1st degree pair was determined as 3rd degree. If unrelated pairs were excluded in comparisons, the accuracies of GEDmatch for profiles with

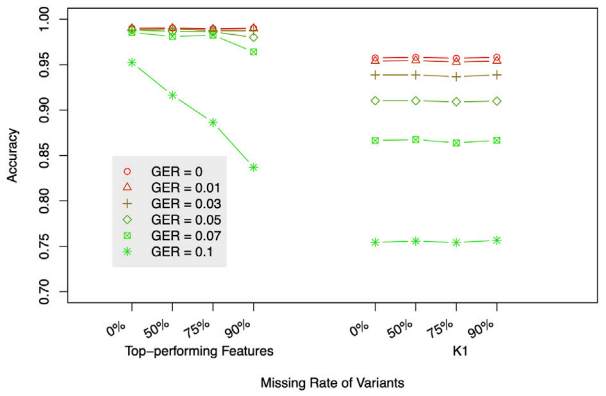


FIGURE 7
The impact of missing data on relationship degree classification. The label of the *x*-axis represents missing rates with two different sets of features (top-performing features vs. K1). GER = genotyping error rate of the test dataset.

500 and 100 pg were 0%. Thus, for profiles with high genotyping errors, GEDmatch may not be a good tool to search the true relatives.

Discussion

This study developed a novel machine learning approach for estimating the relationship types with high error SNP profiles. In this approach, a hierarchical classification strategy was employed first to classify the relationships by degree and then the relationships within each degree separately. For each classification, feature selection was implemented to gain better performance. Both simulated and real data sets were utilized in evaluating this approach, and the accuracies of this approach were higher than K1 alone (the most commonly used measure) and also other individual measures; namely, this approach was more robust than individual measures for SNP profiles with genotyping errors. In addition, the highest accuracy could be obtained by providing the same GERs in the train and test sets, and thus estimating genotyping errors of the SNP profiles is critical to obtaining high accuracy of relationship estimation.

The accuracy for estimating the degrees of the relationships was close to 100% using simulation data, which showed that the feature selection could be helpful in improving the robustness of classification. K1 was sensitive to genotyping errors, particularly when the GER was higher than 0.07 (Figure 4). Adding more features could substantially improve the accuracy (i.e., close to 100%) and the robustness to errors, which implies that most of the errors in kinship estimation due to genotyping errors could be corrected by adding other features.

The accuracy of estimating the degrees of the relationships obtained from the real data was much lower than the simulated data, which indicated that the simulation model used in Ped-sim might not precisely reflect the genotyping errors in the real data, and thus may result in overfitting in the training dataset. Better genotyping error models (de Vries et al., 2022; Nagraj et al., 2022) need to be developed to simulate SNP profiles that better approximate real SNP profiles generated from low-quality samples. However, in the simulations, the most important issue may be "what error rate should we assign to each type of error defined in the model?" As far as we know, limited studies have been done

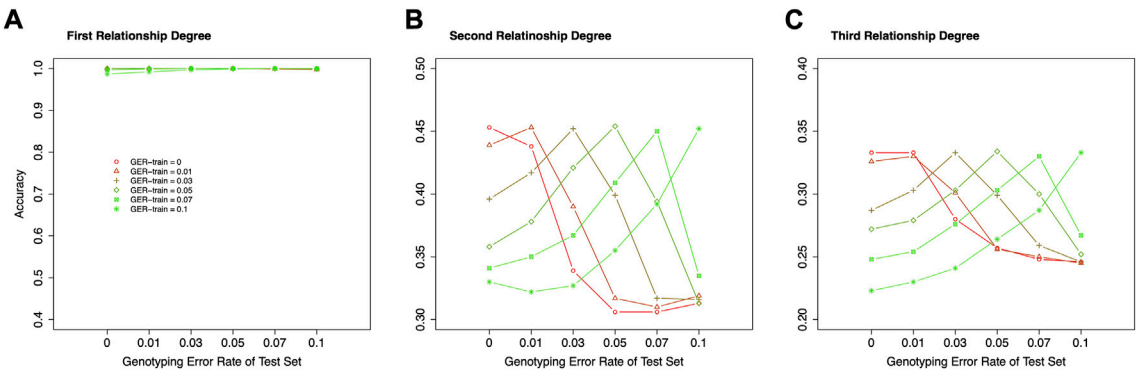


FIGURE 8
The effect of genotyping error rate of the training dataset in relationship type classification. **(A)** the accuracies with the final selected features for the 1st degree relationships in the training and test datasets with different genotyping error rates, **(B)** the accuracies with the final selected features for the 2nd degree relationships in the training and test datasets with different genotyping error rates, **(C)** the accuracies with the final selected features for the 3rd degree relationships in the training and test datasets with different genotyping error rates. The labels of the *x*-axis **(A–C)** represent the different genotyping error rates of the test dataset. The labels of the *y*-axis **(A–C)** represent classification accuracy. GER-train = genotyping error rate of the training dataset.

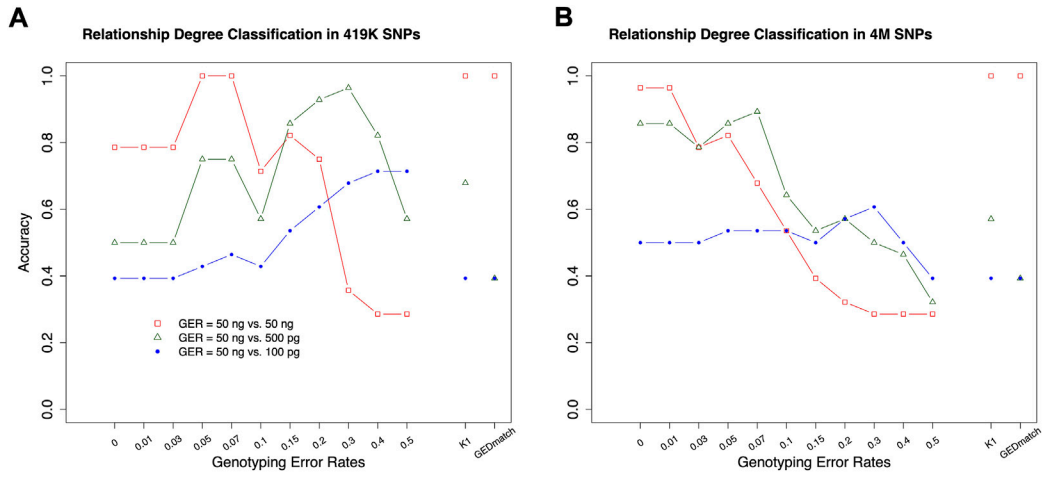


FIGURE 9
The accuracies of classifying relationship degrees with the UTAH family in Figure 1. The accuracies were estimated with the final selected features for the relationship degrees in the simulated training and real test datasets (UTAH family) with different genotyping error rates. **(A)** test dataset with 419K SNPs, **(B)** test dataset with 4M SNPs. The labels of the *x*-axis **(A,B)** represent the different genotyping error rates of the training dataset. The labels of the *y*-axis **(A,B)** represent classification accuracy. In the legend, GER denotes genotyping error rate of the test dataset. The GERs of the test datasets were represented using different colors and icons. GEDmatch denotes the accuracies obtained from GEDmatch website. K1 denotes the accuracies obtained from KING-robust.

to generate empirical data for estimating the error rates (or the range of the error rates) for different types of genotype errors, nor which model best fits the real WGS data generated from the missing persons samples. Thus, the commonly adopted simulation tool, Ped-Sim was used, which a pragmatic solution with one single parameter and the simulations could be better controlled on a reasonable scale. Nevertheless, the classifier trained by the simulated data still outperformed the individual measures. The more sophisticated simulation methods will be tested in future studies.

It is relatively easy to differentiate parent-child and full-sibling within the 1st degree. The combination of IBS0 and K1 could reach 100% accuracy when GERs were less than 0.05. But it dropped quickly when the GER was higher than or equal to 0.05 (Figure 6A). With the selected features, the classification accuracy could reach 100% across all the genotyping error levels. However, the accuracies for estimating the relationship within the 2nd and 3rd degrees were much lower, which was consistent with previous studies (Epstein et al., 2000; Huff et al., 2011; Henn et al., 2012; Ramstetter et al., 2017). In particular, the accuracies of estimating relationships could be equivalent to random guessing (i.e., 33 and 25% within the 2nd and 3rd degree, respectively), if the genotyping errors in the training dataset and the test set were largely different (e.g., 0 for the training dataset and 0.1 for the test set). With more accurate genotyping error estimations, the classifier can be trained with more proper data (i.e., the data with the same GERs as the test set), and the classification accuracy could be improved.

The genotyping errors could come from every step of the genotyping or sequencing process, including the polymerase chain reaction (PCR), sequencing chemistry, hybridization, signal detection, data collection, base-calling, sequence alignment, variant calling, etc. The GER would depend on the quantity and quality of the samples, the genotyping or sequencing protocols, and the bioinformatics analysis pipeline(s). Data cleaning in the bioinformatics analysis (e.g., removing sequence reads with low-quality scores) could reduce the GER with the cost of losing variants. Fortunately, our study and Shcherbina's study (Shcherbina et al., 2016a; Shcherbina et al., 2016a) showed that losing 90% of the SNPs in the GSA panel did not affect the estimation accuracy. Thus, a stringent data cleaning process could be implemented to remove low-quality data and lower the GERs. In addition, a method to precisely estimate the GER of a SNP profile, or at least a range of the GERs with confidence levels, would have practical value (but is yet to be developed).

In this study, two strong classifiers, RF and SVM, were tested. However, more recently developed classification algorithms, such as XGBoost (Chen & Guestrin, 2016) and deep learning (LeCun et al., 2015), may further improve the performance with proper parameter tuning. It is also worth noting that feature selection is very important to increase classification accuracy. The 17 features listed in Supplementary Table S1 were collected from previous literature, and some of these features might be

noisy for certain relationships or relationship degrees. Using the feature selection in relationship degree as an example, when the features were added one by one, the accuracy experienced three stages, raising, platform, and falling. It showed that some features increased the classification accuracy, but some features may be irrelevant, noisy, and even reduce the accuracies for certain relationship degrees or types, partially due to the genotyping errors.

This current method does not require allele frequencies as input, which is the case for many missing person cases. Additional features based on allele frequencies (e.g., the cumulative likelihood of observing a SNP profile given a specific population) may be included in a future study, as allele frequencies of the SNPs could provide more information than just SNP genotypes, and thus higher accuracy could be achieved by combining the features based on genotypes and features based on allele frequencies. However, the effect of inaccurate allele frequencies (e.g., use of African frequencies for Hispanic samples) is yet to be investigated. The features on IBD segments (e.g., the average length of the IBD segments, the total length of the IBD segments, etc.) (Hill & White, 2013; Ramstetter et al., 2018) may not work well with high genotyping errors, as the segments could be easily interrupted by the errors. However, those IBD segment features could be included for cases with negligible errors (i.e., cases with high quality and quantity samples). The GEDmatch results of real data showed that the total length of IBD segments could accurately identify 1st, 2nd, and 3rd pedigree degrees using the samples with low-level genotyping errors (i.e., samples with 50 ng DNA). However, the majority of the related pairs were not detected with GEDMatch when the genotyping error rates were high (i.e., samples with 500 and 100 pg DNA). In addition, in many missing persons cases, the samples' ancestry information may not be available or precisely determined. If the sample is admixed or belongs to an admixture population, the genome-wide relatedness methods such as KING will lead to bias estimation (Thornton et al., 2012; Conomos et al., 2016). The performance of this machine learning approach has not been tested in the admixed population as only European samples were involved in our study. The effect of the admixture population will be considered in our future method development.

To summarize, a novel machine learning-based approach was developed in this study to combine multiple measures and estimate the relationships for profiles with high GERs. Substantial accuracy increase and robustness improvement were observed in determining both relationship degrees and relationship types, which imply that the machine learning approach can increase the robustness of relationship estimations. Further improvement may be conducted by combining more features based on allele frequencies and IBD segments.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: GSE209804.

Ethics statement

The studies involving human participants were reviewed and approved by North Texas Regional Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

Author contributions

JG conceived the idea, designed the study, and finalized the manuscript. MH performed most of the computational analyses, prepared the figures and tables, and wrote the first draft. ML and HL contributed to computational analyses. JK and AS contributed to the wet-lab work. BB contributed to the study design and manuscript. All authors commented on the manuscript, read and approved the final manuscript.

Funding

This study was supported in part by award 2019-DU-BX-0046 (Dense DNA Data for Enhanced Missing Persons Identification), awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.971242/full#supplementary-material

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109.vidual

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., et al. (2015). A global reference for human genetic variation. *Nature* 526 (7571), 68–74. doi:10.1038/nature15393

Boehnke, M., and Cox, N. J. (1997). Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet.* 61 (2), 423–429. doi:10.1086/514862

Browning, B. L., and Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* 88 (2), 173–182. doi:10.1016/j.ajhg.2011.01.010

Caballero, M., Seidman, D. N., Qiao, Y., Sannerud, J., Dyer, T. D., Lehman, D. M., et al. (2019). Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLoS Genet.* 15 (12), 10079799–e1008029. doi:10.1371/journal.pgen.1007979

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data MiningAugu* 13-17, 785. doi:10.1145/2939672.2939785

Conomos, M. P., Miller, M. B., and Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* 39 (4), 276–293. doi:10.1002/gepi.21896

Conomos, M. P., Reiner, A. P., Weir, B. S., and Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* 98 (1), 127–148. doi:10.1016/j.ajhg.2015.11.022

Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J. M., and White, R. (1990). Centre d'Etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. *Genomics* 6 (3), 575–577. doi:10.1016/0888-7543(90)90491-C

de Vries, J. H., Kling, D., Vidaki, A., Arp, P., Kalamara, V., Verbiest, M. M. P. J., et al. (2022). Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy. *Forensic Sci. Int. Genet.* 56, 102625. doi:10.1016/j.fsigen.2021.102625

Epstein, M. P., Duren, W. L., and Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet.* 67 (5), 1219–1231. doi:10.1016/S0002-9297(07)62952-8

Galván-Femenía, I., Barceló-Vidal, C., Sumoy, L., Moreno, V., de Cid, R., and Graffelman, J. (2021). A likelihood ratio approach for identifying three-quarter siblings in genetic databases. *Heredity* 126 (3), 537–547. doi:10.1038/s41437-020-00392-8

Ge, J., Budowle, B., and Chakraborty, R. (2011). Choosing relatives for DNA identification of missing persons. *J. Forensic Sci.* 56 (Suppl. 1), S23–S28. doi:10.1111/j.1556-4029.2010.01631.x

Ge, J., and Budowle, B. (2020). How many familial relationship testing results could be wrong? *PLoS Genet.* 16 (8), 10089299–e1008936. doi:10.1371/JOURNAL.PGEN.1008929

Ge, J., Chakraborty, R., Eisenberg, A., and Budowle, B. (2011). Comparisons of familial DNA Database searching strategies. *J. Forensic Sci.* 56 (6), 1448–1456. doi:10.1111/j.1556-4029.2011.01867.x

Greytak, E. M., Moore, C. C., and Armentrout, S. L. (2019). Genetic genealogy for cold case and active investigations. *Forensic Sci. Int.* 299, 103–113. doi:10.1016/j.forsciint.2019.03.039

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., et al. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* 19 (2), 318–326. doi:10.1101/gr.081398.108

Hares, D. R. (2015). Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 17, 33–34. doi:10.1016/j.fsigen.2015.03.006

Heinrich, V., Kamphans, T., Mundlos, S., Robinson, P. N., and Krawitz, P. M. (2017). A likelihood ratio-based method to predict exact pedigrees for complex families from next-generation sequencing data. *Bioinformatics* 33 (1), 72–78. doi:10.1093/bioinformatics/btw550

Henn, B. M., Hon, L., Macpherson, J. M., Eriksson, N., and Saxonov, S. (2012). Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PloS one* 7 (4), e34267. doi:10.1371/journal.pone.0034267

Hill, W. G., and White, I. M. S. (2013). Identification of pedigree relationship from genome sharing. *G3* 3 (9), 1553–1571. doi:10.1534/g3.113.007500

Huff, C. D., Witherspoon, D. J., Simonson, T. S., Xing, J., Watkins, W. S., Zhang, Y., et al. (2011). Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res.* 21 (5), 768–774. doi:10.1101/gr.115972.110

Kling, D. (2019). On the use of dense sets of SNP markers and their potential in relationship inference. *Forensic Sci. Int. Genet.* 39, 19–31. doi:10.1016/j.fsigen.2018.11.022

Korneliussen, T. S., and Moltke, I. (2015). NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. *G3* 11 (8), 4009–4011. doi:10.1093/bioinformatics/btv509

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553), 436–444. doi:10.1038/nature14539

Li, H., Glusman, G., Hu, H., ShankaracharyaCaballero, J., Hubley, R., et al. (2014). Relationship estimation from whole-genome sequence data. *PLoS Genet.* 10 (1), e1004144. doi:10.1371/journal.pgen.1004144

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26 (22), 2867–2873. doi:10.1093/bioinformatics/btq559

Moltke, I., and Albrechtsen, A. (2014). RelateAdmix : A software tool for estimating relatedness between admixed individuals. *Bioinformatics* 30 (7), 1027–1028. doi:10.1093/bioinformatics/btt652

Morrison, J. (2013). Characterization and correction of error in genome-wide IBD estimation for samples with population structure : *Genetic epidemiology*, 37 (6), 635–641. doi:10.1002/gepi.21737

Nagraj, V. P., Scholz, M., Jessa, S., Ge, J., Woerner, A. E., Huang, M., et al. (2022). vcferr: Development, validation, and application of a SNP genotyping error simulation framework. BioRxiv. Available at: https://www.biorxiv.org/content/10.1101/2022.03.28.485853v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.03.28.485853v1.abstract.

Nøhr, A. K., Hanghøj, K., Erill, G. G., Li, Z., Moltke, I., and Albrechtsen, A. (2021). NGSremix: A software tool for estimating pairwise relatedness between admixed individuals from next-generation sequencing data. *G3* 11 (8), 1–9. doi:10.1093/g3journal/jkab174

Pew, J., Muir, P. H., Wang, J., and Frasier, T. R. (2015). related: An R package for analysing pairwise relatedness from codominant molecular markers. *Mol. Ecol. Resour.* 15 (3), 557–561. doi:10.1111/1755-0998.12323

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Qiao, Y., Sannerud, J. G., Basu-roy, S., Hayward, C., and Williams, A. L. (2021). Distinguishing pedigree relationships via multi-way identity by descent sharing and sex-specific genetic maps. *Am. J. Hum. Genet.* 108 (1), 68–83. doi:10.1016/j.ajhg.2020.12.004

Ramstetter, M. D., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., et al. (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics* 207 (1), 75–82. doi:10.1534/genetics.117.1122

Ramstetter, M. D., Shenoy, S. A., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., et al. (2018). Inferring identical-by-descent sharing of sample ancestors promotes high-resolution relative detection. *Am. J. Hum. Genet.* 103 (1), 30–44. doi:10.1016/j.ajhg.2018.05.008

Seidman, D. N., Shenoy, S. A., Kim, M., Babu, R., Woods, I. G., Dyer, T. D., et al. (2020). Rapid , phase-free detection of long identity- by-descent segments enables effective relationship classification. *Am. J. Hum. Genet.* 106 (4), 453–466. doi:10.1016/j.ajhg.2020.02.012

Shcherbina, A., Ricke, D. O., Schwoebel, E., Boettcher, T., Zook, C., Bobrow, J., et al. (2016a). KinLinks: Software Toolkit for kinship analysis and pedigree generation from HTS datasets." in *IEEE symposium on technologies for homeland security (HST)*, 1. doi:10.1109/THS.2016.7568891

Shcherbina, A., Ricke, D., Schwoebel, E., Boettcher, T., Zook, C., Bobrow, J., et al. 2016b. KinLinks: Software toolkit for kinship analysis and pedigree generation from NGS datasets." in 2016 IEEE Symposium on Technologies for Homeland Security (HST). 10-11 May 2016. Waltham, MA, USA. 1

Sherry, S. T., Feolo, M., Jin, Y., and Scha, A. A. (2017). Quickly identifying identical and closely related subjects in large databases using genotype data. *PloS one* 12 (6), 1–28. doi:10.1371/journal.pone.0179106

Staples, J., Qiao, D., Cho, M. H., Silverman, E. K., Nickerson, D. A., Below, J. E., et al. (2014). PRIMUS: Rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am. J. Human Genet.* 95 (5), 553–564. doi:10.1016/j.ajhg.2014.10.005

Stevens, E. L., Heckenberg, G., Roberson, E. D. O., Baugher, J. D., and Thomas, J. (2011). Inference of relationships in population data using identity-by-descent and identity-by-state *PLoS Genet.* 7 (9), e1002287. doi:10.1371/journal.pgen.1002287

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-balcom, H. M., Caan, B. J., and Risch, N. (2012). Estimating Kinship in Admixed Populations. *Am. J. Human Genet.* 91 (1), 122–138. doi:10.1016/j.ajhg.2012.05.024

Tillmar, A., Fagerholm, S. A., Staaf, J., Sjölund, P., and Ansell, R. (2021). Getting the conclusive lead with investigative genetic genealogy—A successful case study of a 16 year old double murder in Sweden. *Forensic Sci. Int. Genet.* 53, 102525. doi:10.1016/j.fsigen.2021.102525

Turner, S., Nagraj, V. P., Scholz, M., Jessa, S., Acevedo, C., Ge, J., et al. (2022). Evaluating the impact of dropout and genotyping error on SNP-based kinship analysis with forensic samples. *Front. Genet.* 13, 882268. doi:10.3389/fgene.2022.882268

Wall, J. D., Tang, L. F., Zerbe, B., Kvale, M. N., Kwok, P. Y., Schaefer, C., et al. (2014). Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res.* 24 (11), 1734–1739. doi:10.1101/gr.168393.113

Waples, R. K., Albrechtsen, A., and Moltke, I. (2019). Allele frequency—free inference of close familial relationships from genotypes or low—depth sequencing data. *Mol. Ecol.* 28 (1), 35–48. doi:10.1111/mec.14954

Zhou, Y., Browning, S. R., and Browning, B. L. (2020). A fast and simple method for detecting identity-by-descent segments in large-scale data. *Am. J. Hum. Genet.* 106 (4), 426–437. doi:10.1016/j.ajhg.2020.02.010

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 12 October 2022
DOI 10.3389/fgene.2022.947758

Check for updates

OPEN ACCESS

EDITED BY
Ryan Lan-Hai Wei,
Inner Mongolia Normal University, China

REVIEWED BY
Shaoqing Wen,
Fudan University, China
Jun Yao,
China Medical University, China

*CORRESPONDENCE
Youfeng Wen,
wenyf@jzmu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 19 May 2022
ACCEPTED 26 September 2022
PUBLISHED 12 October 2022

CITATION
Hou X, Zhang X, Li X, Huang T, Li W, Zhang H, Huang H and Wen Y (2022), Genomic insights into the genetic structure and population history of Mongolians in Liaoning Province. *Front. Genet.* 13:947758. doi: 10.3389/fgene.2022.947758

COPYRIGHT
© 2022 Hou, Zhang, Li, Huang, Li, Zhang, Huang and Wen. 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.

Genomic insights into the genetic structure and population history of Mongolians in Liaoning Province

Xuwei Hou†, Xianpeng Zhang†, Xin Li, Ting Huang, Wenhui Li, Hailong Zhang, He Huang and Youfeng Wen*

Institute of Biological Anthropology, Jinzhou Medical University, Jinzhou, China

The Mongolian population exceeds six million and is the largest population among the Mongolic speakers in China. However, the genetic structure and admixture history of the Mongolians are still unclear due to the limited number of samples and lower coverage of single-nucleotide polymorphism (SNP). In this study, we genotyped genome-wide data of over 700,000 SNPs in 38 Mongolian individuals from Fuxin in Liaoning Province to explore the genetic structure and population history based on typical and advanced population genetic analysis methods [principal component analysis (PCA), admixture, F_{ST}, f_3-statistics, f_4-statistics, *qp*Adm/*qp*Wave, *qp*Graph, ALDER, and TreeMix]. We found that Fuxin Mongolians had a close genetic relationship with Han people, northern Mongolians, other Mongolic speakers, and Tungusic speakers in East Asia. Also, we found that Neolithic millet farmers in the Yellow River Basin and West Liao River Basin and Neolithic hunter–gatherers in the Mongolian Plateau and Amur River Basin were the dominant ancestral sources, and there were additional gene flows related to Eurasian Steppe pastoralists and Neolithic Iranian farmers in the gene pool of Fuxin Mongolians. These results shed light on dynamic demographic history, complex population admixture, and multiple sources of genetic diversity in Fuxin Mongolians.

KEYWORDS

Mongolian, genetic structure, population admixture, Liaoning Province in China, genome-wide data

Introduction

Northeast Asia is a vast geographical region encompassing the Mongolian Plateau (MP), Yellow River Basin (YRB), West Liao River Basin (WLRB), Amur River Basin (ARB), Russian Far East, Korean Peninsula, and Japanese Islands. Recent studies indicated that frequent and complex population migration, exchange, and admixture events had happened in this region. The West Liao River Basin is considered the cradle of the Transeurasian language family. The "Transeurasian hypothesis" is supported by evidence from linguistics, archaeology, and genetics. It indicates that Japonic, Koreanic,

Tungusic, Mongolic, and Turkic languages are split from a proto-Transeurasian language, and the diffusion of Transeurasian language is related to the expansion of early millet farmers in the West Liao River Basin (Robbeets et al., 2021), but Wang et al. (2021a) hold opposite views: they did not find the West Liao River farmer-related ancestry in ancient populations in the Mongolian Plateau and Amur River Basin. The Yellow River Basin is considered the origin of the Sino-Tibetan language from different perspectives such as archaeology, genetics, and linguistics. It supports the "northern-origin hypothesis," and the diffusion of the Sino-Tibetan language also is considered to conform "farming–language dispersal hypothesis" that the Neolithic YRB millet farmers who are related to Yangshao and/or Majiayao cultures may be the ancestors of Sino-Tibetan language speakers (Sagart et al., 2019; Zhang et al., 2019; Wang et al., 2021a). Paleogenomic research studies showed that there is up to 14,000-year genetic continuity from ancient ARB population to modern Tungusic people, and modern Tungusic people show genetic homogeneity with each other (He et al., 2021; Mao et al., 2021; Wang et al., 2021a), but there is an exception in Tungusic speakers that Manchu people exhibit significant genetic similarity with northern Han people (Zhang et al., 2021b). The Eurasian Steppe zone is the largest steppe zone in the world that stretches from Eastern Europe through Mongolia to Northeast China, Bronze-Age Yamnaya Steppe pastoralists expanded eastward through the Eurasian Steppe, and then Afanasievo, Andronovo, and Sintashta cultures, which are related to the Yamnaya culture, established, respectively (Haak et al., 2015; de Barros Damgaard et al., 2018; Ning et al., 2019). Steppe pastoralists have a genetic influence on East Asian populations, but their genetic contributions are limited (Ning et al., 2019; Wang et al., 2021a; Yang et al., 2021). The influence of steppe pastoralists on the genetic formation of MP populations was discontinuous (Wang et al., 2021a), but the steppe pastoralist-related ancestry has persisted in Northwest China since the Early Bronze Age (~3000 BCE) (Ning et al., 2019; Wang et al., 2021c; Zhang et al., 2021a), and the steppe pastoralist-related ancestry also exists in the genetic makeup of modern populations in South Siberia and Western and Northern China (Feng et al., 2017; He et al., 2021; Ma et al., 2021). In general, previous studies have found dynamic demographic history and multiple sources of genetic diversity in Northeast Asia, but the genetic structure and affinity of modern populations in Northeast Asia remain unclear due to a limited number of samples.

