



CURRENT AND EMERGING TRENDS IN HUMAN IDENTIFICATION AND MOLECULAR ANTHROPOLOGY

EDITED BY: Ozlem Bulbul, Cemal Gurkan and Kenneth K. Kidd
PUBLISHED IN: Frontiers in Genetics



frontiers

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 978-2-88966-913-4

DOI 10.3389/978-2-88966-913-4

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

CURRENT AND EMERGING TRENDS IN HUMAN IDENTIFICATION AND MOLECULAR ANTHROPOLOGY

Topic Editors:

Ozlem Bulbul, Istanbul University- Cerrahpasa, Turkey

Cemal Gurkan, Turkish Cypriot DNA Laboratory (TCDL), Cyprus

Kenneth K. Kidd, Yale University, United States

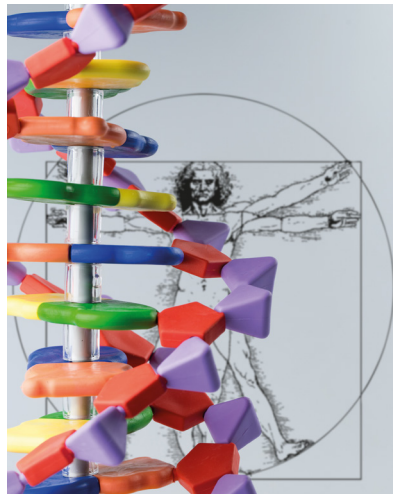


Image: alice-photo/Shutterstock.com

Citation: Bulbul, O., Gurkan, C., Kidd, K. K., eds. (2021). Current and Emerging Trends in Human Identification and Molecular Anthropology. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-913-4

Table of Contents

- 05 Editorial: Current and Emerging Trends in Human Identification and Molecular Anthropology**
Cemal Gurkan, Ozlem Bulbul and Kenneth K. Kidd
- 08 Forensic Autosomal Short Tandem Repeats and Their Potential Association With Phenotype**
Nicole Wyner, Mark Barash and Dennis McNevin
- 15 MicroHapDB: A Portable and Extensible Database of All Published Microhaplotype Marker and Frequency Data**
Daniel S. Standage and Rebecca N. Mitchell
- 26 A Comparison of Forensic Age Prediction Models Using Data From Four DNA Methylation Technologies**
A. Freire-Aradas, E. Pośpiech, A. Aliferi, L. Girón-Santamaría, A. Mosquera-Miguel, A. Pisarek, A. Ambroa-Conde, C. Phillips, M. A. Casares de Cal, A. Gómez-Tato, M. Spólnicka, A. Woźniak, J. Álvarez-Dios, D. Ballard, D. Syndercombe Court, W. Branicki, Ángel Carracedo and M. V. Lareu
- 38 Evaluation of the Precision of Ancestry Inferences in South American Admixed Populations**
Vania Pereira, Roberta Santangelo, Claus Børsting, Torben Tvedebrink, Ana Paula F. Almeida, Elizeu F. Carvalho, Niels Morling and Leonor Gusmão
- 58 Twenty Years Later: A Comprehensive Review of the X Chromosome Use in Forensic Genetics**
Iva Gomes, Nádia Pinto, Sofia Antão-Sousa, Verónica Gomes, Leonor Gusmão and António Amorim
- 75 Development and Validation of a Novel Five-Dye Short Tandem Repeat Panel for Forensic Identification of 11 Species**
Wei Cui, Xiaoye Jin, Yuxin Guo, Chong Chen, Wenqing Zhang, Yijie Wang, Jiangwei Lan and Bofeng Zhu
- 85 Broadening the Applicability of a Custom Multi-Platform Panel of Microhaplotypes: Bio-Geographical Ancestry Inference and Expanded Reference Data**
María de la Puente, Jorge Ruiz-Ramírez, Adrián Ambroa-Conde, Catarina Xavier, Jorge Amigo, María Ángeles Casares de Cal, Antonio Gómez-Tato, Ángel Carracedo, Walther Parson, Christopher Phillips and María Victoria Lareu
- 97 Multi-Indel: A Microhaplotype Marker Can Be Typed Using Capillary Electrophoresis Platforms**
Shengqiu Qu, Meili Lv, Jiaming Xue, Jing Zhu, Li Wang, Hui Jian, Yuqing Liu, Ranran Zhang, Lagabaiyila Zha, Weibo Liang and Lin Zhang

110 *From Identification to Intelligence: An Assessment of the Suitability of Forensic DNA Phenotyping Service Providers for Use in Australian Law Enforcement Casework*

Lauren Atwood, Jennifer Raymond, Alison Sears, Michael Bell and Runa Daniel

121 *Combined Low-/High-Density Modern and Ancient Genome-Wide Data Document Genomic Admixture History of High-Altitude East Asians*

Yan Liu, Mengge Wang, Pengyu Chen, Zheng Wang, Jing Liu, Lilan Yao, Fei Wang, Renkuan Tang, Xing Zou and Guanglin He



Editorial: Current and Emerging Trends in Human Identification and Molecular Anthropology

Cemal Gurkan^{1,2*}, Ozlem Bulbul³ and Kenneth K. Kidd⁴

¹ Turkish Cypriot DNA Laboratory, Committee on Missing Persons in Cyprus Turkish Cypriot Member's Office, Nicosia, Turkey, ² Dr. Fazıl Küçük Faculty of Medicine, Eastern Mediterranean University, Famagusta, Turkey, ³ Institute of Forensic Science, Istanbul University-Cerrahpasa, Istanbul, Turkey, ⁴ Department of Genetics, Yale University School of Medicine, New Haven, CT, United States

Keywords: ancestry informative markers, forensic DNA phenotyping, age prediction, microhaplotypes, genotyping