Mongolians play an indispensable role in the formation of culture and genetic structure in Eurasia, the Mongol Empire controlled vast territories and trade routes stretching from East Asia to Europe in the 13th century, and there are genetic imprints of Mongolians which can be found in modern Eurasian populations (Dulik et al., 2011; Bai et al., 2014; Bai et al., 2018; Jeong et al., 2020). The Mongolian population exceeds six million and is the largest population among the Mongolic speakers in China. Modern Mongolians are widely distributed in China, but they mainly live in Inner Mongolia, Liaoning, Xinjiang, and other northern provinces. Chinese Mongolians from different regions have different genetic profiles, northern Mongolians possess significant genetic contribution from ancient ARB populations, southern Mongolians possess a majority of Neolithic YRB farmer-related ancestry, and Guizhou Mongolians harbor more southern ancestry related to Tai–Kadai, Austroasiatic, and Austronesian speakers. This reveals Mongolians gradually mixed with the local indigenous people along with their migration (Chen et al., 2021; He et al., 2021). Also, there is a different admixture history in western and eastern Chinese Mongolians, the western Mongolians receive more genetic influence of western Eurasians, and the Eastern Mongolians possess more Neolithic YRB- and ARB-related ancestry (He et al., 2021; Yang et al., 2021). Ancient Mongolia is formed by multiple tribes, every ancient Mongolian tribe experience a different origin, exchange, and admixture history, and their descendants live in different regions, which may be one of the reasons why genetic differences existed in modern Mongolians. "Tumet" tribe is a larger Mongolian tribe and has a controversial origin; it has been divided into two groups (Western Tumet group and Eastern Tumet group) according to geographical distribution since the 17th century. The Western Tumet group is mainly distributed in the Inner Mongolian Autonomous Region (Hohhot and Baotou), and the Eastern Tumet group is distributed in Liaoning Province (Fuxin and Chaoyang). Previous studies based on genome-wide data have reported the genetic structure and admixture history of Baotou Mongolians who are descendants of the Western Tumet group, but the genetic structure and admixture history of Liaoning Mongolians is still ambiguous. In this study, we generated genome-wide data of over 700,000 SNPs in 38 Mongolian individuals from Fuxin in Liaoning Province (Figure 1) and combined all available modern and ancient East Asian populations to investigate 1) the genetic profile and structure of Fuxin Mongolians; 2) genetic differences of Chinese Mongolians in different regions; 3) genetic affinity between Fuxin Mongolians and Han people, Tungusic speakers, and Mongolic speakers; 4) how many ancestral sources contributed to Fuxin Mongolians; and 5) to shed light on the proportions of the genetic contribution of the millet farmers in Yellow River Basin and West Liao River Basin, Northern Asian hunter–gatherers, and Western Eurasian populations.

Materials and methods

Sampling and genotyping

In this study, we collected peripheral blood samples from 38 Mongolian individuals (15 male and 23 female) in Fuxin, Liaoning Province, China. The detailed sample information is listed in Supplementary Table S1. Every participant signed the written informed consent before the study begins. The geographical position of the sampling population in this study

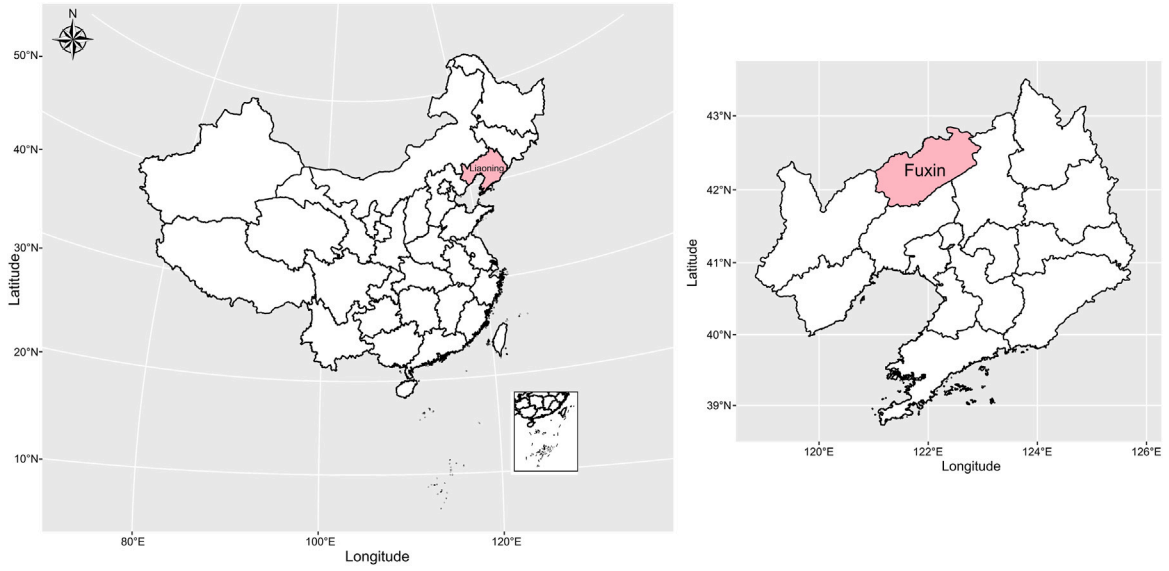


FIGURE 1
Geographical positions of the sampling population in this study.

is shown in Figure 1. All participants were required to be indigenous residents whose ancestors have lived in the sampling site for at least three generations, and self-declared ethnicity and family migration history were recorded. This project was reviewed and approved by the Medical Ethics Committee of Jinzhou Medical University (JZMULL2021101), and all procedures were carried out in accordance with the recommendations of the 2,000 Helsinki Declaration (Helsinki, 2001). DNA extraction was performed by using the Genomic DNA Extraction Kit following manufacturer's instructions, and all DNA samples were genotyped using Illumina WeGene V3 Arrays. The raw data contained 717,228 single-nucleotide polymorphisms (SNPs) and were filtered using PLINK 1.9 (Purcell et al., 2007) based on the predefined threshold (-maf: 0.01, -hwe: 0.0001, -mind: 0.01, and -geno: 0.01). Also, we applied Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011) software to estimate the genetic relationship between newly sampled Mongolian individuals, and the individual who has relatedness up to 0.125 (third-degree relative) with other newly sampled individuals was removed to guarantee studied individuals were unrelated (Supplementary Table S2; Supplementary Figure S1). Finally, we obtained a dataset containing 36 individuals with 477,492 SNPs which was used to perform the following population genetic analysis.

Data merging

In this study, we merged our newly genotyped 36 individuals' data with the Affymetrix Human Origins (HO) array dataset (Patterson et al., 2012) and recently published population data from Baotou Mongolian and Bijie Mongolian (Chen et al., 2021; Yang et al., 2021) to constitute a dataset containing 52,403 SNPs that was used to perform principal component analysis (PCA), F_{ST}, admixture, TreeMix, and ALDER. Also, we merged newly genotyped data with the 1240K dataset from the Reich Lab (https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data) to obtain a dataset with a larger number of SNPs (139,592 SNPs) for performing f_3-statistics, f_4-statistics, *qp*Wave/*qp*Adm, and *qp*Graph.

Principal component analysis and ADMIXTURE

In this study, PCA and admixture were used to explore the population structure in Eurasia. First, we performed the PCA based on the merged-HO dataset containing 52,403 SNPs from 1,488 individuals in 151 populations using the smartpca program in EIGENSOFT software (Patterson et al., 2006) with the following parameters: numoutlieriter: 0 and lsqproject: YES, and the visualization was carried out using R software, the background of 2-D plots was constructed based on modern populations, and all ancient populations were projected onto it. Next, we applied PLINK 1.9 (Purcell et al., 2007) with the parameters "-indep-pairwise 200 25 0.4" to remove SNPs in strong linkage disequilibrium, and then we performed admixture analysis, which is a model-based clustering analysis (Alexander et al., 2009), with a predefined number of ancestral sources (K)

ranging from 2 to 20. Also, the admixture analysis was also performed based on a merged-HO dataset, which contained 2,123 individuals from 179 populations. An optimal value of ancestral sources (K) was selected using the 10-fold cross-validation (CV) errors, which are listed in Supplementary Table S3. The visualization of admixture results was carried out using AncestryPainter (Feng et al., 2018) and R package pophelper.

Pairwise F_{ST} genetic distances

We used the smartpca program in EIGENSOFT software (Patterson et al., 2006) with parameters: fstonly: YES to calculate pairwise F_{ST} genetic distances between Fuxin Mongolian and other modern reference populations. The heatmap and phylogenetic tree were used to shed light on the relationship between Fuxin Mongolians and other populations. The heatmap was plotted using the R package pheatmap, the neighbor-joining (N-J) tree was constructed using MEGA X (Kumar et al., 2018), and the visualization was performed using the online tool iTOL (https://itol.embl.de/).

F-statistics

All *f*-statistics were carried out using ADMIXTOOLS (Patterson et al., 2012). First, we applied qp3pop in ADMIXTOOLS (Patterson et al., 2012) with default parameters to perform the three-population test (f_3-statistics). We calculated outgroup-f_3 (Mongolian_Fuxin, Y; Yoruba) to measure shared genetic drifts between Fuxin Mongolians and reference populations (Y) and computed admixture-f_3 (X, Y; Mongolian_Fuxin) to explore the potential admixture signals. Also, we then performed the four-population test (f_4-statistics) using the qpDstat program in ADMIXTOOLS (Patterson et al., 2012) with default parameters to examine shared alleles and infer the direction of the gene flow.

*qp*Adm/*qp*Wave and *qp*Graph

We used *qp*Wave/*qp*Adm programs in ADMIXTOOLS (Patterson et al., 2012) to determine the minimum number of ancestral sources and quantify the ancestral proportion. In this study, we used the following ten outgroups, namely, Mbuti, Malaysia_LN, Tianyuan, Papuan, Ust_Ishim, GreatAndaman, Kostenki14, Australian, Mixe, and Atayal to test two-way admixture models. Also, we used the *qp*Graph program in ADMIXTOOLS (Patterson et al., 2012) to explore the best fitting phylogenetic framework with population splits and gene flow events and reconstruct the deep population history of Fuxin Mongolians.

TreeMix and ALDER

To further explore the relationship between Fuxin Mongolians and other reference populations, we applied TreeMix v1.13 software (Pickrell and Pritchard, 2012) to construct a rooted maximum likelihood tree with gene flow events varying from 0 to 10, and the best-fitted models were chosen based on the predefined hypothesis and residual values. The admixture time and possible ancestral sources of Fuxin Mongolians were estimated by using multiple admixture-induced linkage disequilibrium for evolutionary relationships (ALDER) (Loh et al., 2013).

Y chromosomal and mtDNA lineages

We used an in-house script to assign the Y-chromosomal and mitochondrial haplogroups following the recommendations of the International Society of Genetic Genealogy (ISOGG; http://www.isogg.org/) and mtDNA PhyloTree17 (http://www.phylotree.org/). The haplogroup information of mtDNA and Y chromosome is listed in Supplementary Table S11.

Results

Population genetic structure

To explore the general patterns of genetic structure between Fuxin Mongolians and other Eurasian reference populations, we conducted PCA first. There were obvious genetic clusters such as Han cline, Mongolian cline, and Turkic speaker cline that could be observed which were consistent with their language categories and geographical distribution (Figure 2). This studied Mongolian people were surrounded by northern East Asian populations such as Han people, Tungusic speakers, Turkic speakers, Mongolic speakers, Tibeto-Burman speakers, and Japanese as well as Korean. We also found some ancient YRB and WLRB populations which were close to Fuxin Mongolians. There were obvious genetic differences between Fuxin Mongolians and Bijie Mongolians, modern Tai–Kadai, Hmong–Mien, and Austronesian speakers as well as ancient populations in Central Asia and Russia. Four present-day northern Mongolians formed a cluster, and their positions in Figure 2 were consistent with geographic distribution. Fuxin Mongolians were close to Hulunbuir Mongolians, Baotou Mongolians were located between Outer Mongolians and Mongolians from Fuxin and Hulunbuir, and Outer Mongolians were adjacent to ancient Mongolians. The same genetic structure could be observed in the results of pairwise-F_{ST} analysis, and we found people who belonged to the same language category always clustered together in East Asian populations. Fuxin Mongolians had a close genetic affinity with Hulunbuir Mongolians, and there were

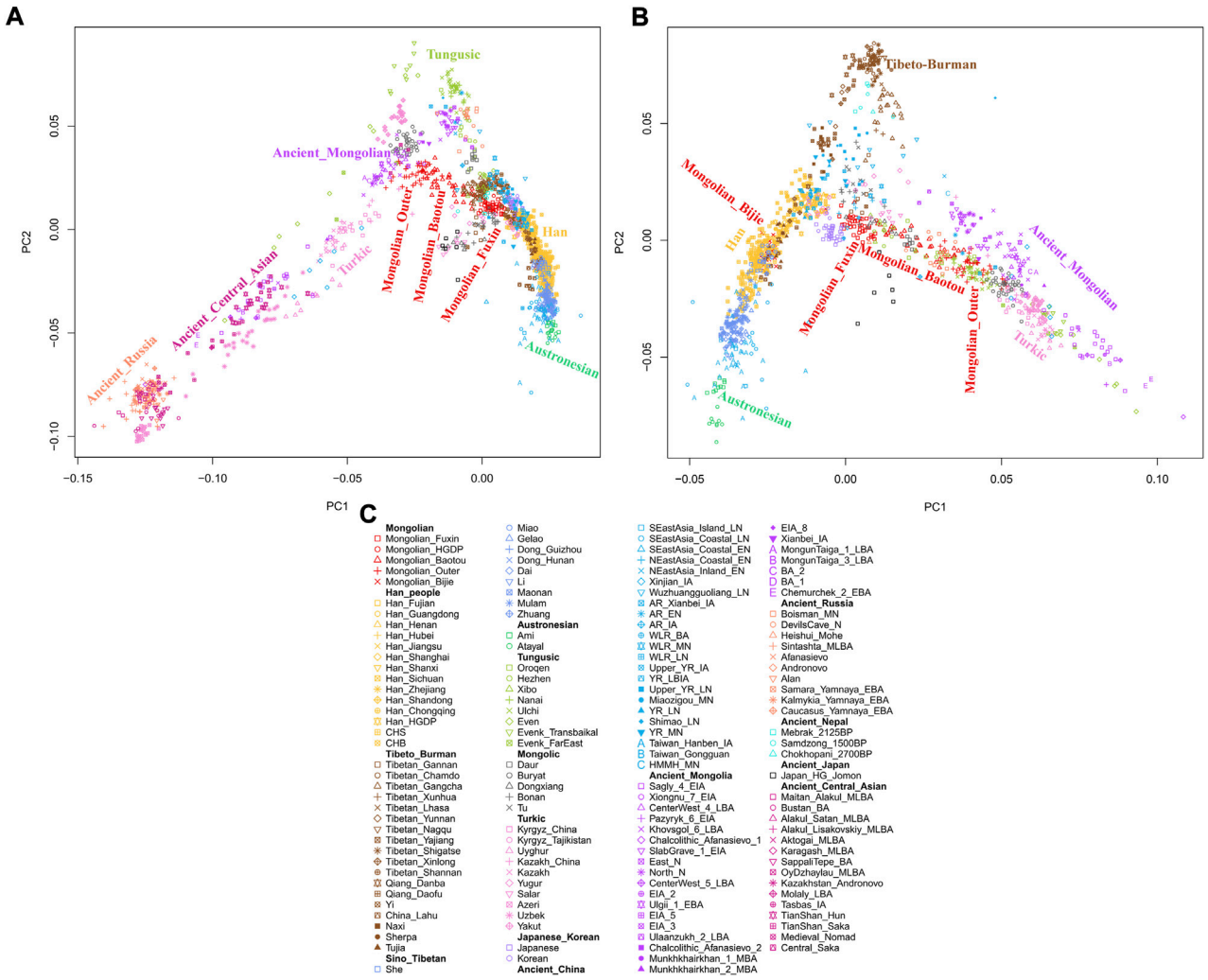


FIGURE 2
The population structure of modern and ancient Eurasian populations inferred by principal component analysis based on genome-wide data. The color of modern population was consistent with language categories, and all ancient populations were projected onto it. **(A~B)** PCA results showed an overview population relationship between Fuxin Mongolians and other modern and ancient populations. **(C)** Population list of PCA.

close genetic distances between Fuxin Mongolians and Han people (FST_{Shanxi} = 0.001 and FST $_{Henan}$ = 0.001), Baotou Mongolian (0.002), Xibo (0.002), Tu (0.002), and Korean (0.002) (Figure 3, Supplementary Table S4). In the model-based admixture clustering analysis, we found that when K = 8, the CV error was the lowest (Supplementary Table S3). The results showed that there were four dominant ancestral components which were marked in yellow, orange, pink, and deep green color in the genetic makeup of Fuxin Mongolians (Figure 4). The yellow ancestry was maximized in the hunter–gatherers from Nepal, it was also enriched in millet farmers from the Yellow River Basin, and it possessed the main proportion in the genetic makeup of modern Tibetans and Han people. The orange ancestry was maximized in the Iron-Age Hanben population from Taiwan and Neolithic southern populations in Fujian, and it also commonly existed in modern Tai–Kadai, Hmong–Mien, and Austronesian speakers. The pink ancestry was maximized in hunter–gatherers from the Amur River Basin and Mongolian Plateau such as Boisman, Devilscave, Mongolia_North_N, and Mongolia_East_N, and it also possessed a dominant proportion in the genetic makeup of ancient Mongolians. In modern populations, the pink ancestry was enriched in Tungusic-speaking populations such as Ulchi and Nanai. The deep green ancestry was maximized in indigenous Nganasan in Siberia and prevalent in northern East Asian populations. In addition to the aforementioned

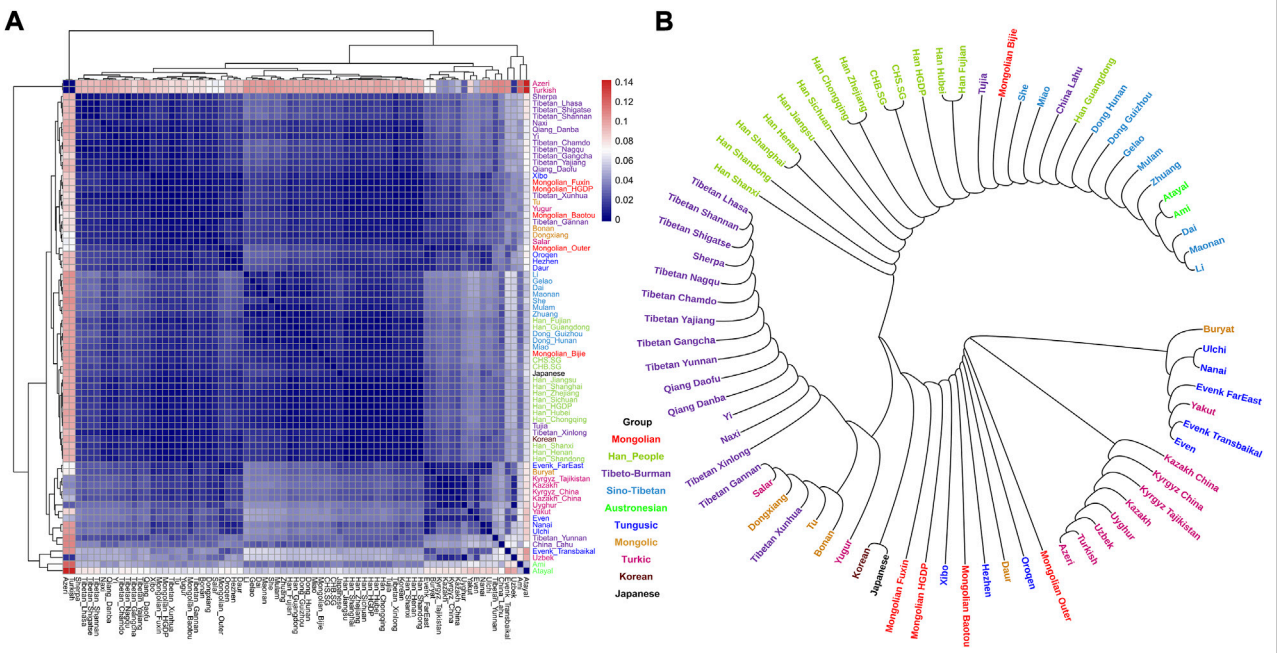


FIGURE 3
Results of pairwise FST among modern East Asian populations. Each population was marked by different colors which was corresponded to their language category. **(A)** Heatmap showed the pairwise FST genetic distance. **(B)** Neighbor-joining (NJ) phylogenetic tree showed the genetic relationship between Fuxin Mongolians and other modern East Asians.

four ancestral components, there were other three ancestral compositions in the genetic makeup of Fuxin Mongolians which were enriched in Western Eurasians, Eurasian Steppe pastoralists, and Central Asian populations, respectively. We found that the genetic profile of Fuxin Mongolians was similar to Hulunbuir Mongolians, but there were more Sino-Tibetan-related ancestry and southern ancestry and less genetic contribution of populations in the Amur River Basin, Siberia, and Mongolian Plateau, relative to Baotou Mongolians and Outer Mongolians. When $K = 7$ (Figure 4), the results were consistent with the aforementioned description, and the YRB farmers and Nepal hunter–gatherers, ARB and MP hunter–gatherers, and southern East Asian populations were the dominant ancestral contributors to Fuxin Mongolians.

f_3- and f_4-statistics

To further test the genetic relationship observed in PCA, FST, and admixture, we performed outgroup-f_3 and admixture-f_3 statistics to examine shared genetic drifts and explore potential admixture signals between Fuxin Mongolians and other reference populations. In outgroup-f_3 (Mongolian_Fuxin, Y; Yoruba) (Figure 5 and Supplementary Tables S5A,B), we found that Fuxin Mongolians showed obvious genetic similarity with Han people, and Fuxin Mongolians also shared more genetic drifts with Korean, Japanese, She, Miao, Tujia, Ulchi, Hezhen, Oroqen, and Daur. When Y represented ancient populations, we found that Fuxin Mongolians shared more genetic drifts with ancient YRB, WLRB, ARB, and MP populations. We next carried out admixture-f_3 (X, Y; Mongolian_Fuxin) (Figure 6; Supplementary Tables S5C,D) to model possible admixture. When X and Y represented modern populations, we found significant negative Z scores ($Z < -3$) between Han people and Tungusic speakers (Hezhen, Oroqen, and Ulchi), Turkic speakers (Uyghur, Turkish, and Altaian), Mongolic speakers (Baotou Mongolian and Buryat), and Western Eurasian populations (Spanish, French, and English). Also, we found negative Z scores when X was ancient YRB populations from the Neolithic to Iron Age and Y represented ancient populations in Eurasian Steppe, Central Asia, and Mongolian Plateau.

We then performed f_4-statistics to explore the asymmetric genetic relationship and gene flow direction between Fuxin Mongolians and other modern/ancient populations in the forms of f_4 (X, Y; Mongolian_Fuxin, Yoruba) and f_4 (X, Mongolian_Fuxin; Y, Yoruba). In the form of f_4 (X, Y; Mongolian_Fuxin, Yoruba) (Supplementary Table S6), we observed significant negative f_4 values when X was Han people and Y represented other modern populations, and it revealed that Fuxin Mongolians shared the most alleles with Han people relative to other modern Eurasian

10.3389/fgene.2022.947758

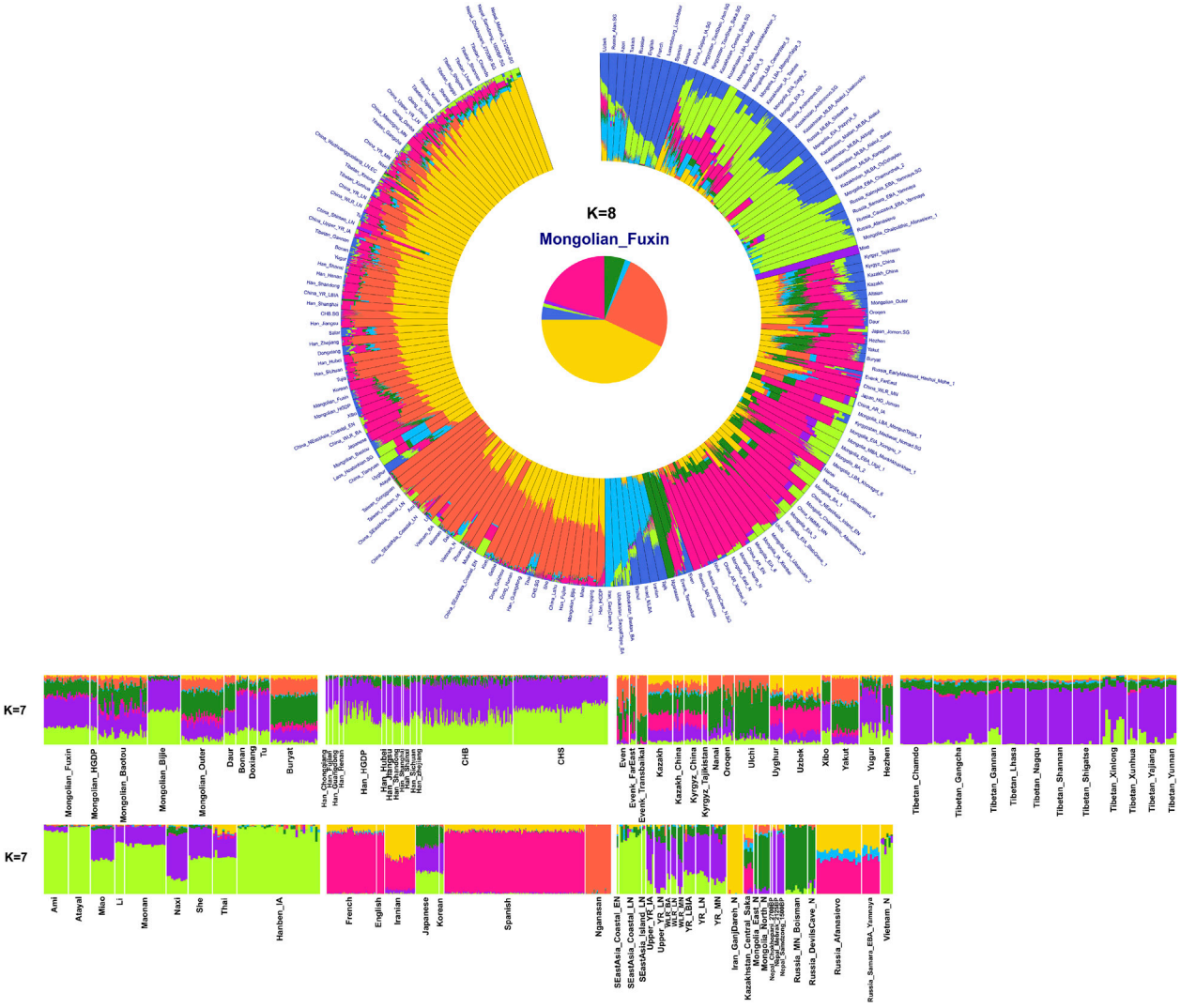


FIGURE 4
Results of model-based admixture clustering analysis showed the ancestral composition of studied Mongolians and reference populations at K = 7 and K = 8.

populations. We also found that Fuxin Mongolians shared more alleles with Hulunbuir Mongolians than Baotou Mongolians, and Daur, Hezhen, Oroqen, Ulchi, and Xibo also shared more alleles with Fuxin Mongolian than Baotou Mongolians. We next performed the f_4-statistic in the form of f_4 (Modern Population1, Mongolian_Fuxin; Modern Population2, Yoruba) (Figure 7; Supplementary Table S7), and the results further indicated that Han, Hulunbuir Mongolian, Xibo, Hezhen, and Daur shared more alleles with Fuxin Mongolian than other modern populations. However, the Han people exhibited a closer genetic relationship with southern East Asian populations, Xibo, Hezhen, and Daur shared more alleles with Tungusic-speaking populations relative to Fuxin Mongolians, and Hulunbuir Mongolians had the closest genetic affinity with Fuxin Mongolians. When X and Y represented ancient populations in the forms of f_4 (X, Y; Mongolian_Fuxin, Yoruba), we found Fuxin Mongolians shared more alleles with ancient YRB, WLRB, ARB, and MP populations (Supplementary Table S6), and the results of f_4-statistic in the form of f_4 (Ancient Population1, Mongolian_Fuxin; Ancient Population2, Yoruba) (Figure 7; Supplementary Table S8) further revealed that there were significant genetic contributions in Fuxin Mongolians from ancient YRB populations related to the Neolithic Yangshao, Miaozigou, and Shimao culture, ancient WLRB populations related to the Hongshan, Lower Xiajiadian, and Upper Xiajiadian culture, Neolithic ARB and MP hunter–gatherers, Bronze-Age Ulaanzukh culture pastoralists (1,448–1,292 calBCE), and Iron-Age Mongolia_EIA_8 individuals (971–830 calBCE).