Editorial on the Research Topic

Current and Emerging Trends in Human Identification and Molecular Anthropology

Recent developments in the DNA analysis technologies such as Massively Parallel Sequencing (MPS) have found prolific use in forensic applications. MPS not only allows the identification of a given individual using traditional forensic genetic markers but also the prediction of his/her age, appearance, and ancestry (Børsting and Morling, 2015). This can be very useful when conventional identification attempts using the comparative DNA profiling systems [e.g., short tandem repeats (STRs)] reach a dead end due to the absence of requisite DNA reference samples and/or potential matches in the respective databases. DNA-based estimation of ancestry and physical characteristics, now collectively described as forensic DNA phenotyping (FDP), has also led to the emergence of the concept of “biological witness” that can potentially provide investigative leads in such cases (Kayser, 2015). For instance, the prediction of the eye and hair color of King Richard III of England (1452–1485) and World War II victims were all made possible with FDP (King et al., 2014; Chaitanya et al., 2017). A growing number of allele frequency data for diverse reference populations has facilitated increasingly better estimates of individual biogeographic ancestry (Pakstis et al., 2015, 2019; Bulbul et al., 2018). In other words, FDP may help us better understand the recent or distant past *via* the “reconstruction” of otherwise unidentifiable human remains. Many facets of human identification in a forensic context overlap with various aspects of human population genetics and molecular anthropology. Accordingly, in this Research Topic, we sought manuscripts that would provide a snapshot of the current and emerging trends in human identification or molecular anthropology, especially at the interface of the two fields (**Figure 1**).

A significant new trend in human identification and molecular anthropology is the emergence of a new genetic marker called *microhaplotypes*. Since the introduction of the concept less than a decade ago (Kidd et al., 2013), these very short segments (<300 bp) of the genome with multiple variants such that the loci are multiallelic have been getting increasing attention (Oldoni et al., 2019). Use of microhaplotypes is exemplified by de la Puente et al. who present a validation of a panel of 113 loci showing their power for biogeographic ancestry inference as well as familial relationship detection. Furthermore, Qu et al. have broadened the concept of the microhaplotype to include multiple small insertions/deletions allowing the markers to be studied using a conventional method like capillary electrophoresis (CE). Accordingly, these new microhaplotype loci may provide a useful supplement to standard CE-based STR typing using the same equipment. The growing number of published microhaplotype loci and the accumulating population data on them has been summarized in a new database called MicroHapDB by Standage and Mitchell. This

OPEN ACCESS

Edited and reviewed by:

Denis Baurain,
University of Liège, Belgium

*Correspondence:

Cemal Gurkan
cemal.gurkan@emu.edu.tr

Specialty section:

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

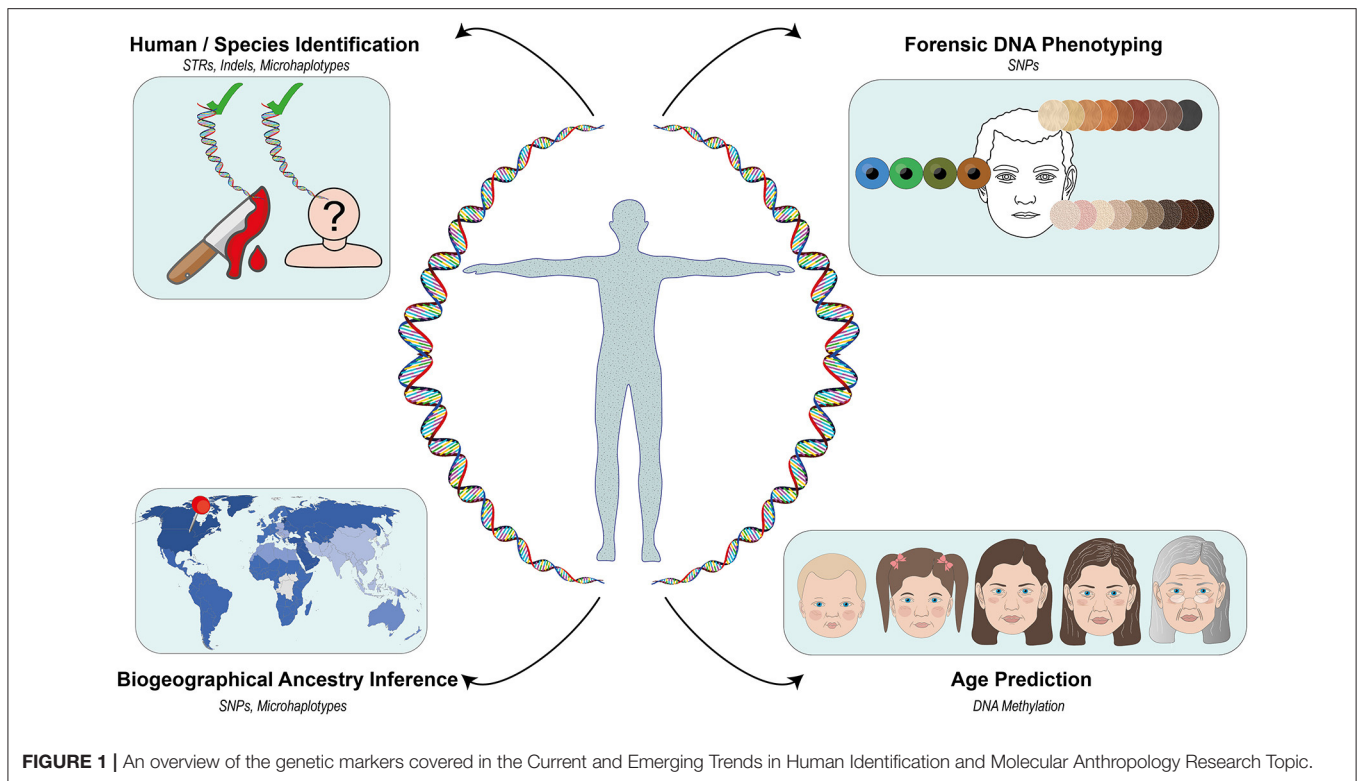
Received: 11 May 2021

Accepted: 24 May 2021

Published: 23 June 2021

Citation:

Gurkan C, Bulbul O and Kidd KK
(2021) Editorial: Current and Emerging
Trends in Human Identification
and Molecular Anthropology.
Front. Genet. 12:708222.
doi: 10.3389/fgene.2021.708222



extensible database will certainly be an important resource as the trend for more microhaplotype studies continues.

Another emerging trend is the gradual introduction of FDP in more routine casework investigations, and to this end, Atwood et al., report the selection process for the provision of FDP services to be used by the Australian law enforcement agencies. Here, a comparison of FDP services offered by six different providers was carried out that led to the successful selection of a provider for the prediction of biogeographical ancestry, hair, and eye color. A parallel trend in forensic research, which can also be regarded complementary to FDP, is the use of molecular techniques to determine the age of the person who left a DNA sample (Naue et al., 2018). Freire-Aradas et al. present a methodological comparison of the current DNA-methylation based forensic age prediction models using four different technologies available, and report largely comparable age predictions results. In fact, DNA methylation-based investigations are part of the growing field of forensic epigenomics whereby it has already been shown to distinguish homo-zygotic twins and the source tissue of a given sample, but also in the near future, even hold the potential for the prediction of lifestyle and environmental exposures of unknown perpetrators (Vidaki and Kayser, 2017).

Of course, studies on more traditional markers such STRs is still ongoing with several new applications. Cui et al. have taken STRs beyond humans to identify loci that can be used to identify different species of interest in a given sample. Their panel can identify DNA samples from 10 different species in

addition to human using pairs of species-specific STR loci, and as such it may not only find use in forensic species identification, but also in the detection of meat fraud and adulteration. Gomes et al. focus on the unique inheritance and population genetic characteristics of markers located on the X chromosome. While noting an unexpected decrease in the number of new forensic investigations in the literature over the last two decades using X chromosome markers, nevertheless, the authors show the utility of these markers, primarily STRs, that so far have been identified. On a different note, Wyner et al. question the validity of the common assumption that the standard forensic STR loci are not associated with any phenotypes. In the light of increasing amount of compelling evidence toward the associations between forensic STR loci and certain phenotypes, authors point out to the presence of numerous legal and ethical implications associated with the already accumulated and expanding data using these markers, and suggest follow-ups and appropriate counter measures to minimize any misuse of such additional information.

SNPs continue to be investigated for their ability to infer ancestry of individuals and genetic affinities between populations. Pereira et al. have studied one admixed population in Brazil and shown dependencies in assessing admixture in ancestry of individuals on the numbers of markers used and the criteria used to identify the markers. Liu et al. have studied many populations in East Asia to assess the ancestry of Tibetans and their relationships to other Asian populations.

AUTHOR CONTRIBUTIONS

CG, OB, and KK co-edited the Research Topic and co-wrote, co-edited, and co-approved the final version of the Editorial for publication. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The guest editors wish to thank all the authors and peer-reviewers for their valuable contributions to this Research Topic, Huseyin Sevay for the rendering of double helix DNA structure in **Figure 1** and Sumeyye Zual Simsek for general help in putting together **Figure 1**.

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
- Bulbul, O., Speed, W. C., Gurkan, C., Soundararajan, U., Rajeevan, H., Pakstis, A. J., et al. (2018). Improving ancestry distinctions among Southwest Asian populations. *Forensic Sci. Int. Genet.* 35, 14–20. doi: 10.1016/j.fsigen.2018.03.010
- Chaitanya, L., Pajnič, I. Z., Walsh, S., Balažic, J., Zupanc, T., and Kayser, M. (2017). Bringing colour back after 70 years: predicting eye and hair colour from skeletal remains of World War II victims using the HIRisPlex system. *Forensic Sci. Int. Genet.* 26, 48–57. doi: 10.1016/j.fsigen.2016.10.004
- Kayser, M. (2015). Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci. Int. Genet.* 18. doi: 10.1016/j.fsigen.2015.02.003
- Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., et al. (2013). Microhaplotype loci are a powerful new type of forensic marker. *Forensic Sci. Int. Genet. Suppl. Ser.* 4, e123–e124. doi: 10.1016/j.fsigss.2013.10.063
- King, T. E., Fortes, G. G., Balaesque, P., Thomas, M. G., Balding, D., Delser, P. M., et al. (2014). Identification of the remains of King Richard III. *Nat. Commun.* 5, 1–8. doi: 10.1038/ncomms6631
- Naue, J., Sanger, T., Hoefslot, H. C. J., Lutz-Bonengel, S., Kloosterman, A. D., and Verschure, P. J. (2018). Proof of concept study of age-dependent DNA methylation markers across different tissues by massive parallel sequencing. *Forensic Sci. Int. Genet.* 36, 152–159. doi: 10.1016/j.fsigen.2018.07.007
- Oldoni, F., Kidd, K. K., and Podini, D. (2019). Microhaplotypes in forensic genetics. *Forensic Sci. Int. Genet.* 38, 54–69. doi: 10.1016/j.fsigen.2018.09.009
- Pakstis, A. J., Gurkan, C., Dogan, M., Balkaya, H. E., Dogan, S., Neophytou, P. I., et al. (2019). Genetic relationships of European, Mediterranean, and SW Asian populations using a panel of 55 AISNPs. *Eur. J. Hum. Genet.* 27, 1885–1893. doi: 10.1038/s41431-019-0466-6
- Pakstis, A. J., Haigh, E., Cherni, L., Elgaaied, A. B. A., Barton, A., Evsanaa, B., et al. (2015). 52 additional reference population samples for the 55 AISNP panel. *Forensic Sci. Int. Genet.* 19, 269–271. doi: 10.1016/j.fsigen.2015.08.003
- Vidaki, A., and Kayser, M. (2017). From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. *Genome Biol.* 18, 1–13. doi: 10.1186/s13059-017-1373-1

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

Copyright © 2021 Gurkan, Bulbul and Kidd. 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.