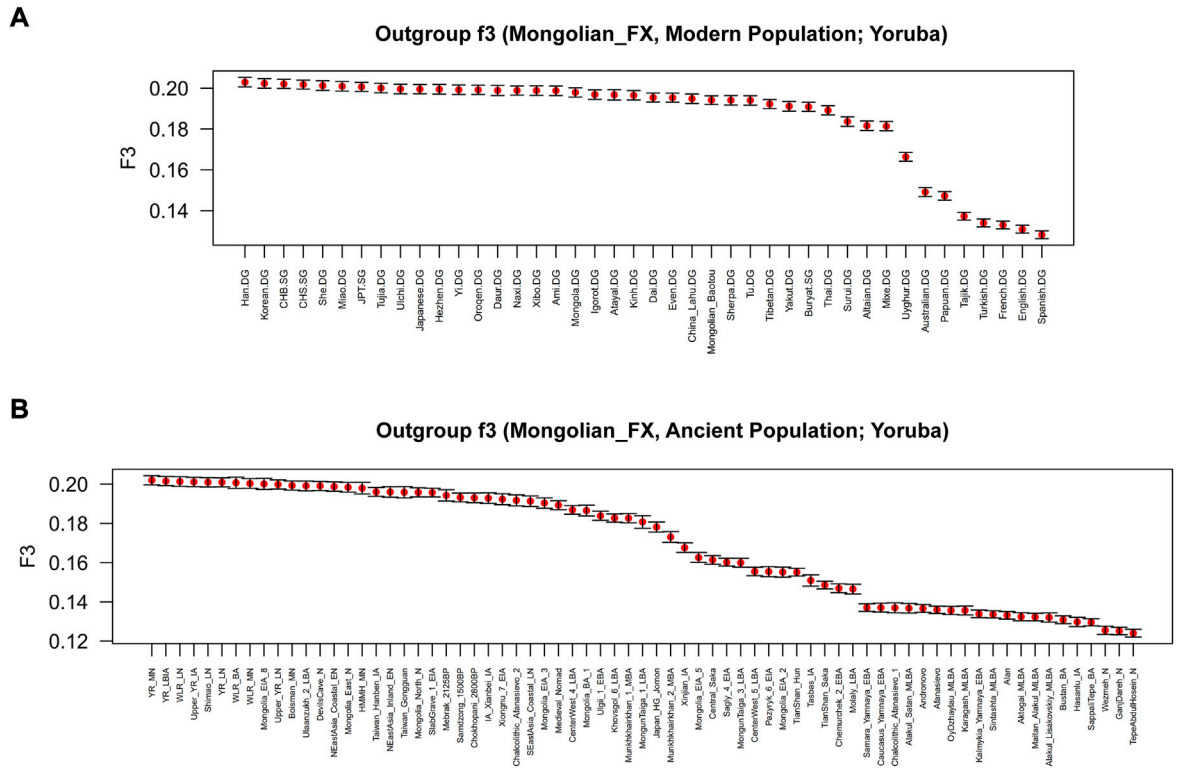


FIGURE 5
Shared genetic drift estimated *via* f_3-statistics of the form f_3 (Mongolian_Fuxin, Y; Yoruba) based on the merged 1240K dataset. **(A)** outgroup-f_3 (Mongolian_Fuxin, Modern population; Yoruba). **(B)** outgroup-f_3 (Mongolian_Fuxin, Ancient population; Yoruba).

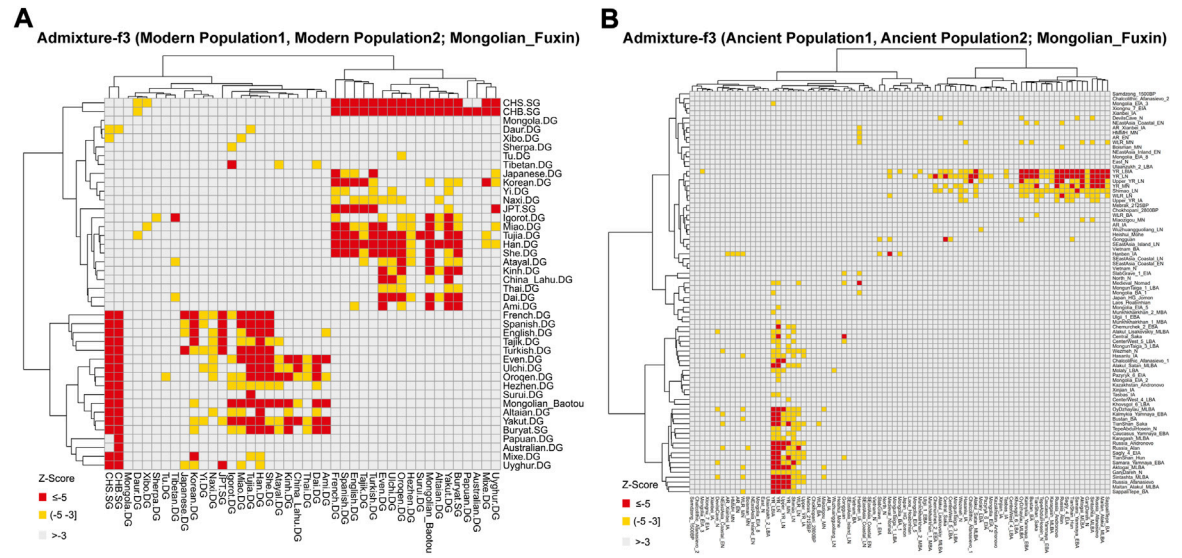


FIGURE 6
Heatmap of the admixture f_3-statistics of the form f_3 (X, Y; Mongolian_Fuxin). The red and yellow color indicated significant admixture signals, and X and Y were the possible ancestral contributors of Fuxin Mongolians. **(A)** Admixture-f_3 (Modern population1, Modern population2; Mongolian_Fuxin). **(B)** Admixture-f_3 (Ancient population1, Ancient population2; Mongolian_Fuxin).

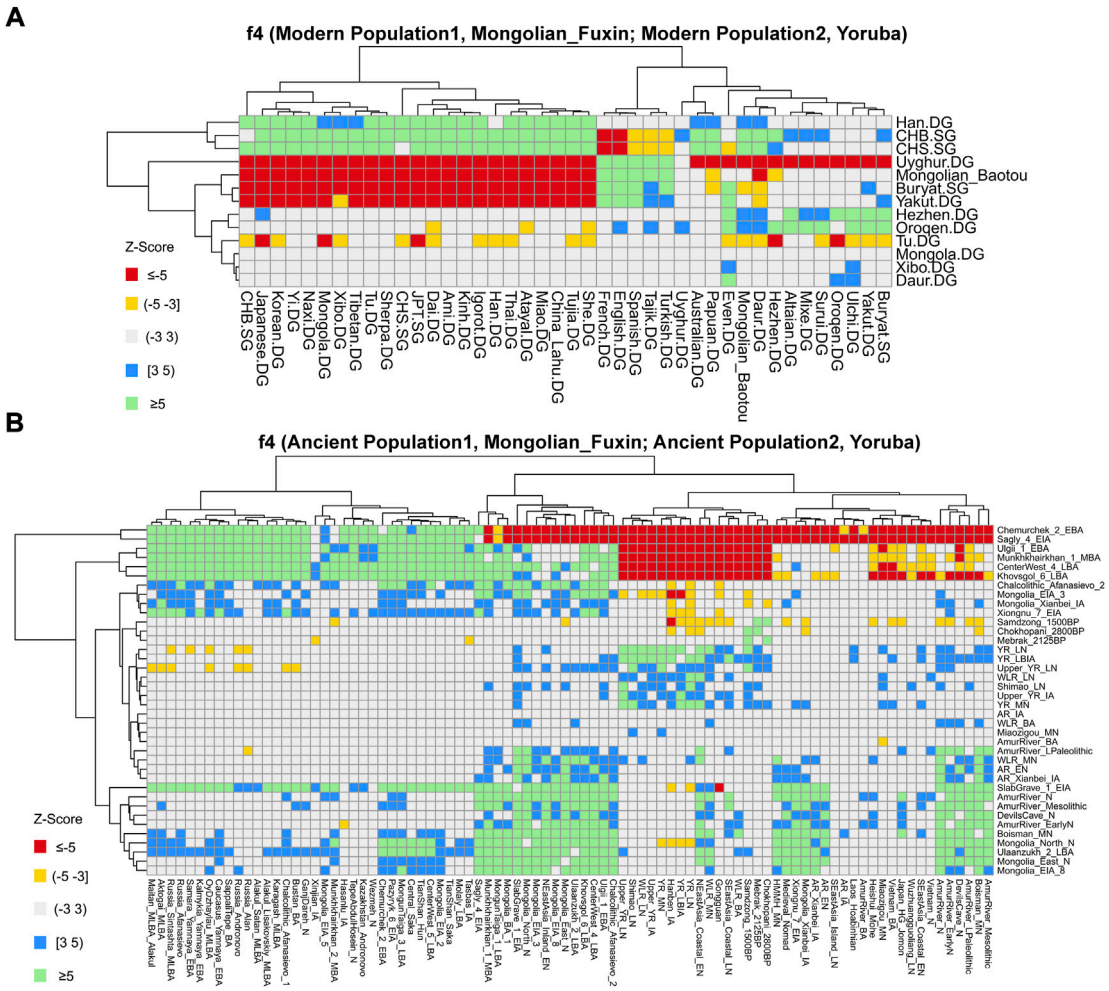


FIGURE 7
Results of f_4-statistics performed in the form of **(A)** f_4 (Modern Population1, Mongolian_Fuxin; Modern Population2, Yoruba); **(B)** f_4 (Ancient Population1, Mongolian_Fuxin; Ancient Population2, Yoruba) exhibited the genetic affinity between Fuxin Mongolians and possible ancestral contributors.

*qp*Wave/*qp*Adm

We applied *qp*Wave/*qp*Adm methods to explore the possible ancestral contributors and estimate their admixture proportions in Fuxin Mongolians. We used millet farmers related to Neolithic Yangshao, Longshan, and Lower Xiajiadian cultures, Neolithic hunter–gatherers in the Amur River Basin and Mongolian Plateau, Iranian Neolithic farmers, Bronze-Age Yamnaya pastoralists, and Bronze-Age Sake peoples in Central Asia as the possible ancestral sources (Figure 8; Supplementary Table S9). We found that the main genetic contributions in Fuxin Mongolians were from ancient populations in the Yellow River Basin, West Liao River Basin, Amur River Basin, and Mongolian Plateau, and there were additional gene flows related to Western Eurasian and Central Asian populations. Also, relative to other Mongolic, Tungusic, and Turkic speakers, Fuxin Mongolians had more millet farmer-related ancestry and ARB hunter–gatherers-related ancestry.

*qp*Graph and TreeMix

We used *qp*Graph and TreeMix methods with gene flow events to further explore possible ancestral sources and potential admixture signals and reconstruct deep phylogenetic structures in Fuxin Mongolians. In the *qp*Graph-based phylogenetic framework, we used Mbuti, Denisovan, Loschbour, GreatAndaman, and Tianyuan to construct the basal model. Neolithic hunter–gatherers in the Mongolian Plateau and Amur River Basin, millet farmers related to Yangshao and Lower Xiajiadian cultures in Yellow River and West River basins, Neolithic Qihe, Iron-Age Hanben, and Bronze-Age

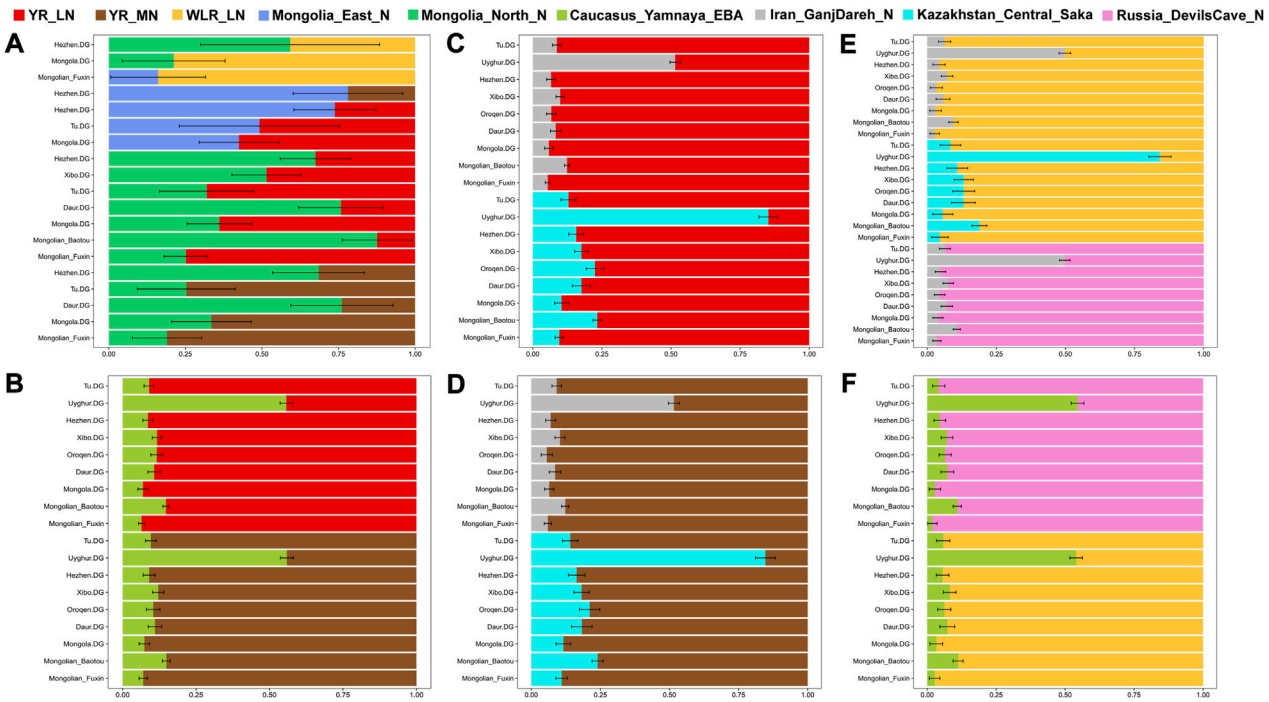


FIGURE 8
*qp*Wave/*qp*Adm results showed the detailed ancestral composition in the best-fitted two-way admixture model. **(A~F)** Different models showed discrepant ancestral composition among Mongolic, Tungusic and Turkic speakers.

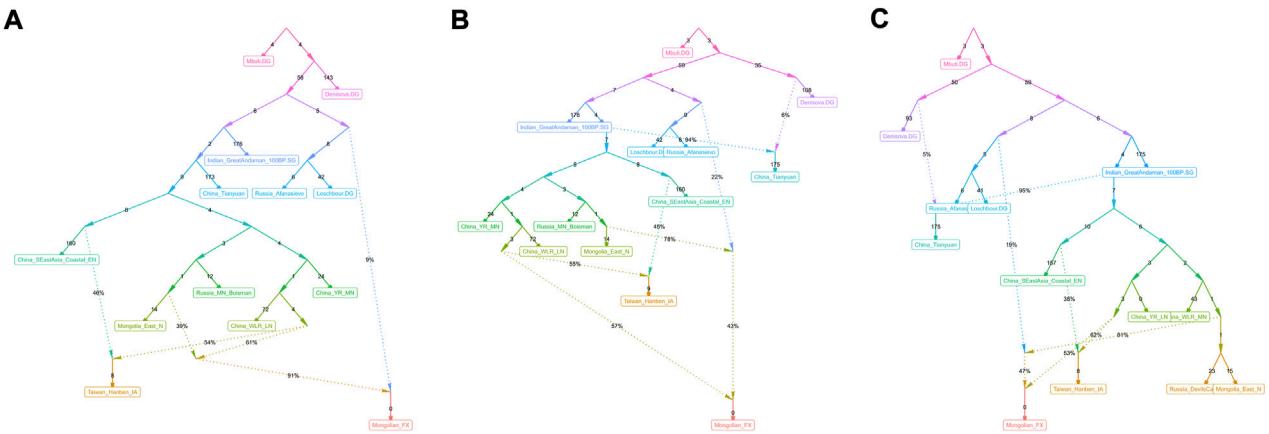


FIGURE 9
*qp*Graph-based phylogeny showed the genetic formation of modern Mongolians. **(A~C)** Different genetic admixture models showed genetic formation of Fuxin Mongolians.

Afanasievo pastoralists were used as different ancestral source proxies (Figure 9). We found that Fuxin Mongolians could be modeled as the mixture of Neolithic millet farmer-related ancestry (53%–57%), Neolithic MP and ARB hunter–gatherer-related ancestry, and Western Eurasian-related ancestry (43%–47%), and we also found that Western Eurasian-related ancestry possessed 9% in the gene pool of Fuxin Mongolians. Our *qp*Graph models were consistent with the *qp*Adm results and further indicated that millet farmers and ARB and MP hunter–gatherers were the dominant ancestral sources and

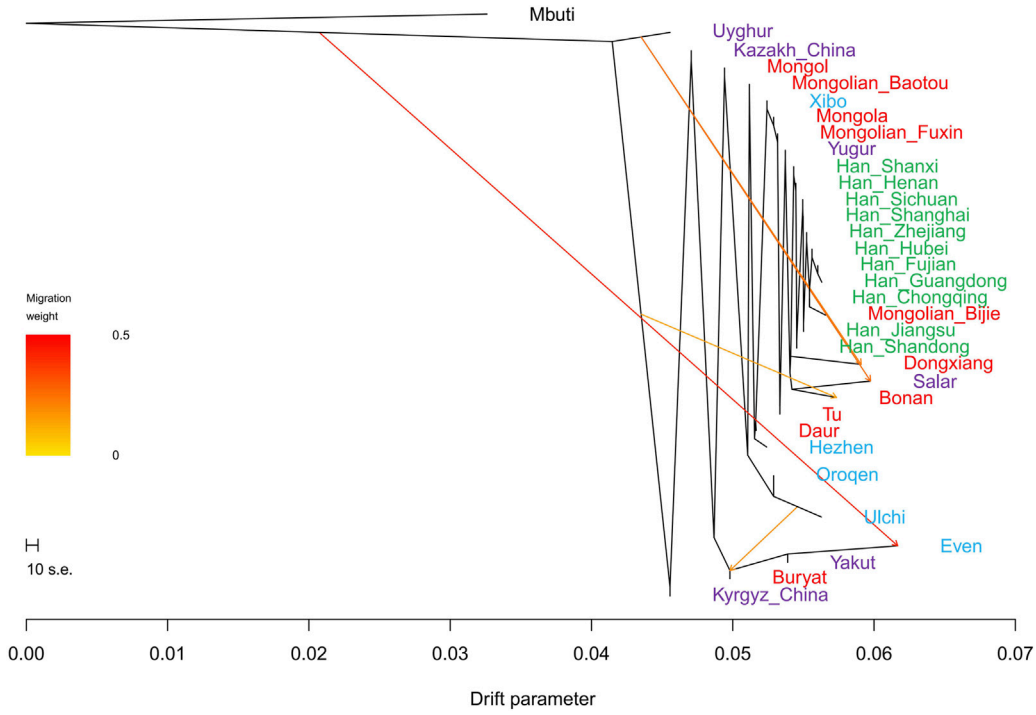


FIGURE 10
TreeMix-based maximum-likelihood phylogenetic tree with four migration events showed the population relationship among Han people and Turkic, Tungusic, and Mongolic speakers. Different colors represent different language categories.

played an indispensable role in the genetic formation of modern Mongolians, and there was a limited genetic influence of Western Eurasian populations in Fuxin Mongolians.

In the TreeMix analysis (Figure 10), there was no significant gene flow event that could be found in Fuxin Mongolians. We found that Fuxin Mongolians clustered together with Mongolians from Hulunbuir and Baotou, Xibo, and Yugur. There was an obvious Han people cline, and Bijie Mongolians clustered with the Southern Han people. These results indicated that although there was an obvious genetic influence of the Han people on Fuxin Mongolians, and Fuxin Mongolians showed more significant genetic similarity with Northern Chinese Mongolians, which revealed genetic differences between Mongolians and Han people.

ALDER and uniparental haplogroups

We next applied the ALDER method based on weighted linkage disequilibrium statistics to estimate the admixture time and explore possible admixture signals. There were multiple sources of admixture signals such as Han, Tungusic speakers, Mongolic speakers, Turkic speakers, and populations who harbored Western Eurasian-related ancestry. The results showed that the obvious eastern–western Eurasian population interaction and admixture occurred during a historic period (~600–~1,300 years ago) which was associated with the prosperity of the Silk Road and the westward expansion of the Mongol Empire (Supplementary Table S10). The time of admixture signals about the Han people could be dated back to ~400–~1,300 years ago which is approximately from the Tang dynasty to the Ming dynasty. The time of admixture signals about Western Eurasians could be dated back to about ~600–~1,500 years ago. Also, the Turkic- and Tungusic-related ancestries flowed into the gene pool of Fuxin Mongolians from ~500 to ~1,800 years ago.

In this study, we successfully identified paternal and maternal lineages in Fuxin Mongolians; 13 paternal haplogroups and 36 maternal haplogroups are listed in the Supplementary Material (Supplementary Table S11). We found that D5a2a, G1c, and R9c1a were the most frequent maternal haplogroups, and D1a1a1a1a2a and N1a2b1b1 were the dominant paternal haplogroups.

Discussion

In this study, we performed a comprehensive population genetic analysis to explore the genetic structure and admixture history of the Mongolian population in Liaoning Province based

on newly generated genome-wide SNP data. The results of the descriptive analysis (PCA, FST, and admixture) indicated that the genetic structure was consistent with the language category and geographical distribution, and populations from the same language group or adjoining geographic position had a close genetic affinity with East Asia. Fuxin Mongolians showed genetic similarity with the Mongolian population in Inner Mongolia (Hulunbuir and Baotou) and other Mongolic speakers such as Daur, Tu, and Bonan. Also, there were close genetic relationships between Fuxin Mongolians and Han people, Tungusic speakers (Hezhen and Xibo), and other geographically adjoining populations. Outgroup-f_3 statistics further supported the aforementioned conclusions. We found that Fuxin Mongolians shared more genetic drifts with the Han people than other East Asians, and f_4-statistics results further indicated the genetic similarity between Fuxin Mongolian and Han people, with insignificant Z-scores in the form of f_4 (Han, Mongolian_Fuxin; Modern Eurasian population, Yoruba). Also, Fuxin Mongolians also shared alleles with Tungusic and Mongolic speakers which are supported by the results of Outgroup-f_3 (Mongolian_Fuxin, Modern Population; Yoruba) and f_4 (Hezhen/Xibo/Mongolia/Daur, Mongolian_Fuxin; Modern population, Yoruba). The results of admixture-f_3 (Tungusic/Mongolic/Turkic speakers, CHS/CHB; Mongolian_Fuxin) which have a significant Z-score further indicated that Han people and Altaic speakers played an indispensable role in the formation of genetic makeup in Fuxin Mongolians. In addition, we also found the genetic contribution of the Western Eurasian population in Fuxin Mongolians. The admixture-f_3 results showed that Fuxin Mongolians could be modeled as the mixture between Han people and Western Eurasian populations, and the ALDER-based admixture time revealed that the eastern–western Eurasian population admixture occurred during ~600–~1,300 years ago which is the period from the Tang dynasty to the Ming dynasty. The Silk Road and Maritime Silk Route were further developed during the Tang Dynasty; the Mongol Empire rose and expanded westward in the 13th century, and ancient Chinese navigators explored the ocean during the Ming dynasty. These events promoted prosperous economic, political, and cultural interactions and genetic exchanges between western and eastern Eurasia (Yao et al., 2004; Jeong et al., 2020). Also, due to special geographical location, Mongolians have become witnessers and promoters of the exchange and admixture of eastern–western Eurasia. The mixed pattern of eastern–western Eurasia is also revealed by mitochondrial and Y-chromosomal haplogroups. B4, D4, and R9 were the main maternal lineages, and O2a1, O2a2, and D1a were the dominant paternal lineages. These haplogroups have a high frequency in East Asian populations such as the Han people (Kivisild et al., 2002; Yao et al., 2002; Lang et al., 2019; Wang et al., 2021b), and Western Eurasian-specific haplogroups U4 also existed in Fuxin Mongolians (Melchior et al., 2010; Soares et al., 2010). Also, except East Asian-dominant haplogroups, there were Siberia-related haplogroups such as C7, N1a, N1b, and Q1a (Malyarchuk and Derenko, 2009; Stoneking and Delfin, 2010; Malyarchuk et al., 2011; Wang et al., 2021b), and these haplogroups further indicated the genetic contribution of Northern Asian populations in the genetic makeup of Fuxin Mongolians. We also found there were genetic differences among Mongolians, Fuxin Mongolians had less ARB and MP hunter–gatherer-related ancestry and Western Eurasian-related ancestry relative to Baotou Mongolian and Outer Mongolian, and Bijie Mongolians in Guizhou Province had more Sino-Tibetan-related ancestry and Southern-related ancestry than other Mongolian people. These results further supported the conclusions of previous studies (He et al., 2021; Yang et al., 2021).