Forensic Autosomal Short Tandem Repeats and Their Potential Association With Phenotype

Nicole Wyner^{1*}, Mark Barash^{1,2} and Dennis McNevin¹

¹ Centre for Forensic Science, School of Mathematical and Physical Sciences, Faculty of Science, University of Technology Sydney, Sydney, NSW, Australia, ² Department of Justice Studies, San José State University, San Jose, CA, United States

OPEN ACCESS

Edited by:

Kenneth K. Kidd,
Yale University, United States

Reviewed by:

Jianye Ge,
The University of North Texas Health
Science Center at Fort Worth,
United States
Adrian Matthew Linacre,
Flinders University, Australia

*Correspondence:

Nicole Wyner
nicole.wyner@alumni.uts.edu.au

Specialty section:

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

Received: 19 May 2020

Accepted: 17 July 2020

Published: 06 August 2020

Citation:

Wyner N, Barash M and
McNevin D (2020) Forensic
Autosomal Short Tandem Repeats
and Their Potential Association With
Phenotype. *Front. Genet.* 11:884.
doi: 10.3389/fgene.2020.00884

Forensic DNA profiling utilizes autosomal short tandem repeat (STR) markers to establish identity of missing persons, confirm familial relations, and link persons of interest to crime scenes. It is a widely accepted notion that genetic markers used in forensic applications are not predictive of phenotype. At present, there has been no demonstration of forensic STR variants directly causing or predicting disease. Such a demonstration would have many legal and ethical implications. For example, is there a duty to inform a DNA donor if a medical condition is discovered during routine analysis of their sample? In this review, we evaluate the possibility that forensic STRs could provide information beyond mere identity. An extensive search of the literature returned 107 articles associating a forensic STR with a trait. A total of 57 of these studies met our inclusion criteria: a reported link between a STR-inclusive gene and a phenotype and a statistical analysis reporting a *p*-value less than 0.05. A total of 50 unique traits were associated with the 24 markers included in the 57 studies. TH01 had the greatest number of associations with 27 traits reportedly linked to 40 different genotypes. Five of the articles associated TH01 with schizophrenia. None of the associations found were independently causative or predictive of disease. Regardless, the likelihood of identifying significant associations is increasing as the function of non-coding STRs in gene expression is steadily revealed. It is recommended that regular reviews take place in order to remain aware of future studies that identify a functional role for any forensic STRs.

Keywords: short tandem repeat, phenotype, forensic marker, DNA profiling, junk DNA, non-coding STRs

INTRODUCTION

Short tandem repeats (STRs) are short repeated sequences of DNA (2–6 bp) that account for approximately 3% of the human genome (Lander et al., 2001). The number of repeat units is highly variable among individuals, which offers a high power of discrimination when analyzed for identification purposes. It is a widely accepted notion that STRs are non-coding in nature and are therefore not implicated in gene expression (Tautz and Schlotterer, 1994; Ramel, 1997; Butler, 2006; Biscotti et al., 2015). There is increasing evidence, however, that non-coding DNA sequences such as STRs may be involved in gene regulation via various mechanisms, hence being associated with phenotype (Sawaya et al., 2013; Chen et al., 2016).

The first STR markers used in forensic casework were selected in 1994 by the Forensic Science Service (FSS) in the United Kingdom for a quadruplex amplification system consisting of four

tetranucleotide STRs—TH01, vWA, FES/FPS, and F13A1 (Kimpton et al., 1994). These markers were deemed suitable for PCR amplification due to their simple repeat sequences and their propensity to display regularly spaced alleles differing by four bases; however, the quadruplex system did not offer a high level of discrimination. In 1997, the Federal Bureau of Investigation (FBI) nominated 13 autosomal STR loci to form the core of the Combined DNA Index System (CODIS), a database consisting of profiles contributed by federal, state, and local forensic laboratories. Two of the markers initially selected by the FSS (vWA and TH01) were included within the core CODIS set, whereas FES/FPS and F13A01 were eventually discarded due to low levels of polymorphism. The core set was reviewed in 2010 with an additional seven STRs being implemented from January 1, 2017. The majority of commercially available DNA profiling kits are manufactured to include the core CODIS STR loci (Butler, 2006). In accordance with the DNA Identification Act of 1994, CODIS is bound by stringent privacy protection protocols, in that the stored DNA samples and subsequent analyses be used strictly for law enforcement identification purposes. The DNA Analysis Backlog Elimination Act of 2000 reaffirms that the markers used for forensic applications were specifically selected because they are not known to be associated with any known physical traits or medical characteristics.

The markers nominated for CODIS were specifically chosen due to their location within non-coding regions of the genome; however, claims that non-coding regions play no functional role have been contested in recent years (Cole, 2007; Kaye, 2007; Sarkar and Adshead, 2010). There is increasing evidence that there may be associations between certain STR alleles and medical conditions (von Wurmb-Schwark et al., 2011; Meraz-Rios et al., 2014). This should not be confused with situations where alleles or loci are diagnostic for medical conditions (e.g., trisomy). Additionally, the ability to infer biogeographical ancestry (BGA) from forensic STRs is possible (Graydon et al., 2009; Algee-Hewitt et al., 2016) with investigators using population-specific STR data as intelligence to guide enquiries (Lowe et al., 2001). BGA is correlated with some phenotypes such as blue eye color in Europeans (Gettings et al., 2014) and lighter skin color with increasing distance from the equator (Relethford, 1997). However, the STR genotype *per se* is not causative of BGA phenotype in any direct sense and is mostly associated with BGA as a result of genetic drift (as STRs for forensic use have been selected to exhibit Hardy Weinberg equilibrium). In the event that any CODIS markers are in future found to be linked to a medical condition or physical trait, the analysis of the DNA sample must still be used only for identification purposes pursuant to the DNA Identification Act of 1994.

Katsanis and Wagner (2013) assessed 24 CODIS loci for phenotypic associations, but found no evidence to support the disclosure of any biomedically relevant information. For example, despite the fact that the locus TH01 was associated with as many as 18 traits: from alcoholism to spinocerebellar ataxia, the authors state that association with these traits does not necessarily imply that individual genotypes are causative or predictive of a particular trait. Following this, a statement issued by the Scientific Working Group of DNA Analysis Methods

[SWGDAM] (2013) restated that although alternate discoveries may be made in the future, current understanding is that the CODIS loci do not reveal any information beyond identity. There has only been one STR to date that has been removed from consideration as a marker used in human identity testing (Szibor et al., 2005). The STR locus HumARA is located within a coding region on the X-chromosome and has been linked to muscular dystrophy. HumARA is a trinucleotide repeat and these are known to be more prone to disease-causing expansions than tetranucleotide repeats (Orr and Zoghbi, 2007; Castet et al., 2010; Hannan, 2018).

MATERIALS AND METHODS

A systematic search of the literature was conducted across three databases (Web of Science, PubMed, and Google Scholar) between August and December 2018. Population data studies, allele frequency studies, validation studies, technique developments, single case reports, mutation analyses, off-ladder allele identification, loss of heterozygosity studies, and locus characterizations were excluded. Additional papers were located by back referencing relevant or similar studies. Following the literature search, each STR was analyzed in the University of California Santa Cruz (UCSC) Genome Browser (Human GRCh38/hg38 Assembly) using the following tracks: Mapping and Sequencing—Base Position-dense; STS Markers-full, Gene and Gene Prediction—GENCODE v29-full; NCBI RefSeq-pack, Phenotype and Literature—OMIM Alleles-full; OMIM Pheno Loci-full; OMIM Genes-full; HGMD Variants-full; GWAS Catalog-full, Regulation—ENCODE Regulation-show; RefSeq Func Elems-full, Variation—Common SNPs(151)-full; FlaggedSNPs(151)-full, Repeats—Microsatellite-full; Simple Repeats-full. The STRs investigated included the 20 CODIS core loci used by the FBI, three extra loci currently used in Australia (Penta E, Penta D, D6S1043), and SE33 which is a core STR in the German national database and has subsequently been incorporated into several European kits.

RESULTS AND DISCUSSION

A total of 57 association studies sourced from three databases met our inclusion criteria: a reported link between a STR-inclusive gene and a phenotype and a statistical analysis reporting a *p*-value less than 0.05. Fifty unique traits were identified across the 24 markers (**Supplementary Table 1**). Schizophrenia was the trait most frequently described with a total of 11 studies reporting data on 14 different polymorphisms potentially associated with eight loci. Two separate articles investigated the allelic frequency amongst people who attempted suicide and reported a significantly higher frequency amongst 10 different alleles of seven forensic loci. The intronic STR TH01 had the greatest number of studies with 26 reports describing 27 traits potentially linked to 40 different genotypes. Five of these studies were investigating a link to schizophrenia, reporting five polymorphisms that are possibly associated with the disease.

No studies associating alleles or genotypes with phenotype were found for Penta E, Penta D, D3S1358, SE33, or D10S1248; however, one study by Shi et al. (2012) investigated the method of diagnosing Down syndrome by testing for a trisomy at the Penta D locus as it is located on chromosome 21. Similarly, six of the 10 articles included for D21S11 were investigating the marker's efficiency in genetic tests for Down syndrome.

Of the 57 articles proposing an association between a forensic STR and a phenotype, none of them confirmed any particular genotype to be solely causative of a phenotype. Despite 13 of the STRs being located within a functional gene, there were no entries in the Online Mendelian Inheritance in Man (OMIM) database relating any STR-inclusive regions of these genes with a disease. A stand-out result is the number of studies reporting an association between a phenotype with polymorphisms at the TH01 locus.

TH01

TH01 is located within the first intron of the tyrosine hydroxylase (TH) gene and is commonly characterized by the repeat motif [AATG]_n or alternatively by the [TCAT]_n motif, according to GenBank top strand nomenclature. TH is the rate-limiting enzyme involved in the biosynthesis of the catecholamines dopamine, epinephrine, and norepinephrine. Catecholamines act as both neurotransmitters and hormones that assist in maintaining homeostasis (Eisenhofer et al., 2004). As such, a strong relationship has been reported in the literature (Eisenhofer et al., 2004; Ng et al., 2015) between variations in the expression of TH and the development of neurological, psychiatric, and cardiovascular diseases.

Previous studies (McEwen, 2002; Antoni et al., 2006; Bastos et al., 2018) have shown that increased levels of epinephrine and norepinephrine are expressed in individuals experiencing acute or chronic stress. Wei et al. (1997) found that individuals carrying the TH01-9 allele showed the highest levels of serum norepinephrine amongst a population of unrelated healthy adults, whereas carriers of the TH01-7 allele showed the lowest. Barbeau et al. (2003) investigated the relationship between the number of TH01 repeats and hemodynamic parameters in subjects at rest and in response to applied stressors. The results of this study indicate that the 6 and 9.3 TH01 alleles are associated with a decrease in the hemodynamic responses to stress, offering a protective effect to individuals carrying those alleles. Carriers of the TH01-6 allele displayed a lower heart rate reactivity when exposed to stressors with increasing age than those without the TH01-6 allele. Furthermore, individuals carrying TH01-9.3 showed no increase in systolic blood pressure in response to stress, whereas those not possessing the TH01-9.3 allele demonstrated a significant increase in systolic blood pressure reactivity with increasing age. Conversely, the TH01-7 allele was found to be detrimental to blood pressure in those with a greater body mass index (BMI). Subjects carrying TH01-7 displayed a higher resting systolic blood pressure as BMI increases and increased heart rate reactivity in response to stressors with increasing BMI.