Paleogenomic studies demonstrated that ancient Mongolians at different times and geographic regions exhibited disparate genetic profiles. There was a dynamic demographic history in the Mongolian Plateau, and multiple populations contributed to genetic ancestries to shape the genetic differences of ancient Mongolians: the main ancient Northeast Asian ancestry, the ephemeral ancient North Eurasian ancestry, the limited genetic contributions of Western Eurasian Steppe pastoralists and Iranian farmers, and recent genetic influence of the Han people (Jeong et al., 2020; Wang et al., 2021a). Our research indicated that these ancestries were still retained in present-day Fuxin Mongolians. The results of outgroup-f_3 (Mongolian_Fuxin, Ancient population; Yoruba) and f_4 (Ancient population1, Ancient population2; Mongolian_Fuxin, Yoruba) indicated that Fuxin Mongolians shared more genetic drifts and alleles with ancient YRB, WLRB, ARB, and MP populations. There were insignificant Z-scores in the form of f_4 (YRB/WLB/ARB/MP, Mongolian_Fuxin; Ancient population2, Yoruba). The genetic contribution of ancient Mongolians also existed in the Fuxin Mongolian gene pool, Mongolia_8_EIA, Ulaanzukh_2_LBA, SlabGrave_1_EIA, Xianbei_IA, Xiongnu_7_EIA, CenterWest_4_LBA, Ulgii_1_EBA, and Khovsgol_6_LBA shared alleles with Fuxin Mongolians, and the results of f_4-statistics in the form of f_4 (Ancient Mongolian, Mongolian_Fuxin; Ancient population2, Yoruba) revealed that Fuxin Mongolians received more genetic influences of Mongolia_8_EIA and Ulaanzukh_2_LBA. According to historical records, modern Mongolians in Liaoning are mainly descended from two ancient Mongolian tribes, the “Mongolia Zhen” tribe, and the “Harqin” tribe. Fuxin Mongolians belong to the “Mongolia Zhen” tribe which has a history of more than 1,200 years, it originated in the Orkhon River Basin and Selenga River Basin, and this tribe expanded westward into the regions of Central Asia and Xinjiang with the development of strength. From the late 15th century to the early 16th century, the “Mongolia Zhen” tribe migrated to Hetao Plain in the Yellow River Basin and became an alliance with the

"Tumet" tribe. In the 16th century, the "Tumet-Mongolia Zhen" tribe migrated eastward, and they finally settled in the western regions of Liaoning Province. The complex migration and admixture history shaped genetic differences between Fuxin Mongolians and other Mongolians. The admixture-f_3 (Ancient population1, Ancient population2; Mongolian_Fuxin) results exhibited that Fuxin Mongolians could be modeled as the mixture between Neolithic and Iron–Age YRB and WLRB populations and Eurasian Steppe pastoralists, ancient Mongolians, and ancient Iranian. The models of the population mixture based on the *qp*Adm method showed YRB and WLRB millet farmers and ARB and MP hunter–gatherers were the dominant ancestral contributors, and there were additional gene flows related to Eurasian Steppe pastoralists and Iranian farmers. The *qp*Graph-based phylogenetic framework further supported the aforementioned results. The millet farmer-related ancestry was 53%–57%, and the MP and ARB hunter–gatherer-related ancestry and Western Eurasian-related ancestry were 43%–47%. We also found the Western Eurasian-related ancestry possessed 9% in the gene pool of Fuxin Mongolians. Also, the *qp*Adm results also indicated genetic differences between Fuxin Mongolians and other Altaic language speakers, and we could find there were more YRB-, WLRB-, and ARB-related ancestries in Fuxin Mongolians, which suggested that Fuxin Mongolians were more influenced by the expansion of millet farmers and retained more genetic contribution of ARB hunter–gatherers. In general, we found there were dynamic demographic history, complex population admixture, and multiple sources of genetic diversity in Fuxin Mongolians.

Conclusion

In this study, we reported genome-wide SNP data of Fuxin Mongolians in Liaoning Province. We applied typical and advanced population genetic analysis methods [principal component analysis (PCA), Admixture, FST, f_3-statistics, f_4-statistics, *qp*Adm/*qp*Wave, *qp*Graph, ALDER, and TreeMix] to explore genetic structure and admixture history of Fuxin Mongolians. We found that Fuxin Mongolians had a close genetic relationship with the Han people, northern Mongolians, other Mongolic speakers, and Tungusic speakers in East Asia. Also, we found that Neolithic millet farmers in the Yellow River Basin and West Liao River Basin and Neolithic hunter–gatherers in the Mongolian Plateau and Amur River Basin were the dominant ancestral sources, and there were additional gene flows related to Eurasian Steppe pastoralists and Neolithic Iranian farmers in the gene pool of Fuxin Mongolians. These results shed light on dynamic demographic history, complex population admixture, and multiple sources of genetic diversity in Fuxin Mongolians.

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://zenodo.org/record/6839600, doi: 10.5281/zenodo.6839600.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of Jinzhou Medical University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

YW designed this study and reviewed and revised the manuscript. XZ analyzed the data, and XH wrote the manuscript. WL, HZ, and HH collected the samples. XL and TH performed the experiment.

Funding

The project was supported by the Key Laboratory of the Human Phenotype Group in Liaoning Province.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.947758/full#supplementary-material

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Bai, H., Guo, X., Zhang, D., Narisu, N., Bu, J., Jirimutu, J., et al. (2014). The genome of a Mongolian individual reveals the genetic imprints of Mongolians on modern human populations. *Genome Biol. Evol.* 6 (12), 3122–3136. doi:10.1093/gbe/evu242

Bai, H., Guo, X., Narisu, N., Lan, T., Wu, Q., Xing, Y., et al. (2018). Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. *Nat. Genet.* 50 (12), 1696–1704. doi:10.1038/s41588-018-0250-5

Chen, J., He, G., Ren, Z., Wang, Q., Liu, Y., Zhang, H., et al. (2021). Genomic insights into the admixture history of mongolic- and tungusic-speaking populations from southwestern east Asia. *Front. Genet.* 12, 685285. doi:10.3389/fgene.2021.685285

de Barros Damgaard, P., Martiniano, R., Kamm, J., Moreno-Mayar, J. V., Kroonen, G., Peyrot, M., et al. (2018). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* 360 (6396), eaar7711. doi:10.1126/science.aar7711

Dulik, M. C., Osipova, L. P., and Schurr, T. G. (2011). Y-chromosome variation in Altaian Kazakhs reveals a common paternal gene pool for Kazakhs and the influence of Mongolian expansions. *PLoS One* 6 (3), e17548. doi:10.1371/journal.pone.0017548

Feng, Q., Lu, Y., Ni, X., Yuan, K., Yang, Y., Yang, X., et al. (2017). Genetic history of xinjiang's uyghurs suggests Bronze age multiple-way contacts in Eurasia. *Mol. Biol. Evol.* 34 (10), 2572–2582. doi:10.1093/molbev/msx177

Feng, Q., Lu, D., and Xu, S. (2018). AncestryPainter: A graphic program for displaying ancestry composition of populations and individuals. *Genomics Proteomics Bioinforma.* 16 (5), 382–385. doi:10.1016/j.gpb.2018.05.002

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522 (7555), 207–211. doi:10.1038/nature14317

He, G., Wang, M., Zou, X., Yeh, H.-Y., Liu, C., Liu, C., et al. (2021). Extensive ethnolinguistic diversity at the crossroads of North China and South Siberia reflects multiple sources of genetic diversity. *J. Syst. Evol.* doi:10.1111/jse.12827

Helsinki, W. M. A. D. o. (2001). World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bull. World Health Organ.* 79 (4), 373–374.

Jeong, C., Wang, K., Wilkin, S., Taylor, W. T. T., Miller, B. K., Bemmann, J. H., et al. (2020). A dynamic 6, 000-year genetic history of eurasia's eastern steppe. *Cell* 183 (4), 890–904. doi:10.1016/j.cell.2020.10.015

Kivisild, T., Tolk, H. V., Parik, J., Wang, Y., Papiha, S. S., Bandelt, H. J., et al. (2002). The emerging limbs and twigs of the East Asian mtDNA tree. *Mol. Biol. Evol.* 19 (10), 1737–1751. doi:10.1093/oxfordjournals.molbev.a003996

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35 (6), 1547–1549. doi:10.1093/molbev/msy096

Lang, M., Liu, H., Song, F., Qiao, X., Ye, Y., Ren, H., et al. (2019). Forensic characteristics and genetic analysis of both 27 Y-STRs and 143 Y-SNPs in Eastern Han Chinese population. *Forensic Sci. Int. Genet.* 42, e13–e20. doi:10.1016/j.fsigen.2019.07.011

Loh, P. R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., et al. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193 (4), 1233–1254. doi:10.1534/genetics.112.147330

Ma, B., Chen, J., Yang, X., Bai, J., Ouyang, S., Mo, X., et al. (2021). The genetic structure and east-west population admixture in Northwest China inferred from genome-wide Array genotyping. *Front. Genet.* 12, 795570. doi:10.3389/fgene.2021.795570

Malyarchuk, B., and Derenko, M. (2009). On the origin of Y-chromosome haplogroup N1b. *Eur. J. Hum. Genet.* 17 (12), 1540–1541. author reply 1541-1543. doi:10.1038/ejhg.2009.100

Malyarchuk, B., Derenko, M., Denisova, G., Maksimov, A., Wozniak, M., Grzybowski, T., et al. (2011). Ancient links between Siberians and Native Americans revealed by subtyping the Y chromosome haplogroup Q1a. *J. Hum. Genet.* 56 (8), 583–588. doi:10.1038/jhg.2011.64

Mao, X., Zhang, H., Qiao, S., Liu, Y., Chang, F., Xie, P., et al. (2021). The deep population history of northern East Asia from the late pleistocene to the holocene. *Cell* 184 (12), 32563256–32563266.e13. doi:10.1016/j.cell.2021.04.040

Melchior, L., Lynnerup, N., Siegismund, H. R., Kivisild, T., and Dissing, J. (2010). Genetic diversity among ancient Nordic populations. *PLoS One* 5 (7), e11898. doi:10.1371/journal.pone.0011898

Ning, C., Wang, C. C., Gao, S., Yang, Y., Zhang, X., Wu, X., et al. (2019). Ancient genomes reveal yamnaya-related ancestry and a potential source of indo-European speakers in Iron age tianshan. *Curr. Biol.* 29 (15), 2526–2532.e4. doi:10.1016/j.cub.2019.06.044

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2 (12), e190. doi:10.1371/journal.pgen.0020190

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* 192 (3), 1065–1093. doi:10.1534/genetics.112.145037

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8 (11), e1002967. doi:10.1371/journal.pgen.1002967

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 (3), 559–575. doi:10.1086/519795

Robbeets, M., Bouckaert, R., Conte, M., Savelyev, A., Li, T., An, D. I., et al. (2021). Triangulation supports agricultural spread of the Transeurasian languages. *Nature* 599 (7886), 616–621. doi:10.1038/s41586-021-04108-8

Sagart, L., Jacques, G., Lai, Y., Ryder, R. J., Thouzeau, V., Greenhill, S. J., et al. (2019). Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proc. Natl. Acad. Sci. U. S. A.* 116 (21), 10317–10322. doi:10.1073/pnas.1817972116

Soares, P., Achilli, A., Semino, O., Davies, W., Macaulay, V., Bandelt, H. J., et al. (2010). The archaeogenetics of Europe. *Curr. Biol.* 20 (4), R174–R183. doi:10.1016/j.cub.2009.11.054

Stoneking, M., and Delfin, F. (2010). The human genetic history of east Asia: weaving a complex tapestry. *Curr. Biol.* 20 (4), R188–R193. doi:10.1016/j.cub.2009.11.052

Wang, C. C., Yeh, H. Y., Popov, A. N., Zhang, H. Q., Matsumura, H., Sirak, K., et al. (2021a). Genomic insights into the formation of human populations in East Asia. *Nature* 591 (7850), 413–419. doi:10.1038/s41586-021-03336-2

Wang, M., He, G., Zou, X., Liu, J., Ye, Z., Ming, T., et al. (2021b). Genetic insights into the paternal admixture history of Chinese Mongolians via high-resolution customized Y-SNP SNaPshot panels. *Forensic Sci. Int. Genet.* 54, 102565. doi:10.1016/j.fsigen.2021.102565

Wang, W., Ding, M., Gardner, J. D., Wang, Y., Miao, B., Guo, W., et al. (2021c). Ancient Xinjiang mitogenomes reveal intense admixture with high genetic diversity. *Sci. Adv.* 7 (14), eabd6690. doi:10.1126/sciadv.abd6690

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88 (1), 76–82. doi:10.1016/j.ajhg.2010.11.011

Yang, X., Sarengaowa, He, G., Guo, J., Zhu, K., Ma, H., et al. (2021). Genomic insights into the genetic structure and natural selection of Mongolians. *Front. Genet.* 12, 735786. doi:10.3389/fgene.2021.735786

Yao, Y. G., Kong, Q. P., Bandelt, H. J., Kivisild, T., and Zhang, Y. P. (2002). Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am. J. Hum. Genet.* 70 (3), 635–651. doi:10.1086/338999

Yao, Y. G., Kong, Q. P., Wang, C. Y., Zhu, C. L., and Zhang, Y. P. (2004). Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in China. *Mol. Biol. Evol.* 21 (12), 2265–2280. doi:10.1093/molbev/msh238

Zhang, M., Yan, S., Pan, W., and Jin, L. (2019). Phylogenetic evidence for Sino-Tibetan origin in northern China in the late neolithic. *Nature* 569 (7754), 112–115. doi:10.1038/s41586-019-1153-z

Zhang, F., Ning, C., Scott, A., Fu, Q., Bjørn, R., Li, W., et al. (2021a). The genomic origins of the Bronze age tarim basin mummies. *Nature* 599 (7884), 256–261. doi:10.1038/s41586-021-04052-7

Zhang, X., He, G., Li, W., Wang, Y., Li, X., Chen, Y., et al. (2021b). Genomic insight into the population admixture history of tungusic-speaking Manchu people in Northeast China. *Front. Genet.* 12, 754492. doi:10.3389/fgene.2021.754492

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 21 October 2022
DOI 10.3389/fevo.2022.985933

Check for updates

OPEN ACCESS

EDITED BY
Guanglin He,
Sichuan University, China

REVIEWED BY
Shaoqing Wen,
Fudan University, China
Xiling Liu,
Academy of Forensic Science, China

*CORRESPONDENCE
Yicong Wang
wangyicong@genomics.cn
Jiang Huang
mmm_hj@126.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 04 July 2022
ACCEPTED 15 September 2022
PUBLISHED 21 October 2022

CITATION
Jin X, Ren Z, Zhang H, Wang Q, Liu Y, Ji J, Yang M, Zhang H, Hu W, Wang N, Wang Y and Huang J (2022) Development and forensic efficiency evaluations of a novel multiplex amplification panel of 17 Multi-InDel loci on the X chromosome. *Front. Ecol. Evol.* 10:985933. doi: 10.3389/fevo.2022.985933

COPYRIGHT
© 2022 Jin, Ren, Zhang, Wang, Liu, Ji, Yang, Zhang, Hu, Wang, Wang 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.

Development and forensic efficiency evaluations of a novel multiplex amplification panel of 17 Multi-InDel loci on the X chromosome

Xiaoye Jin[1,2†], Zheng Ren[1†], Hongling Zhang[1], Qiyan Wang[1], Yubo Liu[1], Jingyan Ji[1], Meiqing Yang[1], Han Zhang[1], Wen Hu[3], Ning Wang[3], Yicong Wang[3*] and Jiang Huang[1*]

[1]Department of Forensic Medicine, Guizhou Medical University, Guiyang, China, [2]Shanghai Key Lab of Forensic Medicine, Key Lab of Forensic Science, Ministry of Justice, Shanghai, China, [3]BGI-Shenzhen, Shenzhen, China

Multi-InDel, as the novel genetic markers, showed great potential in forensic research. Whereas, most scholars mainly focused on autosomal Multi-InDels, which might provide limited genetic information in some complex kinship cases. In this study, we selected 17 Multi-InDels on the X chromosome and developed a multiplex amplification panel based on the next-generation sequencing (NGS) technology. Genetic distributions of these 17 loci in Beijing Han, Chinese Southern Han, and the studied Guizhou Han populations revealed that most loci showed relatively high forensic application values in these Han populations. In addition, more allelic variations of some loci were observed in the Guizhou Han than those in Beijing Han and Southern Han populations. Pairwise F_{ST} values, multi-dimensional analysis, and phylogenetic tree of different continental populations showed that selected 17 loci generally could differentiate African, European, East Asian, and South Asian populations. To sum up, the developed panel in this study is not only viewed as the high-efficient supplementary tool for forensic individual identification and paternity analysis, but it is also beneficial for inferring biogeographical origins of different continental populations.

KEYWORDS

Multi-InDel, X chromosome, NGS, forensic research, biogeographical origin inference

Introduction

Insertion/deletion (InDel) polymorphisms were the third genetic markers that displayed insertion or deletion of different deoxyribonucleic acid (DNA) fragments in the genome. As InDels were firstly identified by Weber et al. (2002), they have been paid a large number of attention by forensic researchers and human geneticists because they had low mutation rates and showed wide distributions in the human genome. Till now, a set of multiplex amplification panels have been developed for various forensic purposes (Chen L. et al., 2021; Jin et al., 2021; Zhang et al., 2021; Fan et al., 2022a). However, InDels commonly showed bi-allelic variations that result in their low genetic diversities in comparison with short tandem repeats (STRs). Therefore, there were defects of these bi-allelic InDels in forensic application. On the one hand, they were not conductive to mixture deconvolution; on the other hand, more InDels need to be incorporated into a multiplex amplification panel if forensic researchers wish to obtain comparable forensic efficiency of commonly used STR kits. To avoid the aforementioned shortcomings, a novel genetic marker (Multi-InDel) was proposed by Huang et al. (2014), formed by some closely linked InDels in the small molecular interval (200 bp). Multi-InDel could exhibit multiple allelic variations in the population and showed great application values in forensic individual identification and ancestry origin inferences (Sun et al., 2019; Qu et al., 2020). Nonetheless, the extant research mainly focused on Multi-InDel loci on the autosomes. These loci might provide relatively limited genetic information in some complex kinship cases like deficiency paternity cases. Hereto, genetic markers on the allosome may get more valuable insights into these complex kinship testing.

Genetic markers on the X chromosome possessed unique genetic patterns: for males, they were only transmitted from father to daughter; on the contrary, they showed a similar inheritance mode with genetic markers on the autosomes for females (Gomes et al., 2020). The specific genetic characteristics made genetic markers on the X chromosome play crucial roles in some complex kinship cases like grandparent-grandchild comparisons, half-sisters testing, paternity analyses in incest cases, and so on (Szibor, 2007). Currently, forensic workers generally employed X-STRs to paternity analysis. However, relatively high mutation rates of X-STRs might exert adverse effects on complex paternity analyses. Therefore, some X-InDel panels have been developed for forensic parentage testing (Zhang et al., 2015; Caputo et al., 2017; Chen L. et al., 2021) because InDels possess some advantageous features in comparison to STRs. In the subsequent studies, Fan et al. (2015, 2016) chose 13 Multi-InDels on the X chromosome and evaluated their forensic application values; obtained results revealed that these 13 loci were not only useful for personal identification and kinship testing, but they could also achieve ancestry resolutions of different continental populations. Even so, they proposed that more Multi-InDels need to select to obtain higher forensic effectiveness and discern genetic substructure of Chinese populations.

In the current study, we selected 17 Multi-InDels on the X chromosome. Next, genetic distributions and forensic efficiency evaluations of these loci in Chinese Beijing Han (CHB) and Southern Han (CHS) populations were conducted based on the previously reported data (Donnelly et al., 2015). Thirdly, a multiplex amplification system of these 17 loci was developed by the NGS technology and was used to genotype 217 Guizhou Han individuals in order to further evaluate its forensic application values. Finally, population genetic analyses of different continental populations were performed to assess the power of these loci to discriminate these continental populations.

Materials and methods

Sample information

We collected bloodstain samples of 217 unrelated healthy Guizhou Han individuals who have lived in Guizhou for at least three generations. All individual participants in this study provided their written informed consent. In addition, genetic data of selected loci in different continental populations were downloaded from 1000 Genome Project Phase III (Donnelly et al., 2015) to evaluate their genetic distributions. Only males were assembled because we could directly determine haplotype information of different InDels in the short DNA region. The general information of 26 reference population like CHB, CHS, and so on was given in **Supplementary Table 1**. This study was performed in line with the guidelines of the Guizhou Medical University ethics committee and warranted by the Guizhou Medical University ethics committee.

Selection of multi-insertion/deletion loci on the X chromosome

Multi-InDel loci on the X chromosome were selected based on previous reports (Donnelly et al., 2015; Fan et al., 2015) according to the following criteria: (1). physical distances between InDels on the short molecular interval were less than 200 bp; (2). fragment length of insertion/deletion was less than 30 bp; (3). the minor allelic frequencies of InDels in Chinese Han populations were greater than 0.1; (4). InDels on the short molecular interval (200 bp) showed different allelic frequency

distributions. (5). Multi-InDel loci on the X chromosome were 1 Mb apart from each other.

Primer design and multiplex polymerase chain reaction of selected X-chromosomal multi-insertion/deletion

Primer sequences of selected X-chromosomal Multi-InDels were designed by the ATOPlex online tool.[1] Detailed primer information was presented in **Supplementary Table 2**.

We conducted two steps of polymerase chain reaction (PCR) to construct the sequencing library. Firstly, one 1.2 mm bloodstain card from each individual was obtained by the punch and pro-processed with 25 μL Clean Buffer at 60°C for 10 min. After removing 16.5 μL Clean Buffer, we added 16.5 μL PCR cocktail including 12.5 μL PCR Enzyme Mix and 4 μL PCR Primer Pool. Next, the PCR mixture was conducted to thermal cycle reaction on the 9700 Thermal cycle PCR System (Thermo Fisher Scientific, MA, USA). The detailed reaction conditions were listed as follows: an initial pre-denaturation at 98°C for 5 min; 14 cycles of 98°C for 15 s, 64°C for 1 min, 60°C for 1 min, and 72°C for 30 s; 72°C for 2 min. And then we purified the amplified product by the MagBead DNA Purification kit (CWBIO, Beijing, China) according to the kit's instructions. We used 6.5 μL TE Buffer to wash the purified DNA off and added second round PCR reagents comprising 12.5 μL PCR Enzyme Mix, 2 μL PCR Block, 2 μL PCR Dual Barcode Primer F, and 2 μL PCR Dual Barcode Primer R. The 9700 Thermal cycle PCR System was also used to conduct thermal cycle reaction according to the below parameters: 98°C for 5 min; 16 cycles of 98°C for 15 s, 64°C for 30 s, 60°C for 30 s, and 72°C for 30 s; 72°C for 2 min. Finally, the amplified product was also purified by the same method mentioned above.

Deoxyribonucleic acid sequencing and data analysis

DNA library was quantitated by the Quant-iT^{TM} PicoGreen dsDNA Assay kit (Thermo Fisher Scientific, MA, USA). Based on DNA quantitation results, DNA libraries of all samples were mixed into a well (30 ng). We constructed the DNA Nano Ball (DNB) by the DNBSEQ OneStep DNB Make Reagent kit (MGI, Shenzhen, China) following its recommended specifications (Li et al., 2021; Fan et al., 2022b).

MGISEQ-2000RS sequencing platform was used to perform DNA sequencing of DNB with the mode of SE400 + 10 + 10. Raw data were processed to filter low-quality sequences and reads with multiple N bases by the Soapnuke software (Chen et al., 2018). Clean data were aligned to hs37d5 reference sequence by the Burrows-Wheeler Alignment software (Li and Durbin, 2009). Next, we used GATK HaplotypeCaller (Schmidt et al., 2010) with the -ERC GVCF parameter to generate GVCF files, and then used GATK CombineGVCFs with default parameters to combine the potential variants and joint genotyping. The mutations with depth more than 100 and frequency more than 0.2 were allowed. Analytical threshold and interpretation threshold were depth of coverage (DoC) × 1.5% and DoC × 4.5% if DoC were greater than 650; if not, they were 15 × and 30 × , respectively.

Statistical analyses

Allelic frequencies and forensic related parameters of 17 Multi-InDels in CHB, CHS, and studied Guizhou Han populations were calculated by the StatsX v2.0 software (Lang et al., 2019). Further, the number of alleles and forensic parameters of these 17 loci in three Han populations were visually shown by *R* v4.1.0 and TBtools v1.0 software (Chen et al., 2020). Linkage disequilibrium analyses of 17 Multi-InDels in the studied Guizhou Han population were conducted by the STRAF online tool (Gouy and Zieger, 2017). Next, forensic parameters of two linkage groups were also estimated by the StatsX. Fixation index (F_{ST}) values of 17 Multi-InDels between different continental populations were calculated by the Arlequin v3.5.2.2 software (Excoffier and Lischer, 2010). Besides, we also estimated pairwise F_{ST} values of different continental populations by the Arlequin. Based on pairwise F_{ST} values, multi-dimensional analysis and phylogenetic tree of these continental populations were conducted by SPSS v18 and MEGA v11.0.9 (Kumar et al., 2018) software, respectively.

Results and discussion

Loci information

In this study, we screened 18 X chromosomal Multi-InDel loci based on established selection conditions. Even so, we found that two Multi-InDel loci (rs112111922_rs35401470 and rs201707878_rs71943052) showed complex sequences in their neighboring regions, which were hard to determine their genotype. In addition, rs58222634 locus of rs58222634_rs200362185 displayed a large number of di-nucleotide repeats, which brought about many noise reads. Thus, the rs58222634 locus and two Multi-InDels were eliminated from the developed multiplex panel. Moreover, we also added a multi-allelic InDel (rs78613336) locus into

1 https://atoplex.mgi-tech.com/

the developed system. Finally, we successfully developed the multiplex amplification system of 17 Multi-InDels. Loci information and physical locations of these loci were given in Table 1 and Figure 1A.

Forensic efficiency evaluations of selected 17 loci in training data

Based on the selected 17 loci on the X chromosome, we firstly assessed the number of alleles and expected heterozygosities (He) of these loci in training data (CHB and CHS), as shown in Figure 1B. Not surprisingly, the M8 locus showed the lowest He and the least number of alleles in CHB and CHS populations since it only included one bi-allelic InDel (rs200362185). In addition, the rs59241399 locus of M12 was not reported in 1000 Genome Project Phase III. Therefore, we also observed that M12 locus showed two allelic variations in CHB and CHS populations. For the remaining 15 loci, more than two alleles could be observed in these loci, especially for M1 and M10. More importantly, most loci displayed relatively high He values (>0.5) in CHB and CHS populations, indicating they showed relatively high genetic diversities in these two Han populations.