TH01-7 was also reported to be significantly more prevalent in patients prone to depression (Chiba et al., 2000). The TH01-8

allele was found more frequently in suicide attempters (Persson et al., 1997), individuals with depression (Serretti et al., 1998), and individuals with delusional disorder (Morimoto et al., 2002). Persson et al. (2000) investigated the influence of the number of TH01 repeats on 30 personality dimensions. Subjects possessing the TH01-8 allele scored higher in the neuroticism facets with significant differences observed between individuals displaying anger, hostility and vulnerability (Persson et al., 2000), compared to non-TH01-8 allele carriers. Nine repeats at the TH01 locus were associated with delusional disorder (Morimoto et al., 2002) and extraversion (Tochigi et al., 2006). Furthermore, Yang et al. (2011) conducted a number of association studies in China and reported that the frequency of TH01-9.3 was higher in those displaying suicidal behavior, and TH01-10 was significantly overrepresented in individuals demonstrating violent behavior including sexual assaults (Yang et al., 2010) and in males with impulsive violent behavior (Yang et al., 2013). TH01 was also linked to various disease states such as schizophrenia (Jacewicz et al., 2006b), predisposition to malaria (Gaikwad et al., 2005; Alam et al., 2011), sudden infant death syndrome (SIDS) (Klitsch et al., 2008; Courts and Madea, 2011), and Parkinson's disease (Sutherland et al., 2008).

As previously mentioned, TH catalyzes the conversion of tyrosine to levodopa (L-DOPA) which is then converted to dopamine. Dopamine can be further converted into norepinephrine and epinephrine. *In vitro* experiments have previously demonstrated that TH01 can regulate TH gene transcription, displaying a quantitative silencing effect (Albanese et al., 2001). TH01 alleles inhibited transcription proportionally to the number of repeats. Given that so many vital functions rely on the presence of dopamine and its metabolites (Wei et al., 1997; Meiser et al., 2013), malfunctions of dopaminergic pathways have been associated with the development of numerous psychological diseases (Meiser et al., 2013), and in this review, TH01 was largely connected with schizophrenia (Kurumaji et al., 2001) and Parkinson's disease (Meiser et al., 2013). The longer TH01-9.3 and TH01-10 alleles, predicted to yield less dopamine, were found more frequently in individuals displaying traits indicative of dopaminergic dysfunction such as impulsive violent behavior (Yang et al., 2013), sexual assault (Yang et al., 2010), and addiction (Sander et al., 1998; Anney et al., 2004).

Some contradictory associations were observed between TH01 and certain phenotypes. For instance, De Benedictis et al. (1998) reported a significant association of >9 TH01 repeats with longevity in male Italian centenarians. Contrariwise, von Wurmb-Schwark et al. (2011) were unable to replicate this result when using the same study design on a German population, just as Bediaga et al. (2015) were also unable to confirm an association in a northern Spanish population. Similarly, there are conflicting reports on the association of TH01-9.3 with SIDS across European populations. In 2008, Klitsch et al. (2008) found that the frequency of the TH01-9.3 allele was significantly higher in SIDS patients than in controls in a German population. This association was further confirmed by Courts and Madea (2011). On the contrary, Studer et al. (2014) were unable to replicate this result in a Swiss population. Further

population-based association studies are needed to confirm the existence of associations between TH01 and these phenotypes.

None of the studies investigating TH01 have identified any of the associated genotypes as being causative of disease; therefore, the associations mentioned should only be considered as possible or potential. Many of the traits reported to be associated with TH01 are multifactorial, meaning they are affected by both genes and the environment, such as in the case of Parkinson's disease (Meiser et al., 2013) and schizophrenia (Zhuo et al., 2019).

Potential Associations of Other STR Markers

Schizophrenia is a complex heritable mental health disorder characterized by delusions, hallucinations, and impaired social cognition. It is understood that schizophrenia is polygenic with disease burdening alleles being distributed across multiple loci (Giusti-Rodríguez and Sullivan, 2013; Zhuo et al., 2019). Consistent with this notion, our study revealed that schizophrenia was associated with the greatest number of STRs: FGA, TH01, vWA, D2S441, D2S1338, D8S1179, D16S539, and D18S51. One study (Jacewicz et al., 2006a) found that longer repeats in D18S51 and D2S1338 were significantly more frequent in patients than in controls. This trend is consistent with the expansion of trinucleotide repeats in other major psychiatric disorders. Although the inherent complexity of the disease has posed a challenge to researchers, neurotransmitter abnormalities have long been acknowledged as a major contributing factor in the pathogenesis of schizophrenia (Mäki et al., 2005; Modai and Shomron, 2016).

Genetic mutations alone are not enough to trigger the onset and development of schizophrenia; therefore, further research is required in order to explore how genetic risk factors interact with environmental risk factors in the development, onset, and progression of the condition.

Venous thromboembolism (VTE) is a disorder defined by the occurrence of deep vein thrombosis and/or pulmonary embolism. vWF is a glycoprotein that plays a role in platelet adhesion during coagulation; therefore, it is understood that alterations in serum levels of vWF can contribute to thrombosis disorders (Laird et al., 2007). Meraz-Rios et al. (2014) found that vWA-18, TPOX-9, and TPOX-12 were observed more frequently in individuals with venous thrombosis in the Mexican mestizo population. Furthermore, vWA and TPOX have been associated with chronic myeloid leukemia (Wang et al., 2012).

Trisomys

Down syndrome, or Trisomy-21, can be diagnosed by the presence of a third allele at chromosome 21. This trisomy can be present at any polymorphic marker found on chromosome 21, and there are several studies evaluating the use of D21S11 and Penta D as effective markers in Down syndrome detection (Yoon et al., 2002; Liou et al., 2004; Shi et al., 2012; Guan et al., 2013). Similarly, D18S51 and D13S317 can be used as genetic markers to diagnose the presence of Edwards syndrome (Trisomy-18) and Patau syndrome (Trisomy-13), respectively. Trisomys are an example of a causal association as all individuals with three

chromosomes will be affected. While the presence of an extra allele at chromosomes 13, 18, or 21 does not reveal a medical condition unknown to the donor, it does provide additional identifiable information to investigators.