Forensic-related parameters of these 17 loci in CHB and CHS populations were also estimated, as given in Figure 1C and Supplementary Table 3. In the CHB population, polymorphic information content (PIC) values of most loci were greater than 0.5 except M8, M12, and M17 loci. Power of discrimination in male (PD_M) and female (PD_F) for these 17 loci ranged from 0.4055 (M8) to 0.6655 (M6) and 0.5643 (M8) to 0.8231 (M1), respectively. Mean exclusion chance of these 17 loci in deficiency cases according to Krüger et al. (1968) (MEC_Krüger) and standard trios according to Kishida et al. (1997) (MEC_Kishida) distributed from 0.1616 (M8) to 0.3961 (M1) and 0.3233 (M8) to 0.5939 (M1), respectively. Further, we also calculated MEC in duos (MEC_Desmarais_duo) and trios (MEC_Desmarais) according to Desmarais's study (Desmarais et al., 1998). The average MEC_Desmarais_duo and MEC_Desmarais values of 17 loci were 0.3927 and 0.5373, respectively. Cumulative PD_M, PD_F, MEC_Krüger, MEC_Kishida, MEC_Desmarais_duo, and MEC_Desmarais of these 17 loci in the CHB population were 0.9999999138, 0.99999999999236, 0.99897, 0.999998, 0.99981, and 0.999998, respectively. In the CHS population, similar results could be discerned from these forensic parameters. The average PIC, PD_M, PD_F, MEC_Krüger, MEC_Kishida, MEC_Desmarais_duo, and MEC_Desmarais values of these 17 loci were 0.5394, 0.6097, 0.7692, 0.3340, 0.5394, 0.3955, and 0.5394, respectively. The M1 locus showed the highest forensic application values; whereas, the M8 locus demonstrated the lowest forensic efficiencies. Cumulative PD_M, PD_F, MEC_Krüger, MEC_Kishida, MEC_Desmarais_duo, and MEC_Desmarais values of these 17 loci in the CHS population were 0.9999999212, 0.99999999999377, 0.99909, 0.99999856, 0.999833, and 0.99999856, respectively. In a nutshell, we proposed that these 17 loci could be used for forensic individual identification and paternity analyses in CHB

TABLE 1 General information of 17 selected Multi-InDel loci on the X chromosome.

Loci	InDel	Position	Reference allele	MAF	Loci	InDel	Position	Allele	MAF
M1	rs58723204	5405799–5405817	TATGTATGTATCT	0.1700	M9	rs79526052	87750028–87750029	TT	0.3800
M1	rs57109987	5405825–5405855	TCTATCTATCTATCTGT	0.3800	M10	rs71938699	116048233–116048237	TATAT	0.3400
M1	rs60804485	5405934–5405941	TATCT	0.4800	M10	rs71953348	116048303–116048306	TGT	0.2900
M2	rs36138671	7176051–7176085	GGATGGATGGATGGATG	0.3100	M10	rs71906073	116048375–116048381	TATAT	0.3800
M2	rs72344557	7176147–7176188	ATAGATAGATAGATAGATAGA	0.4300	M11	rs66680459	117572362	TC	0.2600
M3	rs72434864	32092836–32092841	AA	0.4300	M11	rs200762627	117572375–117572397	TTTCTTTTCTT	0.4200
M3	rs72179143	32092997–32093010	ACAA	0.2300	M12	rs59605609	119837905–119837912	AAATTA	0.5000
M4	rs3831707	41081541–41081543	AAGA	0.2300	M12	rs59241399	119837993–119837994	T	0.3000
M4	rs3831706	41081650	GA	0.4900	M13	rs200453545	122222730–122222731	CTACCTTCCC	0.3900
M5	rs77278564	42814969–42814971	AA	0.5000	M13	rs67487129	122222833–122222838	CTTACT	0.2700
M5	rs72012655	42815056–42815072	CACACTAGAAATC	0.2000	M14	rs34637872	128617383	TAT	0.2600
M6	rs35572306	68198270–68198280	TGT	0.2400	M14	rs35816575	128617476	TT	0.3400
M6	rs11297248	68198342–68198344	TT	0.4000	M15	rs10601024	130541932–130541937	TTTT	0.4800
M7	rs112705035	81328438–81328439	TCT	0.2000	M15	rs79998158	130541947–130541952	TACT	0.2500
M7	rs202133732	81328502–81328504	TCTC	0.2000	M16	rs78613336	133539846–133539847	ATATATAT	0.2500
M7	rs368767691	81328516	CA	0.4100	M17	rs200613017	150221805–150221811	TTTGT	0.2600
M8	rs200362185	84955192–84955202	ATATA	0.2200	M17	rs201990767	150221879–150221901	TTTCT	0.2700
M9	rs113311720	87749998–87750001	TAAATCT	0.3100					

Loci information is referenced to the GRCh37.p13 genome sequence. MAF is the minor allele frequency of selected InDel locus in the East Asian population reported in the 1000 Genome Project Phase III.

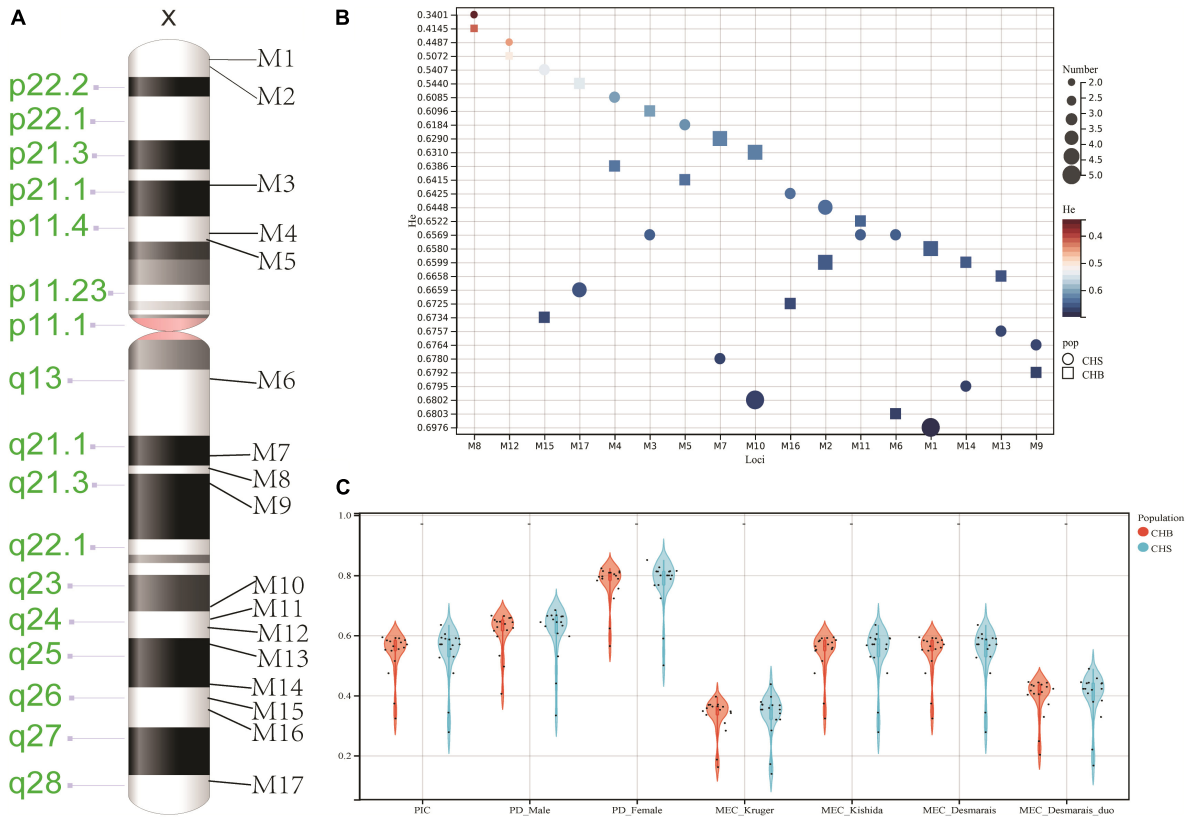


FIGURE 1
Physical locations **(A)**, genetic diversities **(B)**, and forensic efficiency evaluations **(C)** of 17 X-chromosomal Multi-InDels in Beijing Han (CHB) and Southern Han (CHS) populations. For genetic diversities, different colors represented different expected heterozygosity (He) values: red and blue denote small and large He values, respectively; the size of shape is proportional to the number of alleles observed in each locus.

and CHS populations since they showed high cumulative PD and MEC values.

Genetic distributions and forensic-related parameters of selected 17 loci in the studied Guizhou Han population

Even though capillary electrophoresis was widely used in forensic practice, it only detected length-based genetic variants that might reduce genetic diversities of studied genetic markers to some extent. Instead, not only can NGS detect length-based genetic variants, it can also discern sequence-based and additional genetic variants in surrounding regions of targeted markers (Børsting and Morling, 2015). In a previous study, Chen C. et al. (2021) investigated genetic polymorphisms of 231 genetic markers in the Chinese Hui group by the NGS, and they found that sequence-based genetic markers could show more allelic variations than length-based genetic markers. Thus, we developed the multiplex amplification system of 17 X chromosomal Multi-InDels based on the NGS technology. In this study, we chose the MGISEQ-2000RS sequencing platform to construct the multiplex amplification panel because previous studies pointed out that the MGISEQ-2000RS showed comparable or better performance than Illumina sequencing platforms (Jeon et al., 2021; Zhu et al., 2021).

Quality control results of 217 Guizhou Han individuals were presented in **Supplementary Table 4**. We found that Q20, Q30, and Coverage > = 100 of these individuals were greater than 0.89, 0.80, and 0.94, respectively. In addition, we also displayed DoC values of selected 17 loci in the Guizhou Han population (**Figure 2A**). Results revealed that most loci showed high DoC values, implying the developed panel possessed relatively balanced amplification performance. Even so, we found that the M10 locus displayed low DoC values in comparison to other loci.

The number of alleles and He values of these 17 loci in the Guizhou Han population were shown in **Figure 2B** and **Supplementary Table 5**. We found that most loci showed at least three allelic variations except the M8 locus. In addition, more than 10 alleles were observed at M1, M2, M7, and M17 loci. Compared to results from CHB and CHS populations, some loci showed more allelic variations in the studied

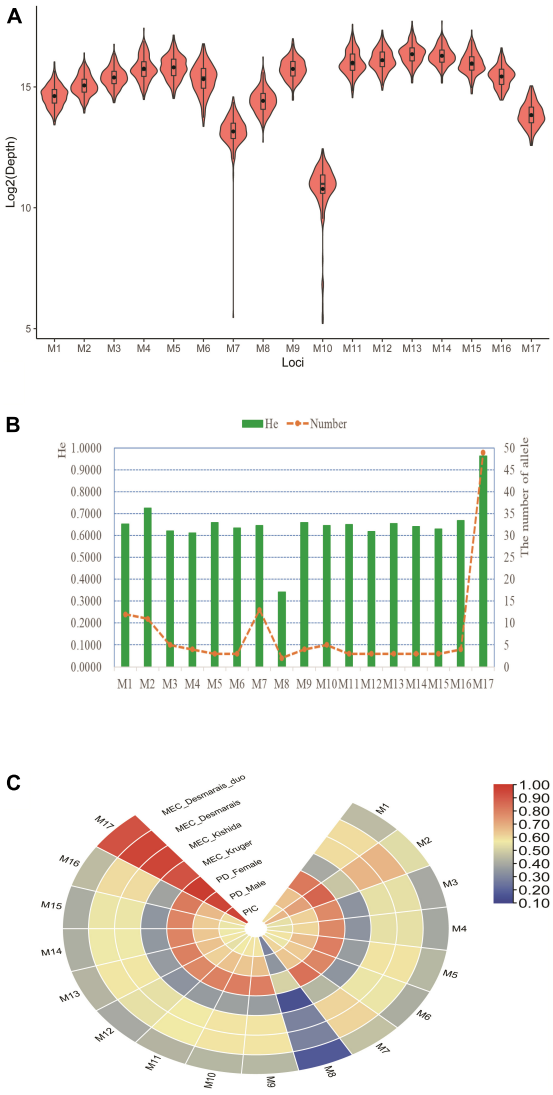


FIGURE 2
Depths of coverage **(A)**, expected heterozygosities **(B)**, and forensic parameters **(C)** of 17 X-chromosomal Multi-InDels in the studied Guizhou Han population.

population. We thought that NGS could detect potential sequence variations and other variations in the neighboring area, which made these loci possess higher genetic diversities. In addition, we observed that He values of 16 loci were larger than 0.5, implying that these loci exhibited relatively high genetic polymorphisms in the studied Han populations. Forensic-related parameters of these 17 loci were displayed in **Figure 2C** and **Supplementary Table 5**. The average PIC, PD_Male, PD_Female, MEC_Kruger, MEC_Kishida, MEC_Desmarais, and MEC_Desmarais_duo values of these 17 loci were 0.5828, 0.6449, 0.7999, 0.3875, 0.5827, 0.5828, and 0.4440, respectively. Besides, we also assessed linkage disequilibrium of these 17 loci in the Guizhou Han population (**Supplementary Table 6**). After applying Bonferroni correction (P = 0.05/136 = 0.0004), two linkage groups (LGs) were discerned from the studied Guizhou Han group. Among these two linkage groups, the number of haplotypes observed in LG1 (M2 and M3 loci) and LG2 (M10 and M17 loci) were 22 and 94, respectively. Haplotype diversities, PIC, PD_Male, PD_Female, MEC_Kruger, MEC_Kishida, MEC_Desmarais, and MEC_Desmarais_duo of LG1 were 0.8930, 0.8790, 0.8889, 0.9778, 0.7777, 0.8789, 0.8790, and 0.7936, respectively; and they were 0.9860, 0.9811, 0.9815, 0.9993, 0.9614, 0.9797, 0.9811, and 0.9634 at the LG2 locus, respectively. Next, we evaluated the cumulative forensic efficiency of two LGs and the remaining 13 X-chromosomal Multi-InDels in the Guizhou Han population. Accumulative PD_Male, PD_Female, MEC_Kruger, MEC_Kishida, MEC_Desmarais, and MEC_Desmarais_duo were 0.999999993691, 0.999999999999976, 0.999967, 0.999999936, 0.99999994, and 0.99999186, respectively. Compared to 17 autosomal Multi-InDels, 13 X chromosomal Multi-InDels, and 38 X-InDels (Fan et al., 2015; Qu et al., 2020; Chen L. et al., 2021), we found that these 17 loci presented in this study showed higher forensic application values in individual identification and paternity analysis, implying that the developed panel could be viewed as a more valuable tool for forensic research, especially for those complex paternity testing cases. Nonetheless, we found that the M8 locus showed relatively low genetic diversity in the Guizhou Han population. Besides, the M10 locus exhibited relatively low DoC values in comparison to other loci. Therefore, the developed system needs to be further improved in the future. Furthermore, developmental validation of the novel panel including mixture deconvolution, species specificity, stability, and concordance studies should be performed in the following study.

Population genetic analyses of different continental populations based on selected 17 loci

In a previous study, Fan et al. (2016) explored hierarchical genetic structure of different continental populations via 13 X-chromosomal Multi-InDels and found that these 13 loci could be used to differentiate East Asian, European, and African populations. Further, they also stated that Multi-InDels might be more appropriate for inferring biogeographical origins of different continental populations than multi-allelic InDels and STRs (Fan et al., 2016). Therefore, we also evaluated the power of these 17 loci to differentiate continental populations.

As shown in **Figure 3A**, we found that those populations from the same continent showed low F_{ST} values; whereas, populations from different continents had large F_{ST} values. Next, an MDS was also plotted based on pairwise F_{ST} values (**Figure 3B**). We found that these 26 populations formed four population clusters: five European populations located in the

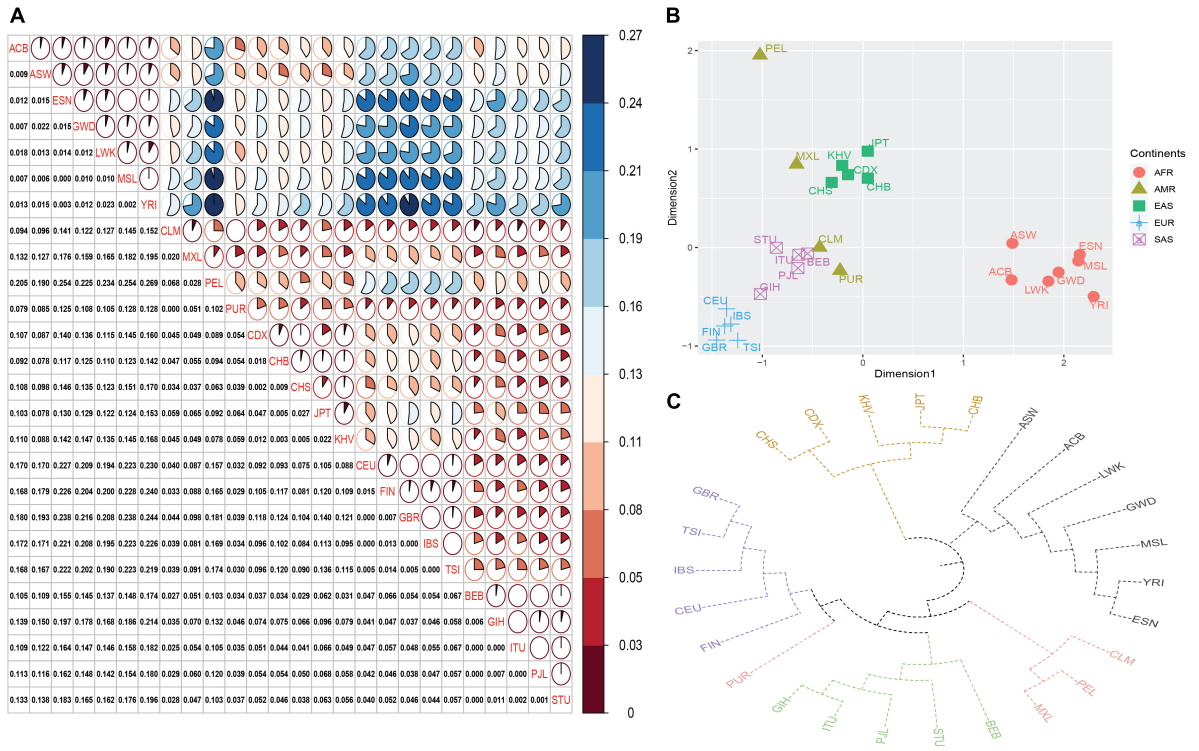


FIGURE 3
Population genetic analyses of different continental populations based on selected 17 loci. **(A)** Pairwise F_{ST} distances of different continental populations. Different colors represented different F_{ST} values: red denoted small F_{ST} values; blue denoted large F_{ST} values. **(B)** Multi-dimensional analysis of different continental populations. **(C)** The phylogenetic tree of different continental populations.

left bottom region; seven African populations positioned in the right area; five East Asian populations situated in the left top region; five South Asian populations clustered closely and distributed between East Asian and European populations. In addition, four American populations scattered among South Asian and East Asian populations. The phylogenetic tree of these 26 populations was given in **Figure 3C**. Similar to population distributions in MDS, we found that those populations from the same continent formed a branch except American populations. For American populations, abnormal population distribution patterns in MDS and phylogenetic tree might be related to their complex genetic components. On the one hand, the study about ancient DNA revealed that there were western Eurasian genetic signatures in modern-day Native Americans (Raghavan et al., 2014). On the other hand, four populations (MXL, PEL, PUR, and CLM) from American possessed different ancestral components from African, European, and indigenous American populations (Donnelly et al., 2015).

To further discern those loci showing large genetic differentiations among different continental populations, we also estimated pairwise F_{ST} values of each locus between different populations, as given in **Supplementary Table 7**. A previous study stated that genetic markers with high F_{ST} values between compared populations were conductive to differentiating these two populations (Phillips, 2015). We found that M2, M5, M6, M11, M14, and M15 loci displayed relatively high F_{ST} values (>0.1) between African population and other continental populations, implying that these loci could be viewed as potential ancestry informative markers for inferring biogeographical origins of African populations. Likewise, M14 locus showed relatively high F_{ST} values between East Asian and European, American, and South Asian populations; M13 locus displayed relatively high F_{ST} values between European and American and South Asian populations, which indicated that these loci could be utilized for differentiating these continental populations. In the following study, we need to further evaluate whether these loci can discern population substructure of Chinese different ethnic groups.

Conclusion

To conclude, we developed a novel multiplex amplification panel of 17 X-chromosomal Multi-InDels via the NGS platform. The majority of these loci showed relatively high genetic diversities in CHB, CHS, and studied Han populations, and they can be viewed as a valuable supplementary tool for forensic personal identification and paternity analyses, especially for

those deficiency cases. In addition, we found that some out of these 17 loci were also beneficial to differentiating African, European, East Asian, and South Asian populations.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: CNGB Sequence Archive (CNSA) of the China National Genebank Database, accession number CNP0003564.

Ethics statement

The studies involving human participants were reviewed and approved by the Guizhou Medical University Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

Author contributions

XJ and ZR wrote this manuscript. HLZ, QW, YL, and JJ collected samples and performed the experiment. MY, HZ, WH, and NW conducted data analysis. YW and JH designed the work and provided the conception. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Guizhou Provincial Science and Technology Projects (No. ZK [2022] General 355), the Guizhou Education Department Young Scientific and Technical Talents Project, Qian Education KY (No. [2022] 215), the Guizhou Scientific Support Project, Qian Science Support (No. [2021] General 448), the Shanghai Key Lab of Forensic Medicine, the Key Lab of Forensic Science, Ministry of Justice, China (Academy of Forensic Science), Open Project, (No. KF202207), the Guizhou Province Education Department, Characteristic Region Project, Qian Education KY (No. [2021] 065), the Guizhou "Hundred" High-level Innovative Talent Project, Qian Science Platform Talents (No. [2020] 6012), the Guizhou Scientific Support Project, Qian Science Support (No. [2020] 4Y057), the Guizhou Science Project, Qian Science Foundation (No. [2020] 1Y353), the Guizhou Scientific Support Project, Qian Science Support (2019) 2825, the Guizhou Scientific Cultivation Project, Qian Science Platform Talent (No. [2018] 5779-X), the Guizhou Engineering Technology Research Center Project, the Qian High-Tech of Development and Reform Commission (No. [2016] 1345), the Guizhou Innovation training program for college students (No. [2019] 5200926), and the National Natural Science Foundation of China (No. 82160324).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.985933/full#supplementary-material

References

Børsting, C., and Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. *Forensic Sci. Int. Genet.* 18, 78–89. doi: 10.1016/j.fsigen.2015.02.002

Caputo, M., Amador, M. A., Santos, S., and Corach, D. (2017). Potential forensic use of a 33 X-InDel panel in the Argentinean population. *Int. J. Legal Med.* 131, 107–112. doi: 10.1007/s00414-016-1399-z

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, C., Jin, X., Zhang, X., Zhang, W., Guo, Y., Tao, R., et al. (2021). Comprehensive insights into forensic features and genetic background of chinese northwest Hui group using six distinct categories of 231 molecular markers. *Front. Genet.* 12:705753. doi: 10.3389/fgene.2021.705753

Chen, L., Pan, X., Wang, Y., Du, W., Wu, W., Tang, Z., et al. (2021). Development and validation of a forensic multiplex system with 38 X-InDel loci. *Front. Genet.* 12:670482. doi: 10.3389/fgene.2021.670482

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and

preprocessing of high-throughput sequencing data. *Gigascience* 7, 1–6. doi: 10.1093/gigascience/gix120

Desmarais, D., Zhong, Y., Chakraborty, R., Perreault, C., and Busque, L. (1998). Development of a highly polymorphic STR marker for identity testing purposes at the human androgen receptor gene (HUMARA). *J. Forensic Sci.* 43:14355J. doi: 10.1520/jfs14355j

Donnelly, P., Green, E. D., Knoppers, B. M., Mardis, E. R., Nickerson, D. A., Wilson, R. K., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

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

Fan, G., Ye, Y., Luo, H., and Hou, Y. (2015). Use of multi-InDels as novel markers to analyze 13 X-chromosome haplotype loci for forensic purposes. *Electrophoresis* 36, 2931–2938. doi: 10.1002/elps.201500159

Fan, G. Y., Ye, Y., and Hou, Y. P. (2016). Detecting a hierarchical genetic population structure via Multi-InDel markers on the X chromosome. *Sci. Rep.* 6:32178. doi: 10.1038/srep32178

Fan, H., He, Y., Li, S., Xie, Q., Wang, F., Du, Z., et al. (2022a). Systematic evaluation of a novel 6-dye direct and multiplex PCR-CE-based InDel typing system for forensic purposes. *Front. Genet.* 12:744645. doi: 10.3389/fgene.2021.744645

Fan, H., Wang, L., Liu, C., Lu, X., Xu, X., Ru, K., et al. (2022b). Development and validation of a novel 133-plex forensic STR panel (52 STRs and 81 Y-STRs) using single-end 400 bp massive parallel sequencing. *Int. J. Legal Med.* 136, 447–464. doi: 10.1007/s00414-021-02738-1

Gomes, I., Pinto, N., Antão-Sousa, S., Gomes, V., Gusmão, L., and Amorim, A. (2020). Twenty years later: a comprehensive review of the X chromosome use in forensic genetics. *Front. Genet.* 11:926. doi: 10.3389/fgene.2020.00926

Gouy, A., and Zieger, M. (2017). STRAF—A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci. Int. Genet.* 30, 148–151. doi: 10.1016/j.fsigen.2017.07.007

Huang, J., Luo, H., Wei, W., and Hou, Y. (2014). A novel method for the analysis of 20 multi-Indel polymorphisms and its forensic application. *Electrophoresis* 35, 487–493. doi: 10.1002/elps.201300346

Jeon, S. A., Park, J. L., Park, S. J., Kim, J. H., Goh, S. H., Han, J. Y., et al. (2021). Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. *Genes Genom.* 43, 713–724. doi: 10.1007/s13258-021-01096-x

Jin, R., Cui, W., Fang, Y., Jin, X., Wang, H., Lan, Q., et al. (2021). A novel panel of 43 insertion/deletion loci for human identifications of forensic degraded DNA samples: development and validation. *Front. Genet.* 12:610540. doi: 10.3389/fgene.2021.610540

Kishida, T., Wang, W., Fukuda, M., and Tamaki, Y. (1997). Duplex PCR of the Y-27H39 and HPRT loci with reference to japanese population data on the HPRT locus. *Japanese J. Leg. Med.* 51, 67–69.

Krüger, J., Fuhrmann, W., Lichte, K., and Steffens, C. (1968). On the utilization of erythrocyte acid phosphatase polymorphism in paternity evaluation. *Dtsch Z Gesamte Gerichtl Med.* 64, 127–146.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lang, Y., Guo, F., and Niu, Q. (2019). StatsX v2.0: the interactive graphical software for population statistics on X-STR. *Int. J. Legal Med.* 133, 39–44. doi: 10.1007/s00414-018-1824-6

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–1760.