Cancer

Forensic STRs have been used as genetic markers in several studies to screen for cancer-related alleles. Hui et al. (2014) found that two pairs of alleles (D8S1179-16 with D5S818-13 and D2S1338-23 with D6S1043-11) were found more frequently in gastric cancer patients. Furthermore, a study from China identified a significant association between homozygous alleles at D6S1043 and an increased risk of invasive cervical cancer (Wu et al., 2008). Loss of heterozygosity (LOH) is a genetic mutation that results in the loss of one copy of a heterozygous gene, often resulting in cancer due to loss of functional tumor suppressor genes. LOH in different cancer tissues have been observed at a number of forensic loci such as CSF1PO, FGA, vWA, D3S1358, D5S818, D8S1179, D13S317, and D18S51 in patients with laryngeal cancer (Rogowski et al., 2004). LOH may alter the results of a DNA profile and should be taken into consideration in cases where only cancerous tissue is available for analysis (Peloso et al., 2003; Zhou et al., 2017).

Qi et al. (2018) conducted a study investigating the possibility of using genetic markers rather than related genes to screen for predisposition to lung and liver cancer. This study used CODIS markers to examine the theory of programmed onset which hypothesizes that the occurrence of a chronic disease is independent of age and may instead depend on a programmed onset pattern. The results showed a significant difference in the occurrence of lung cancer between those who carried the D18S51-20 allele and those who did not, and the incidence of liver cancer between those carrying D21S11-30.2 and D6S1043-18 alleles and those who did not. While these results demonstrate CODIS markers being used to predict an individual's predisposition to cancer, there are an extensive number of cancer-related genes in the genome; therefore, the risk of breaching genetic privacy with this information remains low.

Y and X STRs

The Y chromosome has accumulated male advantage and fertility genes (Lahn and Page, 1997; Graves, 2006) and so it is possible that phenotypes associated with maleness are associated with Y STRs. X-linked phenotypes (as a result of recessive genes on the X chromosome) are more prevalent in males (because there is no dominant Y chromosome homolog) so there may also be associations with X STRs. In fact, X-linked genes have recently been shown to influence male fertility and sex ratio of offspring in mice (Kruger et al., 2019).

Association Versus Causation

The association of a STR with a trait or disease does not infer causation. Moreover, some alleles seem to have opposite effects: TH01 allele 9.3 may help with stress (Zhang et al., 2004) but also has a potential link with suicide (Persson et al., 1997; Yang et al., 2011). A genetic variant is considered causative when it is known that the presence of the variant will produce an effect that in turn

causes disease (Hu et al., 2018). None of the associations reported in this study offer proof of causation (except for trisomies), rather they propose a general relationship between some STRs used in forensic applications and a phenotype. These relationships may also be explained by confounding variables, bias, or by chance in cases where a significant finding is unable to be replicated by another study. In fact, this review could be seen as a reflection of the broader so-called “replication crisis” in science (Schooler, 2014). Many of the studies reported in this review may not have sufficiently mitigated against the “multiple comparison problem” where a number of comparisons will be significant by chance. By setting our p -value threshold to 0.05, we run the risk that 5% of significant results are significant by chance.

Many of the traits that can be predicted by genetic analysis are the result of epistatic interactions between genes and environmental factors. When considering the associations in this review, it is not reasonable to suggest that an individual possessing the more frequently observed allele associated with a trait will express a specific phenotype. There are many underlying mechanisms involved in the development of complex diseases and while the risk of forensic STRs being found to expose revealing medical information is minimal, the presence of a particular allele may indicate heightened potential or risk for a phenotype.

Molecular Mechanisms

While it remains true that forensic markers are located within non-coding regions, there is growing evidence that STRs in introns and up- or down-stream of genes may affect phenotype. STR mutations in the 5′ untranslated region (UTR) are known to modify gene expression, probably because they serve as protein binding sites (Li et al., 2004). Mutations in the 3′ UTR result in extended mRNA which can be toxic to the cell (Li et al., 2004; La Spada and Taylor, 2010). There are 13 CODIS STRs located in introns (**Supplementary Table 2**). Mutations in introns can affect mRNA splicing which can result in gene silencing or loss of function (Li et al., 2004; La Spada and Taylor, 2010). The TCAT repeat in the first intron of TH01 acts as a transcription regulatory element *in vitro* (Meloni et al., 1998). Albanèse et al. (2001) reported a reduction in transcriptional activity of TH as the TCAT repeat number varied from three to eight. STRs are also found at high density in promoter regions and it is highly likely that some are implicated in gene expression by modulating spacing of regulatory elements (Gemayel et al., 2012;

Sawaya et al., 2013; Gymrek et al., 2016; Quilez et al., 2016; Gymrek, 2017).

There is now etiological support for STRs as causative agents for disease in that they are quite plausibly epigenetic regulators for gene expression when located in introns or up- or down-stream of genes. This may increase prior support for the hypotheses of association and thus reduce the required significance level, as described by Kidd (1993), which is a counter to the “multiple comparison problem” discussed earlier.

CONCLUSION

While the results of this study did indicate a large number of phenotypic traits associated with forensic STRs, none were found to be independently causative or predictive of disease. Nevertheless, as there are numerous reported instances of tetranucleotide repeats being implicated in disease and molecular mechanisms have been demonstrated, there remains a strong chance that this inference may change in the near future. One limitation of this study was the sole use of the UCSC genome browser. Future studies may benefit from using a wider range of resources and investigating additional markers such as SNPs in flanking regions, mtDNA and Y-STRs. In the event that a statistically significant association, causal or predictive relationship is discovered, it is not necessarily a valid cause for removal from STR panels, but additional protective measures, such as tightening legislation surrounding genetic privacy, may need to be considered to prevent abuse of this information.