Li, R., Shen, X., Chen, H., Peng, D., Wu, R., and Sun, H. (2021). Developmental validation of the MGIEasy signature identification library prep kit, an all-in-one multiplex system for forensic applications. *Int. J. Legal Med.* 135, 739–753. doi: 10.1007/s00414-021-02507-0

Phillips, C. (2015). Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci. Int. Genet.* 18, 49–65. doi: 10.1016/j.fsigen.2015.05.012

Qu, S., Lv, M., Xue, J., Zhu, J., Wang, L., Jian, H., et al. (2020). Multi-indel: a microhaplotype marker can be typed using capillary electrophoresis platforms. *Front. Genet.* 11:567082. doi: 10.3389/fgene.2020.567082

Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., et al. (2014). Upper palaeolithic Siberian genome reveals dual ancestry of native Americans. *Nature* 505, 87–91. doi: 10.1038/nature12736

Schmidt, S., McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., 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

Sun, K., Yun, L., Zhang, C., Shao, C., Gao, T., Zhao, Z., et al. (2019). Evaluation of 12 Multi-InDel markers for forensic ancestry prediction in Asian populations. *Forensic Sci. Int. Genet.* 43:102155. doi: 10.1016/j.fsigen.2019.102155

Szibor, R. (2007). X-chromosomal markers: past, present and future. *Forensic Sci. Int. Genet.* 1, 93–99. doi: 10.1016/j.fsigen.2007.03.003

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., and Marth, G. (2002). Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.* 71, 854–862. doi: 10.1086/342727

Zhang, S., Sun, K., Bian, Y., Zhao, Q., Wang, Z., Ji, C., et al. (2015). Developmental validation of an X-Insertion/Deletion polymorphism panel and application in HAN population of China. *Sci. Rep.* 5:18336. doi: 10.1038/srep18336

Zhang, X., Shen, C., Jin, X., Guo, Y., Xie, T., and Zhu, B. (2021). Developmental validations of a self-developed 39 AIM-InDel panel and its forensic efficiency evaluations in the Shaanxi Han population. *Int. J. Legal Med.* 135, 1359–1367. doi: 10.1007/s00414-021-02600-4

Zhu, K., Du, P., Xiong, J., Ren, X., Sun, C., Tao, Y., et al. (2021). Comparative performance of the MGISEQ-2000 and Illumina X-ten sequencing platforms for paleogenomics. *Front. Genet.* 12:745508. doi: 10.3389/fgene.2021.745508

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/fevo.2022.1014024

Check for updates

OPEN ACCESS

EDITED BY
Guanglin He,
Sichuan University, China

REVIEWED BY
Xiaoye Jin,
Guizhou Medical University, China
Jun Yao,
China Medical University, China

*CORRESPONDENCE
Youfeng Wen
wenyf@jzmu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 08 August 2022
ACCEPTED 09 September 2022
PUBLISHED 03 November 2022

CITATION
Zhou J, Zhang X, Li X, Sui J, Zhang S, Zhong H, Zhang Q, Zhang X, Huang H and Wen Y (2022) Genetic structure and demographic history of Northern Han people in Liaoning Province inferred from genome-wide array data. *Front. Ecol. Evol.* 10:1014024. doi: 10.3389/fevo.2022.1014024

COPYRIGHT
© 2022 Zhou, Zhang, Li, Sui, Zhang, Zhong, Zhang, Zhang, Huang and Wen. 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.

Genetic structure and demographic history of Northern Han people in Liaoning Province inferred from genome-wide array data

Jingbin Zhou†, Xianpeng Zhang†, Xin Li, Jie Sui, Shuang Zhang, Hua Zhong, Qiuxi Zhang, Xiaoming Zhang, He Huang and Youfeng Wen*

Institute of Biological Anthropology, Jinzhou Medical University, Jinzhou, China

In this study, we used typical and advanced population genetic analysis methods [principal component analysis (PCA), ADMIXTURE, F_{ST}, f_3-statistics, f_4-statistics, *qp*Adm/*qp*Wave, *qp*Graph, ALDER (Admixture-induced Linkage Disequilibrium for Evolutionary Relationships) and TreeMix] to explore the genetic structure of 80 Han individuals from four different cities in Liaoning Province and reconstruct their demographic history based on the newly generated genome-wide data. We found that Liaoning Han people have genetic similarities with other northern Han people (Shandong, Henan, and Shanxi) and Liaoning Manchu people. Millet farmers in the Yellow River Basin (YRB) and the West Liao River Basin (WLRB) (57–98%) and hunter-gatherers in the Mongolian Plateau (MP) and the Amur River Basin (ARB) (40–43%) are the main ancestral sources of the Liaoning Han people. Our study further supports the "northern origin hypothesis"; YRB-related ancestry accounts for 83–98% of the genetic makeup of the Liaoning Han population. There are clear genetic influences of northern East Asian populations in the Liaoning Han people, ancient Northeast Asian-related ancestry is another dominant ancestral component, and large-scale population admixture has happened between Tungusic Manchu people and Han people. There are genetic differences among the Liaoning Han people, and we found that these differences are associated with different migration routes of Hans during the "Chuang Guandong" period in historical records.

KEYWORDS

genetic structure, population history, Han people, genome-wide data, Liaoning Province of China

Introduction

The origin of the Sino–Tibetan language family is a controversial issue. Recently, an increasing number of studies supported the "northern origin hypothesis" from different perspectives, such as archeology, genetics, and linguistics. These studies also reported that the dispersal of the Sino–Tibetan language supports the "farming-language dispersal

hypothesis", Neolithic millet farmers in the Yellow River Basin (YRB), who are associated with Yangshao and/or Majiayao cultures, may have been the ancestors of Sino–Tibetan language speakers (Sagart et al., 2019; Zhang et al., 2019; Wang C. C. et al., 2021). The "northern origin hypothesis" is consistent with historical records that the Han population originated from the Huaxia tribe in the YRB. Modern Han people comprise the largest population among Sino–Tibetan language speakers and are the most populous ethnic group in China and East Asia, with a current population of a staggering 1.286 billion individuals (Seventh Census). Previous studies found that the Han population can be divided into two distinct groups, the northern Han population and the southern Han population, and Han people in different regions have different genetic profiles, but they all exhibit mixed characteristics: dominant ancestry related to Neolithic YRB farmers and minor genetic contributions from geographically different indigenous peoples (He et al., 2020; Wang et al., 2021a,b; Yao et al., 2021; Zhang X. et al., 2021; He et al., 2022b). These studies reveal a complex demographic history of the Han people. The Han population or their ancestors frequently expanded northward and southward throughout history, mixed with different indigenous ethnic groups, and accumulated genetic diversity, and they played an important role in the formation of the genetic structure of East Asian populations. Previous studies described in depth the genetic contributions of the Han population and its ancestors to modern southern ethnic groups, such as Tai–Kadai, Austroasiatic, Austronesian, and Hmong–Mien-speaking populations (He et al., 2020; Bin et al., 2021; Luo et al., 2021; Wang C. C. et al., 2021; Wang et al., 2022; Guo et al., 2022), but the genetic affinities and structure of northern Han people and northern ethnic groups are still unclear due to the sparse sampling of present-day populations and a low coverage of single-nucleotide polymorphisms (SNPs).

Liaoning is a northern province in China, and the largest ethnic group in this province is Han, with 36.16 million people (Seventh Census), followed by the Manchu population, with more than 5 million people. In addition, 50 other ethnic groups live in Liaoning Province. The dynamic population history shows rich cultural, linguistic, and genetic diversities. The West Liao River Basin (WLRB) and the Amur River Basin (ARB) geographic regions are adjacent to Liaoning Province, and modern and ancient populations in these regions have played an indispensable role in the formation of the genetic structure, culture, and language in East Asia, especially northern East Asia. Previous studies found that the WLRB may be the cradle of the Transeurasian language family according to evidence from linguistics, archeology, and genetics, and the diffusion of the Transeurasian language followed the expansion of WLRB millet farmers (Robbeets et al., 2021), but there are opposing views. Wang et al. did not find a genetic contribution of the WLRB millet farmers in modern and ancient populations in the Mongolian Plateau (MP) and ARB areas (Wang C. C. et al., 2021). Many linguists did not recognize the proto-Transeurasian language, and they argued that the commonalities between different Transeurasian language groups were caused by historical exchange and interaction (Robbeets and Savelyev, 2020). Paleogenomic studies show that there are up to 14,000 years of genetic continuity from the ancient ARB population to the modern Tungusic people, and the latter also maintain genetic homogeneity with each other (He et al., 2021; Mao et al., 2021; Wang C. C. et al., 2021); but there is an exception, in that the Tungusic Manchu people exhibit significant genetic similarity with the northern Han people, which reveals that a large-scale population admixture occurred between the Manchus and the Hans (Zhang X. et al., 2021). Genetic influences of Han people and their ancestors can also be observed in other populations in Liaoning Province. Previous studies based on autosomal short tandem repeat (STR), Y/X chromosome short tandem repeat (Y/X-STR), and mitochondrial DNA (mtDNA) revealed close genetic relationships between the Liaoning Han population and the Manchu, Mongolian, Xibo and Chinese Korean populations in Liaoning Province (Yao and Wang, 2016; Guo, 2017a; Du et al., 2020). However, compared with other ethnic groups, the Liaoning Han people exhibit a closer genetic affinity with the northern Han people in other regions, such as Heilongjiang, Jilin, Hebei, and Shandong (Yao and Wang, 2016; Yao et al., 2016; Guo, 2017b). Nevertheless, due to limited sampling and few genetic markers, the genetic structure and profile of the Liaoning Han people are still unclear. In this study, we obtained genome-wide data for Han people from Dalian, Dandong, Shenyang, and Panjin in Liaoning Province to perform population genetic analysis. This study mainly aims to explore (1) the genetic profile and structure of the Liaoning Han population; (2) the genetic affinity between the Liaoning Han population and the Tungusic- and Mongolic-speaking populations; (3) the number of ancestral sources contributing to the Liaoning Han population; (4) the genetic contributions of ancient millet farmers in the YRB and WLRB; and finally, (5) the origin of the Liaoning Han population.

Materials and methods

Sampling and genotyping

In this study, we collected saliva samples from 80 Han individuals (19 Shenyang Hans, 20 Dandong Hans, 21 Dalian Hans and 20 Panjin Hans) from four cities in Liaoning Province, China (Supplementary Table 1). Every participant signed the informed consent form before the start of the study. The geographical positions of the sampling population in this study are displayed in Figure 1. All participants were required to be indigenous residents whose ancestors had lived at the sampling sites for at least three generations, and the ethnic groups of

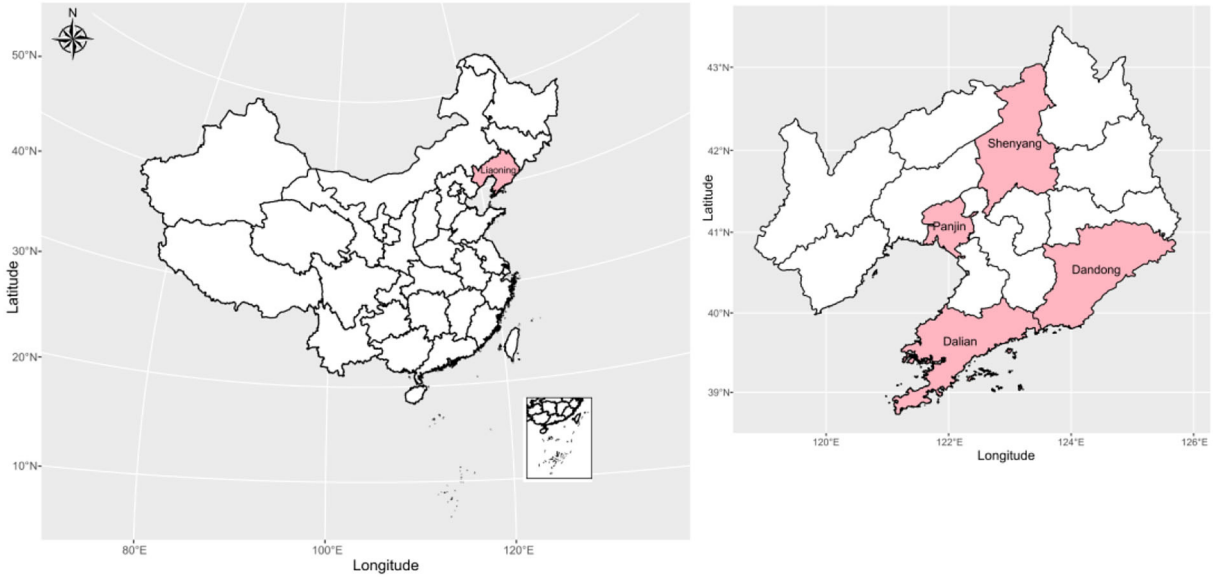


FIGURE 1
Geographical positions of the sampling population in this study.

all participants were assigned according to the self-declaration based on the family migration history and corresponding family records. This research was reviewed and approved by the Medical Ethics Committee of Jinzhou Medical University (JZMULL2021101), and all procedures were carried out by following the recommendations of the Declaration of Helsinki of 2000 (Helsinki, 2001). The DNA extraction was performed with a genomic DNA extraction kit following the manufacturer's instructions. DNA sequencing and genotyping were carried out by using Illumina WeGene Arrays. Finally, we obtained the raw dataset in bplink format (bed, bim and fam), which contained 717,228 SNPs. Then, we used PLINK 1.9 software (Purcell et al., 2007) to perform initial filtering based on our predefined parameter thresholds (-maf: 0.01, -hwe: 0.0001, -mind: 0.01 and -geno: 0.01), and a dataset containing 456,201 SNPs from 80 Han individuals was obtained. Then, we used genome-wide complex trait analysis (GCTA) (Yang et al., 2011) software to infer the genetic relationships between newly sampled Han individuals and the individuals who showed close kinship up to the third degree (kinship value $>$ 0.125), with other newly sampled individuals excluded to ensure that all members of the sample population in this study were unrelated. The results are listed in Supplementary Table 2 and Supplementary Figure 1, showing no kinship within the third degree between newly sampled Han individuals.

Data merging

In this study, we used two merged datasets:

1) "Merged-HO" dataset: We merged our newly generated dataset of 80 Han individuals with previously published modern and ancient population data from the Affymetrix Human Origins (HO) Array dataset (Patterson et al., 2012).
2) "Merged-1240K" dataset: We also merged new genotype data with the 1240K dataset from the Reich Lab (https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data).

In addition, we merged these two datasets with our previous Liaoning Manchu and Mongolian datasets (Zhang X. et al., 2021; Hou et al., 2022). Finally, we obtained the "Merged-HO" dataset containing 41,585 SNPs and the "Merged-1240K" dataset containing 113,144 SNPs to perform the population genetic analysis.

Principal component and ADMIXTURE analyses

In this study, we applied the principal component analysis (PCA) and the ADMIXTURE analysis based on the "Merged-HO" dataset to explore the population structure in East Asia. The PCA was conducted by using the smartpca package of EIGENSOFT software (Patterson et al., 2006) with the default options lsqproject: YES and numoutlieriter: 0. Visualization was performed by the R package ggplot2. Then, we used PLINK 1.9 (Purcell et al., 2007) with the parameter "-indep-pairwise 200 25 0.4" to remove SNPs in strong linkage

disequilibrium, and we then applied ADMIXTURE (Alexander et al., 2009), which is a model-based clustering analysis, with a predefined number of ancestral sources (K) ranging from 2 to 20. The optimal number of ancestral sources (K) was selected using 10-fold cross-validation (CV) errors, which are listed in Supplementary Table 3. Visualization of the ADMIXTURE results was carried out by using the R package pophelper.

Pairwise-F_{ST} genetic distance

We used the smartpca program in EIGENSOFT software (Patterson et al., 2006) with the parameter fstonly: YES to calculate pairwise F_{ST} genetic distances between Liaoning Han and other modern reference populations. Visualization was carried out by using the R package pheatmap.

f-statistics

We used ADMIXTOOLS (Patterson et al., 2012) to calculate all f-statistics. First, a three-population test (f_3-statistics) was carried out by using the qp3pop program in ADMIXTOOLS with default parameters. We calculated outgroup-f_3 (Han_Liaoning, Y; Mbuti) values based on the "Merged-HO" dataset and the "Merged-1240K" dataset to measure shared genetic drift between the Liaoning Han and other Eurasian populations (Y). We also conducted admixture-f_3 (X, Y; Han_Liaoning) analysis to explore the possible genetic contributors and potential admixture signals. Then, we carried out a four-population test (f_4-statistics) by using the qpDstat program in ADMIXTOOLS with default parameters. We calculated f_4 (Modern/Ancient population1, Modern/Ancient population2; Han_Liaoning, Mbuti) and f_4 (Modern/Ancient population1, Han_Liaoning; Modern/Ancient population2, Mbuti) values to examine shared alleles and explore the direction of gene flow.

QpAdm and QpWave

The *qp*Adm and *qp*Wave programs in ADMIXTOOLS (Patterson et al., 2012) were applied to the "Merged-HO" dataset and the "Merged-1240K" dataset to explore the minimum number of ancestral sources and quantify ancestral proportions. We used 10 outgroups, namely, Mbuti, Malaysia_LN, Tianyuan, Papuan, Ust_Ishim, GreatAndaman, Kostenki14, Australian, Mixe, and Atayal, to test the two-way admixture model.

*Qp*Graph and TreeMix

To identify the best-fitting phylogenetic framework with population splits and gene flow events, we used the qpGraph program in ADMIXTOOLS and TreeMix software to reconstruct the deep population history of the Liaoning Han people.

ALDER and Y-chromosomal and MtDNA lineages

The admixture time and possible ancestral sources of Liaoning Han were estimated by using multiple Admixture-induced Linkage Disequilibrium for Evolutionary Relationships (ALDERs) (Loh et al., 2013). Y-chromosomal and mitochondrial haplogroups were assigned by using an in-house script and following the recommendations of the International Society of Genetic Genealogy (ISOGG; http://www.isogg.org/) and mtDNA PhyloTree17 (http://www.phylotree.org/). The haplogroup information based on mtDNA and the Y chromosome is provided in Supplementary Table 11.

Results

Population structure and genetic affinity of East Eurasian populations

The structure of the East Eurasian population was indicated by the results of PCA and ADMIXTURE analysis. Figure 2A shows that the population structure is consistent with language category and geographic location. Populations who belong to the same language group or have adjoined geographic distributions show a close genetic relationship, and there are clear Han clines, Tibeto–Burman clines, and Tai–Kadai clines. The Liaoning Han people are located along the Han cline and show a close relationship with other northern Han people. In addition, modern Liaoning Manchu people and the ancient YRB population overlap with the Liaoning Han people, as shown in Figure 2A, revealing close genetic relationships among them. We also found that some populations with adjoining geographic locations with the Liaoning Han people, such as Liaoning Mongolians, Japanese, and Koreans, exhibit close genetic relationships. There are no significant genetic differences among the Liaoning Han populations in Dandong, Dalian, Shenyang, and Panjin. The ADMIXTURE results also support the aforementioned findings (Figure 2B). In this study, we found that the CV error is the lowest when $K = 9$. There are four dominant ancestral components in the genetic makeup of the Liaoning Han population. The blue ancestry is maximized in modern Han people and ancient YRB populations; the yellow ancestry is enriched in modern Tibetans and ancient Nepalese

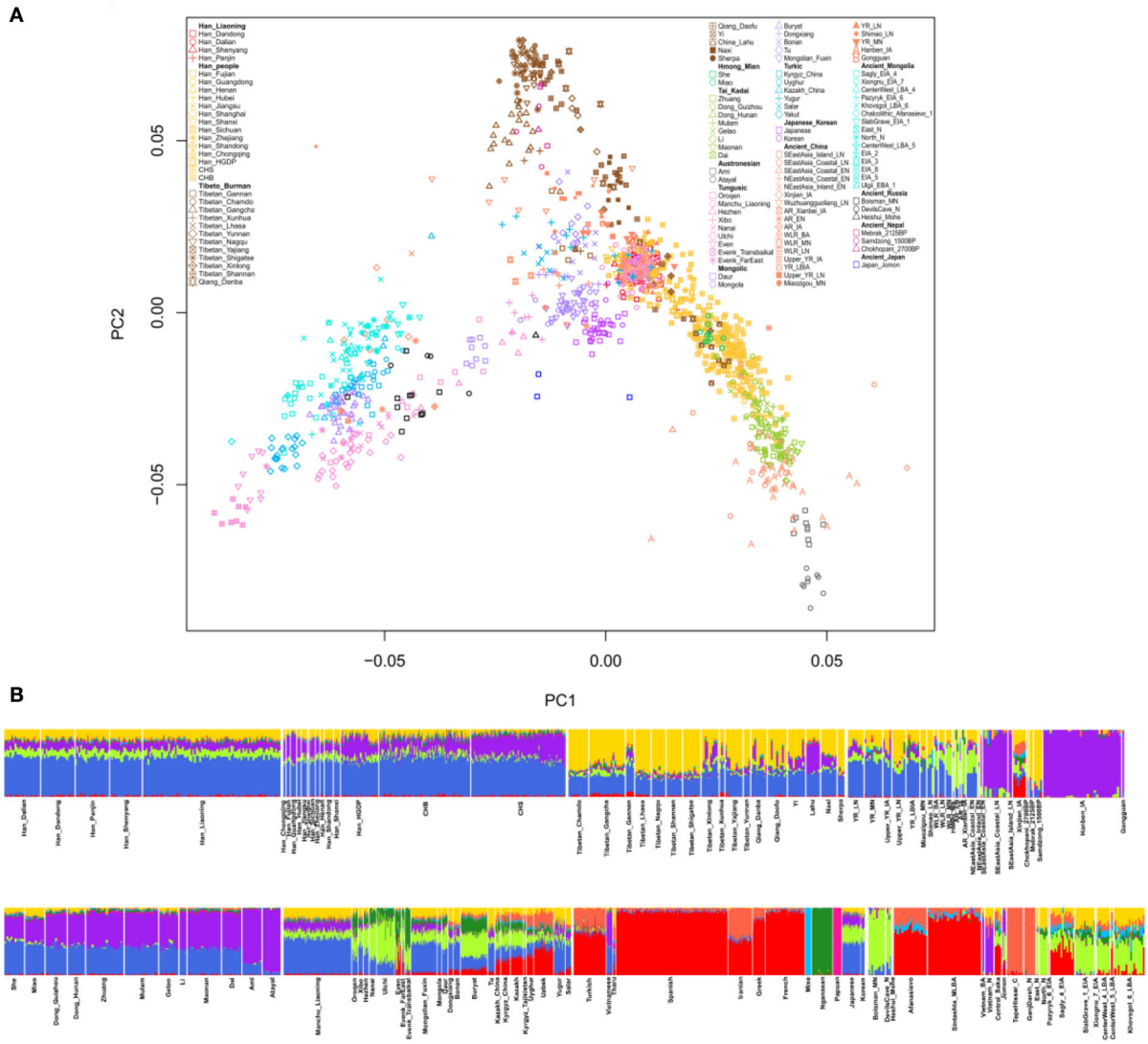


FIGURE 2
Population structure of ancient and modern Eurasian populations. **(A)** Principal component analysis; **(B)** ADMIXTURE-based model cluster analysis.

populations; the purple ancestry is maximized in Iron Age Hanben individuals and Neolithic Fujian coastal and island populations and enriched in modern Tai–Kadai speakers and Austronesian speakers; the green ancestry is maximized in ancient ARB hunter-gatherers and modern Tungusic speakers such as Ulchi and Nanai; there is a small amount of West Eurasian ancestry in the genetic makeup of the Liaoning Han people. The genetic profile of the Liaoning Han people is similar to those of other northern Han people and Liaoning Manchu people, and there is more YRB- and ARB-related ancestry and less southern ancestry relative to those in the southern Han people. The pairwise F_{ST} genetic distances also showed the same genetic profile of the Liaoning Han people (Figure 3). The shortest genetic distances were observed between the Liaoning Han people and other northern Han people, such as the Shandong Han and Henan Han people and the Liaoning Manchu people. Fuxin Mongolian, Korean, Japanese, Xibo, and other northern populations also exhibit closer genetic relationships with the Liaoning Han people (Supplementary Table 4).

f_3-statistics and f_4-statistics

To further explore genetic affinity, potential admixture signals, and genetic flow direction among East Asian

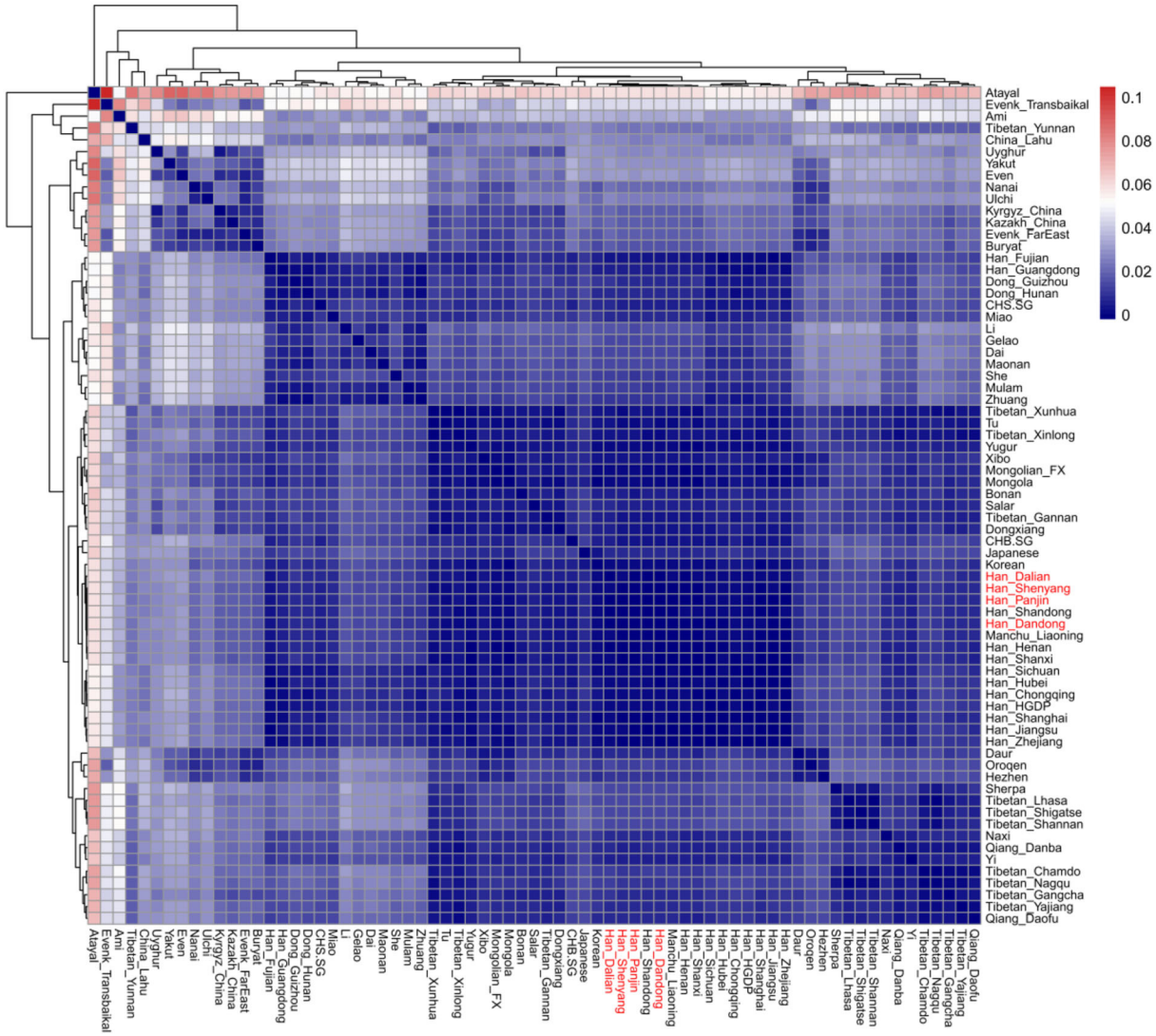


FIGURE 3
Heatmap of genetic affinity among modern East Asian populations based on pairwise F_{ST} genetic distance.