AUTHOR CONTRIBUTIONS

NW designed the study, performed the literature review, and wrote the manuscript. MB conceived the project, designed the study, and reviewed and edited the manuscript. DM conceived and managed the project, designed the study, and reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00884/full#supplementary-material>

REFERENCES

- Alam, S., Ferdous, A., Ali, M. E., Ahmed, A., Naved, A. F., and Akhteruzzaman, S. (2011). Forensic microsatellite TH01 and malaria predisposition. *Dhaka Univ. J. Biol. Sci.* 20, 1–6. doi: 10.3329/dujbs.v20i1.8831
- Albanèse, V., Biguet, N. F., Kiefer, H., Bayard, E., Mallet, J., and Meloni, R. (2001). Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite. *Hum. Mol. Genet.* 10, 1785–1792. doi: 10.1093/hmg/10.17.1785
- Algee-Hewitt, B. F. B., Michael, E. D., Kim, J., Li, J. Z., and Rosenberg, N. A. (2016). Individual identifiability predicts population identifiability in forensic microsatellite markers. *Curr. Biol.* 26, 935–942. doi: 10.1016/j.cub.2016.01.065
- Anney, R. J. L., Olsson, C. A., Lotfi-Miri, M., Patton, G. C., and Williamson, R. (2004). Nicotine dependence in a prospective population-based study of adolescents: the protective role of a functional tyrosine hydroxylase polymorphism. *Pharmacogenetics* 14, 73–81. doi: 10.1097/01.fpc.0000054157.92680.b6
- Antoni, M. H., Lutgendorf, S. K., Cole, S. W., Dhabhar, F. S., Sephton, S. E., McDonald, P. G., et al. (2006). The influence of bio-behavioural factors on tumour biology: pathways and mechanisms. *Nat. Rev. Cancer* 6, 240–248. doi: 10.1038/nrc1820

- Barbeau, P., Litaker, M. S., Jackson, R. W., and Treiber, F. A. (2003). A tyrosine hydroxylase microsatellite and hemodynamic response to stress in a multi-ethnic sample of youth. *Ethn. Dis.* 13, 186–192.
- Bastos, D. B., Sarafim-Silva, B. A. M., Sundefeld, M., Ribeiro, A. A., Brandao, J. D. P., Biasoli, E. R., et al. (2018). Circulating catecholamines are associated with biobehavioral factors and anxiety symptoms in head and neck cancer patients. *PLoS One* 13:e0202515. doi: 10.1371/journal.pone.0202515
- Bediaga, N. G., Aznar, J. M., Elcoroaristizabal, X., Alboniga, O., Gomez-Busto, F., Artabe, I. A., et al. (2015). Associations between STR autosomal markers and longevity. *Age* 37:95. doi: 10.1007/s11357-015-9818-5
- Biscotti, M. A., Olmo, E., and Heslop-Harrison, J. S. (2015). Repetitive DNA in eukaryotic genomes. *Chromosome Res* 23, 415–420. doi: 10.1007/s10577-015-9499-z
- Butler, J. M. (2006). Genetics and genomics of core short tandem repeat loci used in human identity testing. *J. Forensic Sci.* 51, 253–265. doi: 10.1111/j.1556-4029.2006.00046.x
- Castel, A. L., Cleary, J. D., and Pearson, C. E. (2010). Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat. Rev. Mol. Cell Biol.* 11, 165–170. doi: 10.1038/nrm2854
- Chen, H. Y., Ma, S. L., Huang, W., Ji, L., Leung, V. H. K., Jiang, H., et al. (2016). The mechanism of transactivation regulation due to polymorphic short tandem repeats (STRs) using IGF1 promoter as a model. *Sci. Rep.* 6:38225. doi: 10.1038/srep38225
- Chiba, M., Suzuki, S., Hinokio, Y., Hirai, M., Satoh, Y., Tashiro, A., et al. (2000). Tyrosine hydroxylase gene microsatellite polymorphism associated with insulin resistance in depressive disorder. *Metabolism* 49, 1145–1149. doi: 10.1053/meta.2000.8611
- Cole, S. A. (2007). Is the “Junk”. DNA designation bunk? *Nw. UL Rev. Colloquy* 102, 54–63.
- Courts, C., and Madea, B. (2011). Significant association of TH01 allele 9.3 and SIDS. *J. Forensic Sci.* 56, 415–417. doi: 10.1111/j.1556-4029.2010.01670.x
- De Benedictis, G., Carotenuto, L., Carrieri, G., De Luca, M., Falcone, E., Rose, G., et al. (1998). Gene/longevity association studies at four autosomal loci (REN, THO, PARP, SOD2). *Eur. J. Hum. Genet.* 6, 534–541. doi: 10.1038/sj.ejhg.5200222
- Eisenhofer, G., Kopin, I. J., and Goldstein, D. S. (2004). Catecholamine metabolism: a contemporary view with implications for physiology and medicine. *Pharmacol. Rev.* 56, 331–349. doi: 10.1124/pr.56.3.1
- Gaikwad, S., Ashma, R., Kumar, N., Trivedi, R., and Kashyap, V. K. (2005). Host microsatellite alleles in malaria predisposition? *Malar. J.* 4, 331–349. doi: 10.1186/1475-2875-4-50
- Gemayel, R., Cho, J., Boeynaems, S., and Verstrepen, K. J. (2012). Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. *Genes* 3, 461–480. doi: 10.3390/genes3030461
- Gettings, K. B., Lai, R., Johnson, J. L., Peck, M. A., Hart, J. A., Gordish-Dressman, H., et al. (2014). A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population. *Forensic Sci. Int. Genet.* 8, 101–108. doi: 10.1016/j.fsigen.2013.07.010
- Giusti-Rodríguez, P., and