populations, we performed outgroup-f_3, admixture-f_3, and f_4-statistical analyses. The results of Outgroup-f_3 (Han_Liaoning, Y; Mbuti) based on the "Merged-1240K" dataset and the "Merged-HO" dataset indicate that the Liaoning Han people have genetic similarity with the northern and southern Han people, the Liaoning Manchu people, Japanese people, and Korean people relative to other East Asian populations (Figures 4A,D, Supplementary Tables 5A–F). Some Tibeto–Burman speakers, such as Naxi, Tujia, and Yi, and southern East Asian populations also have a close genetic relationship with the Liaoning Han people. We also found genetic differences between the Liaoning Han and western Eurasian populations and the Turkic-speaking populations. The above results can be found in the Dandong, Dalian, Shenyang, and Panjin Han peoples, and there is clear genetic homogeneity among the Liaoning Han populations. When Y represented ancient populations, we found that the Liaoning Han people shared more genetic drift with the YRB and WLRB populations during the Neolithic to Iron Age. We also found that ARB populations, such as Mohe, Boisman, and Devils Cave, Iron Age Hanben, and Neolithic Fujian coastal populations, shared genetic drift with the Liaoning Han people (Figure 4B, Supplementary Tables 5G–K). The results of admixture-f_3 (Modern/Ancient population1, Modern/Ancient population2; Mbuti) indicated that the Liaoning Han people could be modeled as a mixture between southern populations

10.3389/fevo.2022.1014024

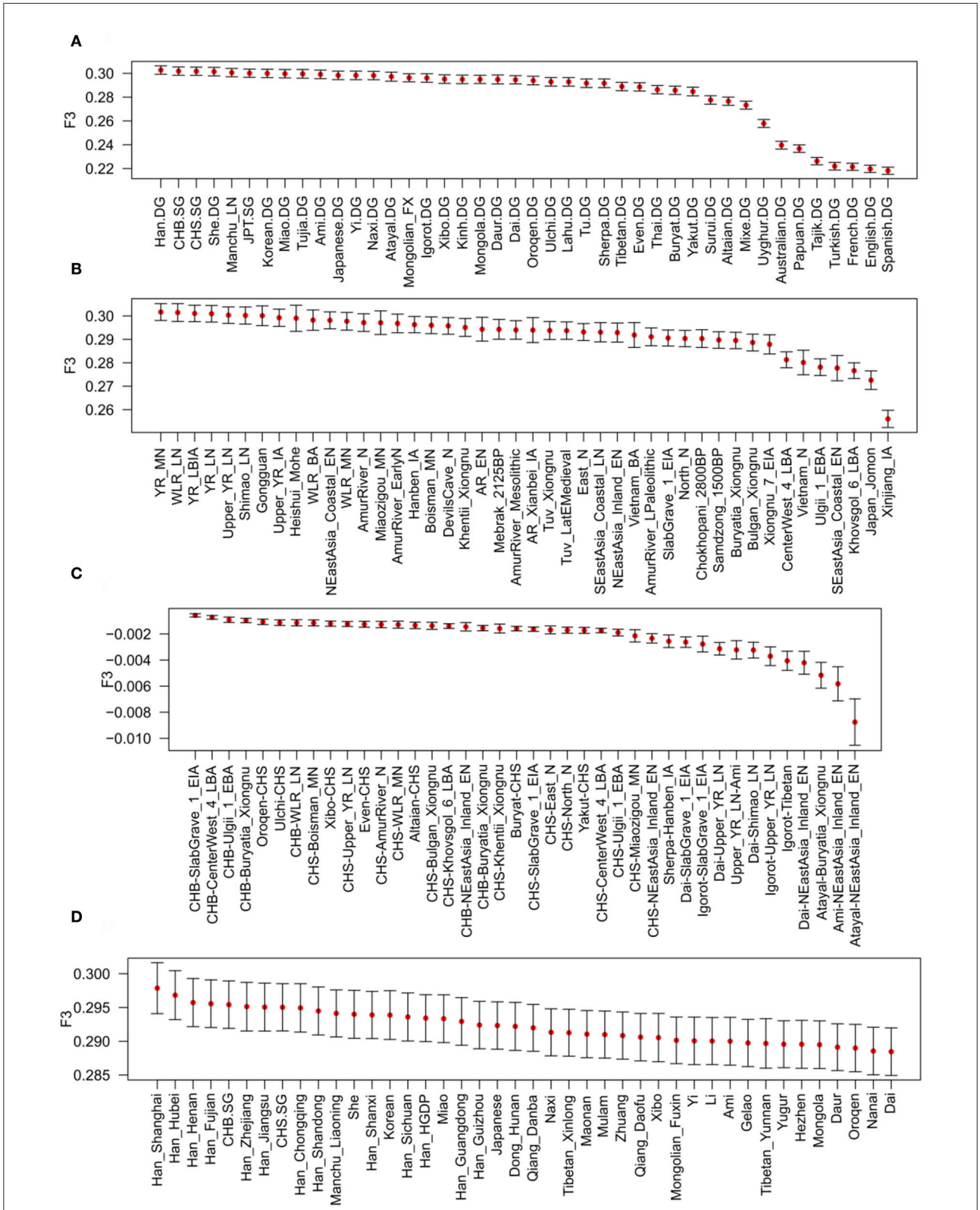


FIGURE 4
Shared genetic drift and genetic affinity measured *via* f_3-statistics analysis. **(A)** Outgroup-f_3 (Han_Liaoning, Modern population; Mbuti) based on the "Merged-1240K" dataset; **(B)** outgroup-f_3 (Han_Liaoning, Ancient population; Mbuti) based on the "Merged-1240K" dataset; **(C)** admixture-f_3 (Modern/Ancient population1, Modern/Ancient population2; Han_Liaoning) based on the "Merged-1240K" dataset; **(D)** outgroup-f_3 (Han_Liaoning, Modern population; Mbuti) based on the "Merged-HO" dataset.

represented by southern Han (CHS), Dai, and Ami and northern populations in the YRB and in the ARB and on the Mongolian Plateau (MP) (Figure 4C, Supplementary Table 5L).

The results of f_4-statistics in the form of f_4 (Modern population1, Modern population2; Han_Liaoning, Mbuti) indicated that the Liaoning Han people shared more alleles with northern and southern Han people, Liaoning Manchu people, Liaoning Mongolian, Japanese, Korean, and some southern East Asian populations, such as Dai, She, Yi, and Miao, than with western Eurasian populations and other East Asian populations (Supplementary Table 6A). The results of f_4-statistics in the form of f_4 (Ancient population1, Ancient population2; Han_Liaoning, Mbuti) revealed that Liaoning Han people shared more alleles with YRB and WLRB populations than with other ancient Eurasian populations, and there was gene flow related to the ARB and MP populations and ancient southern populations (Supplementary Table 6B). To further determine asymmetric genetic relationships, we calculated f_4-statistics in the form of f_4 (Han population/Manchu_Liaoning/Mongolian_Liaoning, Han_Liaoning; Modern population, Mbuti) based on the "Merged-HO" dataset and f_4 (Han population/Manchu_Liaoning/Mongolian_Liaoning/Xibo/Daur /Oroqen/Hezhen, Han_Liaoning; Modern population, Mbuti) based on the "Merged-1240K" dataset (Supplementary Table 7, Figures 5A,B). The Liaoning Han people showed significant genetic similarities with other northern Han people (Henan, Shandong and Shanxi) and Liaoning Manchu people, southern Han populations shared more alleles with Tai–Kadai, Hmong–Mien, and Austronesian-speaking populations than Liaoning Hans, modern Tungusic speakers exhibited genetic similarities with each other and showed clear genetic differences with Liaoning Hans, and Liaoning Mongolians had more genetic influence from western Eurasian populations than Liaoning Hans. The Liaoning Han people have genetic homogeneity, but we found that Dalian Hans showed a closer genetic affinity with Shandong Hans than Shenyang Hans, and Dalian Hans shared more alleles with some Han people (Shandong Hans, Fujian Hans, CHS and Han Chinese Beijing (CHB)), Tibeto–Burman speakers (Naxi, Yi, Lhasa, and Yajiang Tibetan), Liaoning Manchu people, Liaoning Mongolians, and Ami than Panjin Hans. In the form of f_4 (Ancient YRB/WLRB/ARB/MP/Nepal population, Han_Liaoning; Ancient population, Mbuti), we found that the Liaoning Han people received more genetic contributions from ancient YRB, WLRB, and ARB populations. Ancient individuals from archeological sites, such as Dacaozi, Luoheguxiang, Haojiatai, and Jiaozuoniecun, around the YRB and ancient cultures, such as Mohe, Mebrak, Upper and Lower Xiajiadian, Yangshao, and Longshan, shared more alleles with the Liaoning Han people. The Liaoning Han people also exhibited clear genetic similarities with each other but relative to Panjin Hans, Dalian Hans had more shared alleles with Neolithic WLRB and ARB populations, and Shenyang Hans shared more alleles with Xianbei-culture individuals in the ARB during the Iron Age (Supplementary Table 8, Figure 5C).

Two-way admixture model based on *Qp*Wave/*Qp*Adm

To further explore potential ancestral sources and determine the portions of different ancestries, we performed the *qp*Wave/*qp*Adm method. We found clear genetic differences among northern and southern Han populations, and there was more southern ancestry and less WLRB/YRB/ARB-related ancestry in southern Han populations than in northern Han people. Relative to Shandong and Shanxi Han peoples, Liaoning Hans had more southern ancestry, and the genetic profile of Henan Hans was more similar to that of Liaoning Hans (Figure 6). We found the same genetic makeup between the Liaoning Han people and Liaoning Manchu people; Liaoning Mongolians had more ancient Northeast Asian-related ancestry and western Eurasian-related ancestry (Figures 6, 7). In comparison with Tungusic speakers, the Liaoning Han people have more southern ancestry and less western Eurasian-related ancestry (Figure 7). We also found no significant genetic differences among the Liaoning Han people (Figure 8).

Phylogenetic framework with gene flow events constructed by the *Qp*Graph and TreeMix methods

In the *qp*Graph-based phylogenetic framework with gene flow events, Mbuti, Denisovan, Loschbour, GreatAndaman, and Tianyuan were used to construct the basal model, and Neolithic hunter-gatherers on the Mongolian Plateau (MP) and in the ARB, millet farmers related to the Qijia, Yangshao and Lower Xiajiadian cultures in the YRB and WLRB, Neolithic Qihe, Iron Age Hanben, Bronze Age Afanasievo pastoralists, and Kostenki14 and Ishim_Ust from Russia were used as different potential ancestral contributors (Figure 9, Supplementary Table 9). We found that Liaoning Han people have 57–98% Neolithic millet farmer-related ancestry, 40–43% ancient Northeast Asian-related ancestry, and 2% western Eurasian-related ancestry.

In the TreeMix analysis (Figure 10), no gene flow events related to the Liaoning Han populations were found, and the phylogenetic tree based on TreeMix showed that Liaoning Han people clustered with other northern Han populations. Dandong Hans, Panjin Hans, and Shenyang Hans clustered with Henan Hans, Shanxi Hans, and Liaoning Manchus and Dalian Hans clustered with Shandong Hans showed a close relationship with Japanese and Koreans.

10.3389/fevo.2022.1014024

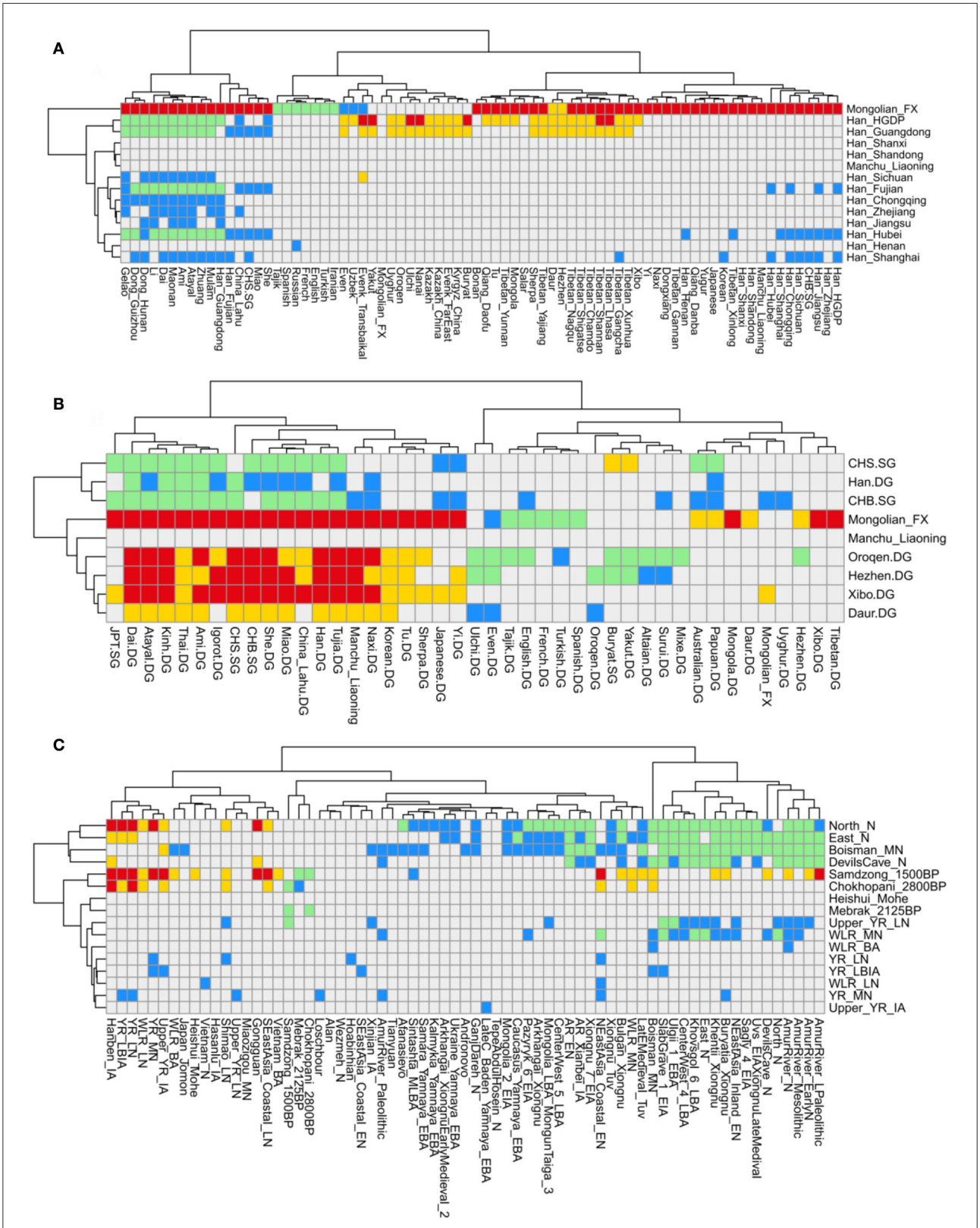


FIGURE 5
Shared allele and gene flow direction explored *via* f_4-statistics analysis. **(A)** f_4 (Modern population1, Han_Liaoning; Modern population2, Mbuti) based on the "Merged-HO" dataset; **(B)** f_4 (Modern population1, Han_Liaoning; Modern population2, Mbuti) based on the "Merged-1240K" dataset; **(C)** f_4 (Ancient population1, Han_Liaoning; Ancient population2, Mbuti) based on the "Merged-1240K" dataset.

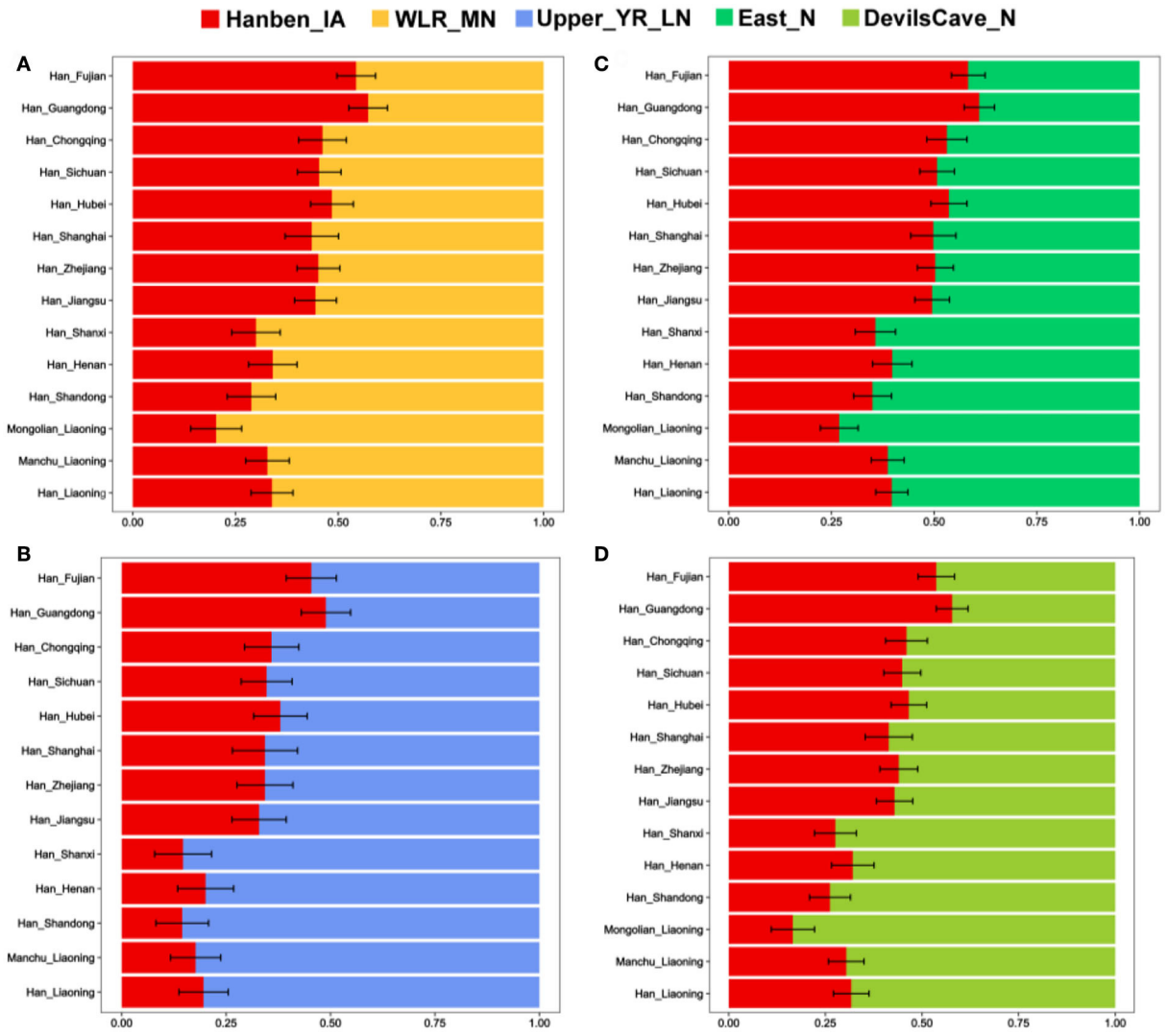


FIGURE 6
(A–D) Two-way admixture model among different Han populations constructed based on the "Merged-HO" dataset.

Admixture time and uniparental lineages of Liaoning Han people

We next used the ALDER method based on weighted linkage disequilibrium statistics to estimate admixture time. We found that western Eurasian-related ancestry flowed into the gene pool of Liaoning Hans populations 4,500–1,200 years in the past, northern East Asian- and southern Siberia-related ancestry flowed into the gene pool of Liaoning Hans populations 4,500–1,300 years in the past, and southern East Asian-related ancestry flowed into the gene pool of Liaoning Hans populations approximately 2,000–1,100 years in the past (Supplementary Table 10).

In this study, we successfully identified 40 paternal haplogroups and 80 maternal haplogroups in the Liaoning Han people, which are listed in Supplementary Table 11. We found that D4, D5, and B4 are the most frequent maternal haplogroups, and O1b1a2a1, C2c1a2a2, O2a2b1a1a1a1, and O2a2b1a2a1a2 are the dominant paternal haplogroups. These paternal and maternal haplogroups are common in East Asia, and they are the main haplogroups of the Han population. The makeup of uniparental lineages also reveals population admixture; O1 and O2 are the dominant paternal haplogroups in Han people, and C2 and N1 are common haplogroups in northern East Asian populations, such as Tungusic and Mongolic speakers (Wei et al., 2018; Wang et al., 2019, 2021a). D4, D5, B4, B5, M7, N9, F1, and A are the common maternal haplogroups in Han people (Li et al., 2019), and G1, G2, Z3, and Z4 are dominant haplogroups in northern East Asian and Siberian populations (Wang et al., 2007; Dryomov et al., 2020).

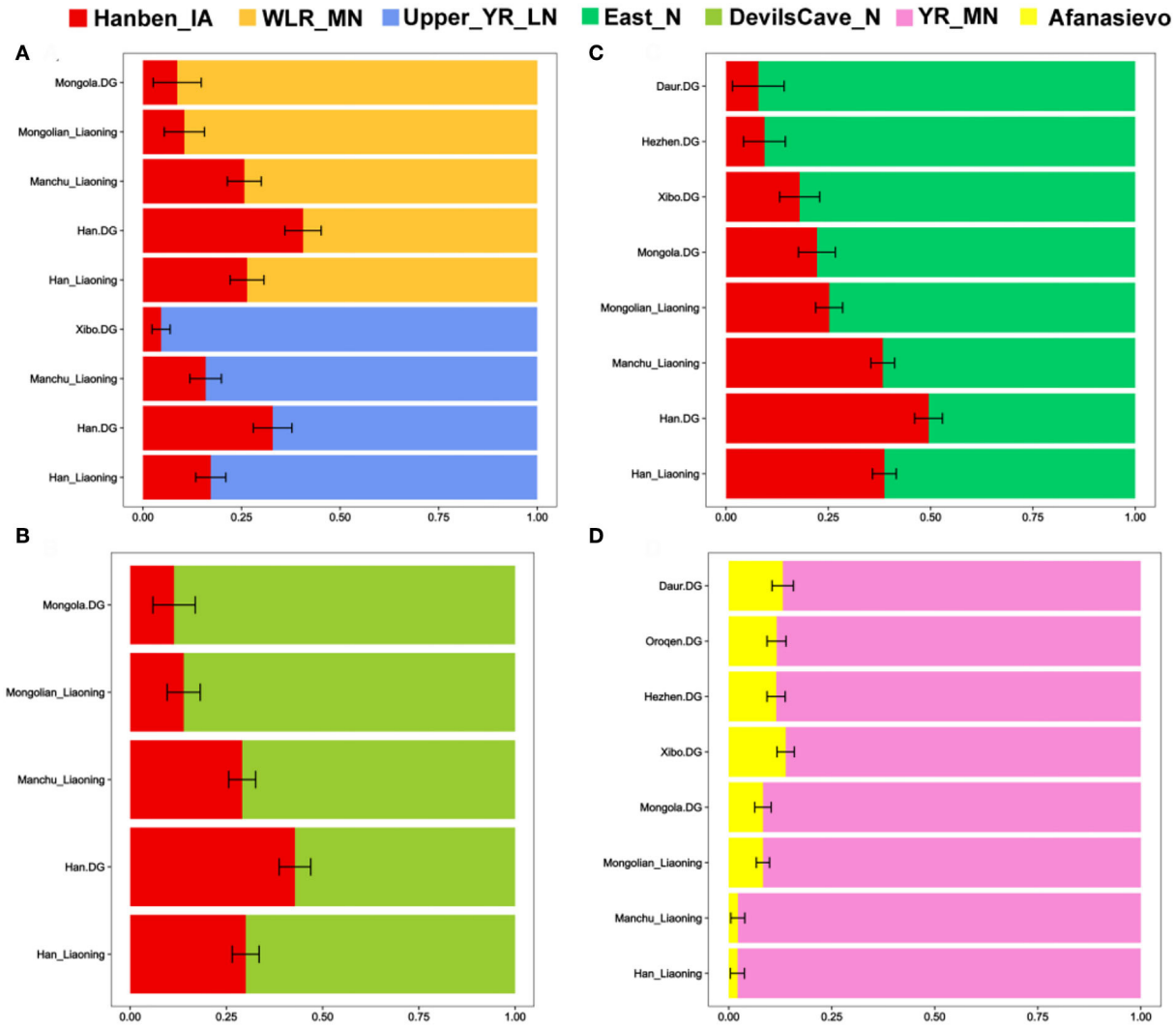


FIGURE 7
(A–D) Two-way admixture model between Liaoning Han and other northern populations constructed based on the "Merged-1240K" dataset.

Discussion

In this study, we explored the genetic structure and affinity and reconstructed the population history of the Liaoning Han people based on genome-wide data. We found that the Liaoning Han people have genetic homogeneity and exhibited significant genetic similarity with the northern Han people (Henan, Shandong, and Shanxi) and Liaoning Manchu people. This finding reveals that the Liaoning Han people and other northern Han people have a common ancestor or origin, and there has been large-scale population admixture between the Han people and Manchu people in Liaoning Province, which further confirms our previous studies' results (Zhang X. et al., 2021), but the Liaoning Han people still exhibit clear genetic differences from other Tungusic speakers, such as Xibo, Oroqen, and Hezhen. Fuxin Mongolians in Liaoning Province show a close genetic relationship with the Liaoning Han people, but there are also clear genetic differences. Fuxin Mongolians have more northern East Asian-related ancestry and Western Eurasian-related ancestry, which indicates that although Mongolians carry clear genetic influences of the Han people, they still retain their own genetic characteristics. In addition, populations that have adjoining geographic locations with the Liaoning Han people also show close genetic distances, such as Japanese and Koreans. Previous studies revealed a coastal expansion route during the Late Pleistocene, the southerly Tianyuan-related lineage and the Onge-related lineage are the dominant ancestral sources of Jomon hunter-gatherers in Japan, and the Tianyuan-related lineage is distributed in ancient populations in the YRB and Yangtze River Basin (Wang C. C. et al., 2021). Robbeets et al.

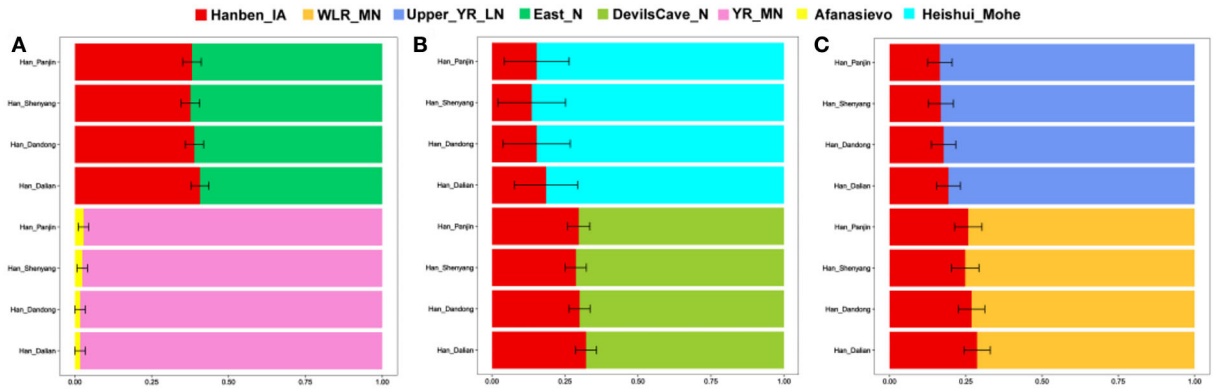


FIGURE 8
(A–C) Two-way admixture model among Liaoning Han constructed based on the "Merged-1240K" dataset.

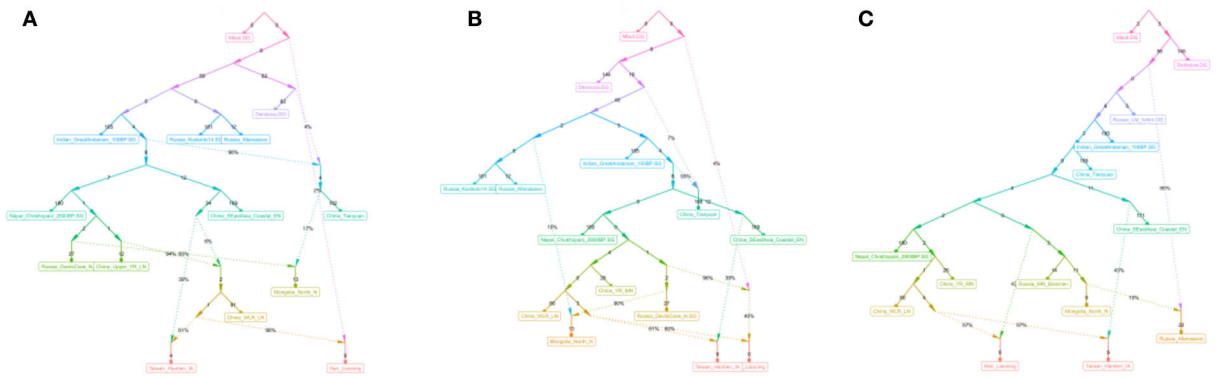


FIGURE 9
(A–C) *qp*Graph-based phylogenetic framework showing the genetic formation of the modern Liaoning Han people.

reported that modern Japanese and Koreans are associated with ancient WLRB populations based on evidence from archeology, language, and genetics, and upper Xiajiadian-related ancestry can be found in modern Japanese and Koreans (Robbeets et al., 2021). Overall, there may be common ancestors who originated from central East Asia among Hans, Japanese, and Koreans, but due to differences in historical development and geographic locations after population divergence, Han people have a much greater expansion speed than Japanese and Koreans based on advanced agriculture, technology and culture, and geographical advantages, and genetic differences among them have gradually appeared (Wang et al., 2018).

Han people are characterized by a north–south genetic cline, and there are many shared components in the gene pool of Han people from different geographic regions, but significant differences can also be detected in these populations. Previous studies indicated that Han people mixed gradually with indigenous residents with their expansion, southern Han people show clear admixture signals with Tai–Kadai and Austronesian-speaking populations (He et al., 2020), southwestern Han people have a close genetic relationship with Hmong–Mien speakers (Wang et al., 2021b), and northwestern Han people show more western Eurasian-related ancestry in their genetic makeup (Yao et al., 2021). In this study, we found that the Liaoning Han people have genetic similarities with the Tungusic Manchu people, and the Liaoning Han people have more YRB-related ancestry and ancient Northeast Asian-related ancestry than the southern Han people. Our results further support the northern origin hypothesis, in which ancient YRB and WLRB populations share more alleles with Liaoning Han people and are the main ancestral contributors, the Neolithic millet-related ancestry is 57–98%, and there is no significant Z score in the f_4 (YR_MN/YR_LN/YR_LBIA/Upper_YR_IA /WLR_MN/WLR_BA, Han_Liaoning; Ancient population, Mbuti) analysis. However, relative to ancient WLRB populations, Liaoning Han people have more YRB-related

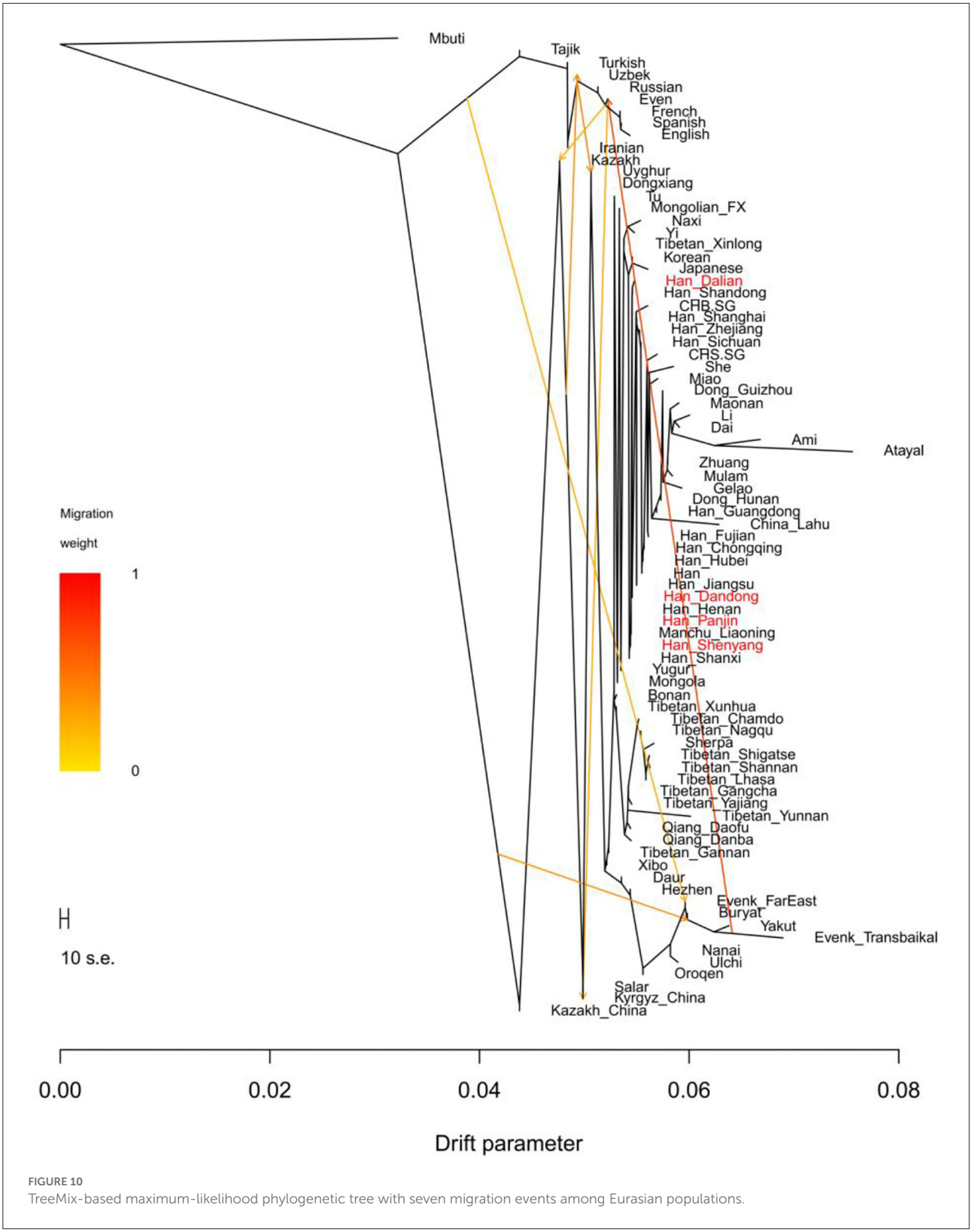


FIGURE 10
TreeMix-based maximum-likelihood phylogenetic tree with seven migration events among Eurasian populations.

ancestry which according for 83%–98%. In addition, 40–43% ancient Northeast Asian-related ancestry could be found in the genetic makeup of Liaoning Hans. Hunter-gatherers on the Mongolian Plateau (MP) and in the ARB are another main genetic contributor. We also found that the Mohe people in the ARB show a close genetic relationship with the Liaoning Han people, there are no significant Z scores in f_4 (Heishui_Mohe, Han_Liaoning; Ancient population, Mbuti) analysis, and Mohe-related ancestry is 82–86%. We observed western Eurasian-related ancestry in the genetic makeup of the Liaoning Han people, and there was approximately 2% western Eurasian-related ancestry. Admixture time based on ALDER methods indicated that western Eurasian-related ancestry and ancient Northeast Asian-related ancestry flowed into the gene pool of Han people 4,500–1,200/1,300 years in the past, from the Xia, Shang, and Zhou dynasties to the Tang dynasty. Paleogenomic studies indicated that northern East Asian-related ancestry has expanded southward since the Neolithic period (Ning et al., 2020; Yang et al., 2020), and western Eurasian-related ancestry has expanded eastward since the early Bronze Age (Wang W. et al., 2021; Zhang F. et al., 2021). Previous studies found that western and northern East Asian pastoralists played an important role in the formation of early China, Chinese culture, and Huaxia people (Sun et al., 2019; Ma et al., 2021). Frequent population exchange and dynamic population history promote population admixture, but based on advanced agriculture, technology, and culture, the Han people or their ancestors often had a greater demographic advantage over ancient ethnic groups in East Asia, so they often assimilated with the population and culture of other ethnic groups. Southern East Asian-related ancestry flowed into Liaoning Hans' gene pool approximately 2,000–1,100 years in the past from the Eastern Han Dynasty to the Tang Dynasty. According to historical records, there were many southward expansions of Han people from the central plain, and Han people mixed gradually with indigenous residents during their migration (Wen et al., 2004; He et al., 2022a). In addition, the northward expansions of southern East Asian populations also occurred during the historical period; finally, the genetic structure and north–south genetic cline of Han people were constructed.

Although the Liaoning Han people exhibit clear genetic homogeneity, some genetic differences could be found. Dalian Hans show closer genetic affinity with Shandong Hans than Shenyang Hans, and Dalian Hans share more alleles with some Han people (Shandong Hans, Fujian Hans, CHS and CHB), Tibeto–Burman speakers (Naxi, Yi, Lhasa and Yajiang Tibetans), Liaoning Manchu people, Liaoning Mongolians, and Ami than Panjin Hans (Supplementary Table 7). Relative to Panjin Hans, Dalian Hans share more alleles with the Neolithic WLRB and ARB populations, and Shenyang Hans share more alleles with Xianbei individuals in the ARB (Supplementary Table 8). Two-way admixture models based on the *qp*Adm method indicated that the Liaoning Han population has a genetic makeup similar to that of the Henan Han population (Figure 6). The Dalian Han people have less millet farmer-related ancestry, hunter-gatherer-related ancestry, and western Eurasian-related ancestry than other Liaoning Han people (Figure 8). In the phylogenetic framework of East Asian populations based on the TreeMix analysis, we found that the Dalian Han populations clustered with Shandong Han populations and had a close genetic relationship with Japanese and Koreans, but Dandong Hans, Panjin Hans, and Shenyang Hans clustered with Henan Hans, Shanxi Hans, and Liaoning Manchus (Figure 10). According to historical records, "Chuang Guandong" was a large-scale population migration event that occurred beginning in 1877. More than 30 million people moved to Northeast China from modern Shandong, Hebei, and Henan Provinces, and Han people were the main ethnic group in this population migration. They played an indispensable role in the formation of the population structure of northern East Asians. According to historical records, the Shandong Han people arrived mainly in Northeast China by sea and mainly resided in Dalian in Liaoning Province, and Han people from Hebei and Henan Provinces migrated to Northeast China by land and were distributed mainly in the northern and western regions of Liaoning Province. In general, we propose that the genetic differences among Liaoning Han people are associated with migration routes during the "Chuang Guandong" period in historical records. We further reconstructed the demographic history of the Liaoning Han people based on genome-wide data, but more studies from different perspectives are needed to test this hypothesis.

Conclusion

In this study, we used typical and advanced population genetic analysis methods (PCA, ADMIXTURE, F_{ST}, f_3-statistics, f_4-statistics, *qp*Adm/*qp*Wave, *qp*Graph, ALDER and TreeMix) to explore the genetic structure of 80 Han individuals from four different cities in Liaoning Province and reconstruct their demographic history based on newly generated genome-wide data. We found that the Liaoning Han people have genetic similarities with other northern Han people (Shandong, Henan, and Shanxi) and Liaoning Manchu people. Millet farmers in the YRB and WLRB (57–98%) and hunter-gatherers on the Mongolian Plateau (MP) and in the ARB (40–43%) are the main ancestral sources of the Liaoning Han people. Our study further supports the northern origin hypothesis; the YRB-related ancestry accounted for 83–98% of the genetic makeup of the Liaoning Han population. There are clear genetic influences of northern East Asian populations in the Liaoning Han people, ancient Northeast Asian-related ancestry is another dominant ancestral component, and large-scale population

admixture has occurred between the Tungusic Manchu people and Han people. Interestingly, there are genetic differences among the Liaoning Han people, and we found that these differences are associated with different migration routes of the Han people during the Chuang Guandong period in historical records.

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: doi: 10.5281/zenodo.7068554.

Ethics statement

The studies involving human participants were reviewed and approved by the Medical Ethics Committee of Jinzhou Medical University (JZMULL2021101). The patients/participants provided their written informed consent to participate in this study.

Author contributions

YW designed this study and revised the manuscript. XianZ analyzed the data. JZ wrote this manuscript. XL and JS performed the experiment. SZ, HZ, QZ, and XiaoZ collected the samples. HH provided computer technical assistance. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Human Phenome Laboratory Project of Liaoning Province.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.1014024/full#supplementary-material

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genom. Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Bin, X., Wang, R., Huang, Y., Wei, R., Zhu, K., Yang, X., et al. (2021). Genomic insight into the population structure and admixture history of tai-kadai-speaking Sui People in Southwest China. *Front. Genet.* 12, 735084. doi: 10.3389/fgene.2021.735084

Dryomov, S. V., Starikovskaya, E. B., Nazhmidenova, A. M., Morozov, I. V., and Sukernik, R. I. (2020). Genetic legacy of cultures indigenous to the Northeast Asian coast in mitochondrial genomes of nearly extinct maritime tribes. *BMC Evo.l Biol.* 20, 83. doi: 10.1186/s12862-020-01652-1

Du, J., Diao, Y., Rakha, A., Ameen, F., AlKahtani, M. D. F., and Adnan, A. (2020). Forensic applications and genetic characterization of Liaoning Han population revealed by extended set of autosomal STRs. *Mol. Genet. Genom. Med.* 8, e1517. doi: 10.1002/mgg3.1517

Guo, F. (2017a). Population genetic data for 12 X-STR loci in the Northern Han Chinese and StatsX package as tools for population statistics on X-STR. *Forensic Sci. Int. Genet.* 26, e1–e8. doi: 10.1016/j.fsigen.2016.10.012

Guo, F. (2017b). Population genetics for 17 Y-STR loci in Northern Han Chinese from Liaoning Province, Northeast China. *Forensic Sci. Int. Genet*. 29, e35–e37. doi: 10.1016/j.fsigen.2017.04.012

Guo, J., Wang, W., Zhao, K., Li, G., He, G., Zhao, J., et al. (2022). Genomic insights into Neolithic farming-related migrations in the junction of east and southeast *Asia* 177, 328–342. doi: 10.1002/ajpa.24434

He, G., Wang, M., Zou, X., Yeh, H.-Y., Liu, C., Liu, C., et al. (2021). Extensive ethnolinguistic diversity at the crossroads of North China and South Siberia reflects multiple sources of genetic diversity. *J. Syst. Evol*. doi: 10.1111/jse.12827

He, G., Wang, Z., Guo, J., Wang, M., Zou, X., Tang, R., et al. (2020). Inferring the population history of Tai-Kadai-speaking people and southernmost Han Chinese on Hainan Island by genome-wide array genotyping. *Eur. J. Hum. Genet.* 28, 1111–1123. doi: 10.1038/s41431-020-0599-7

He, G.-L., Li, Y.-X., Zou, X., Yeh, H.-Y., Tang, R.-K., Wang, P.-X., et al. (2022a). Northern gene flow into southeastern East Asians inferred from genome-wide array genotyping. *n/a(n/a)* doi: 10.1111/jse.12826

He, G.-L., Wang, M.-G., Li, Y.-X., Zou, X., Yeh, H.-Y., Tang, R.-K., et al. (2022b). Fine-scale north-to-south genetic admixture profile in Shaanxi Han Chinese revealed by genome-wide demographic history reconstruction. *J. Syst. Evol.* 60, 955–972. doi: 10.1111/jse.12715

Helsinki, W. M. A. D. (2001). World medical association declaration of helsinki. Ethical principles for medical research involving human subjects. *Bull. World Health Organ.* 79, 373–374.

Hou, X., Zhang, X., Li, X., Huang, T., Li, W., Zhang, H., et al. (2022). Genomic insights into the genetic structure and population history of Mongolians in Liaoning Province. *Front. Genet.* 13, 947758. doi: 10.3389/fgene.2022.947758

Li, Y. C., Ye, W. J., Jiang, C. G., Zeng, Z., Tian, J. Y., Yang, L. Q., et al. (2019). River valleys shaped the maternal genetic landscape of Han Chinese. *Mol. Biol. Evol.* 36, 1643–1652. doi: 10.1093/molbev/msz072

Loh, P. R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., et al. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193, 1233–1254. doi: 10.1534/genetics.112.147330

Luo, T., Wang, R., and Wang, C. C. (2021). Inferring the population structure and admixture history of three Hmong-Mien-speaking Miao tribes from southwest China based on genome-wide SNP genotyping. *Ann. Hum. Biol.* 48, 418–429. doi: 10.1080/03014460.2021.2005825

Ma, P., Yang, X., Yan, S., Li, C., Gao, S., Han, B., et al. (2021). Ancient Y-DNA with reconstructed phylogeny provides insights into the demographic history of paternal haplogroup N1a2-F1360. *J. Genet. Genom.* 48, 1130–1133. doi: 10.1016/j.jgg.2021.07.018

Mao, X., Zhang, H., Qiao, S., Liu, Y., Chang, F., Xie, P., et al. (2021). The deep population history of northern East Asia from the Late Pleistocene to the Holocene. *Cell* 184, 3256–3266.e3213. doi: 10.1016/j.cell.2021.04.040

Ning, C., Li, T., Wang, K., Zhang, F., Li, T., Wu, X., et al. (2020). Ancient genomes from northern China suggest links between subsistence changes and human migration. *Nat. Commun.* 11, 2700. doi: 10.1038/s41467-020-16557-2

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* 192, 1065–1093. doi: 10.1534/genetics.112.145037

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2, e190. doi: 10.1371/journal.pgen.0020190

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

Robbeets, M., Bouckaert, R., Conte, M., Savelyev, A., Li, T., An, D. I., et al. (2021). Triangulation supports agricultural spread of the Transeurasian languages. *Nature* 599, 616–621. doi: 10.1038/s41586-021-04108-8

Robbeets, M., and Savelyev, A. (2020). *The Oxford Guide to the Transeurasian Languages.* Oxford: Oxford University Press.

Sagart, L., Jacques, G., Lai, Y., Ryder, R. J., Thouzeau, V., Greenhill, S. J., et al. (2019). Dated language phylogenies shed light on the ancestry of Sino–Tibetan. *Proc. Natl. Acad. Sci. U.S.A.* 116, 10317–10322. doi: 10.1073/pnas.1817972116

Sun, N., Ma, P. C., Yan, S., Wen, S. Q., Sun, C., Du, P. X., et al. (2019). Phylogeography of Y-chromosome haplogroup Q1a1a-M120, a paternal lineage connecting populations in Siberia and East Asia. *Ann. Hum. Biol.* 46, 261–266. doi: 10.1080/03014460.2019.1632930

Wang, C. C., Yeh, H. Y., Popov, A. N., Zhang, H. Q., Matsumura, H., Sirak, K., et al. (2021). Genomic insights into the formation of human populations in East Asia. *Nature* 591, 413–419. doi: 10.1038/s41586-021-03336-2

Wang, C. Z., Wei, L. H., Wang, L. X., Wen, S. Q., Yu, X. E., Shi, M. S., et al. (2019). Relating Clans Ao and Aisin Gioro from northeast China by whole Y-chromosome sequencing. *J. Hum. Genet.* 64, 775–780. doi: 10.1038/s10038-019-0622-4

Wang, H., Ge, B., Mair, V. H., Cai, D., Xie, C., Zhang, Q., et al. (2007). Molecular genetic analysis of remains from Lamadong cemetery, Liaoning, China. *Am. J. Phys. Anthropol.* 134, 404–411. doi: 10.1002/ajpa.20685

Wang, M., He, G., Zou, X., Chen, P.-Y., Wang, Z., Tang, R., et al. (2022). Reconstructing the genetic admixture history of Tai-Kadai and Sinitic people: insights from genome-wide SNP data from South China. *J. Syst. Evol.* doi: 10.1111/jse.12825

Wang, M., He, G., Zou, X., Liu, J., Ye, Z., Ming, T., et al. (2021a). Genetic insights into the paternal admixture history of Chinese Mongolians via high-resolution customized Y-SNP SNaPshot panels. *Forensic Sci. Int. Genet.* 54, 102565. doi: 10.1016/j.fsigen.2021.102565

Wang, M., Yuan, D., Zou, X., Wang, Z., Yeh, H. Y., Liu, J., et al. (2021b). Fine-scale genetic structure and natural selection signatures of southwestern hans inferred from patterns of genome-wide allele, haplotype, and haplogroup lineages. *Front. Genet.* 12, 727821. doi: 10.3389/fgene.2021.727821

Wang, W., Ding, M., Gardner, J.D., Wang, Y., Miao, B., Guo, W., et al. (2021). Ancient Xinjiang mitogenomes reveal intense admixture with high genetic diversity. *Sci. Adv.* 7, eabd6690. doi: 10.1126/sciadv.abd6690

Wang, Y., Lu, D., Chung, Y. J., and Xu, S. (2018). Genetic structure, divergence and admixture of Han Chinese, Japanese and Korean populations. *Hereditas* 155, 19. doi: 10.1186/s41065-018-0057-5

Wei, L. H., Yan, S., Lu, Y., Wen, S. Q., Huang, Y. Z., Wang, L. X., et al. (2018). Whole-sequence analysis indicates that the Y chromosome C2*-Star Cluster traces back to ordinary Mongols, rather than Genghis Khan. *Eur. J. Hum. Genet.* 26, 230–237. doi: 10.1038/s41431-017-0012-3

Wen, B., Li, H., Lu, D., Song, X., Zhang, F., He, Y., et al. (2004). Genetic evidence supports demic diffusion of Han culture. *Nature* 431, 302–305. doi: 10.1038/nature02878

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yang, M. A., Fan, X., Sun, B., Chen, C., Lang, J., Ko, Y. C., et al. (2020). Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* 369, 282–288. doi: 10.1126/science.aba0909

Yao, H., Wang, M., Zou, X., Li, Y., Yang, X., Li, A., et al. (2021). New insights into the fine-scale history of western-eastern admixture of the northwestern Chinese population in the Hexi Corridor via genome-wide genetic legacy. *Mol. Genet. Genom.* 296, 631–651. doi: 10.1007/s00438-021-01767-0

Yao, J., and Wang, B. J. (2016). Genetic Variation of 25 Y-Chromosomal and 15 Autosomal STR Loci in the Han chinese population of Liaoning Province, Northeast China. *PLoS ONE* 11, e0160415. doi: 10.1371/journal.pone.0160415

Yao, J., Wang, L. M., Gui, J., Xing, J. X., Xuan, J. F., and Wang, B. J. (2016). Population data of 15 autosomal STR loci in Chinese Han population from Liaoning Province, Northeast China. *Forensic Sci. Int. Genet.* 23, e20–e21. doi: 10.1016/j.fsigen.2016.04.012

Zhang, F., Ning, C., Scott, A., Fu, Q., Bjørn, R., Li, W., et al. (2021). The genomic origins of the Bronze Age Tarim Basin mummies. *Nature* 599, 256–261. doi: 10.1038/s41586-021-04052-7

Zhang, M., Yan, S., Pan, W., and Jin, L. (2019). Phylogenetic evidence for Sino–Tibetan origin in northern China in the Late Neolithic. *Nature* 569, 112–115. doi: 10.1038/s41586-019-1153-z

Zhang, X., He, G., Li, W., Wang, Y., Li, X., Chen, Y., et al. (2021). Genomic insight into the population admixture history of tungusic-speaking Manchu people in Northeast China. *Front. Genet.* 12, 754492. doi: 10.3389/fgene.2021.754492

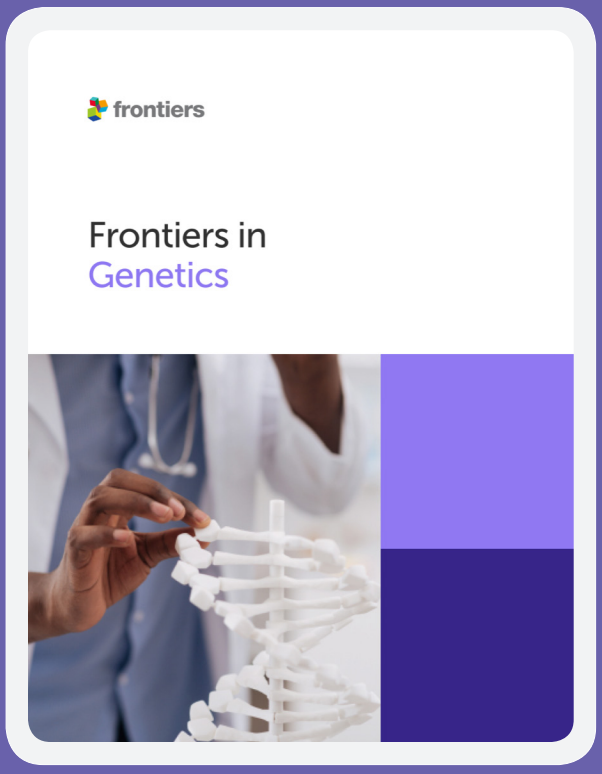
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact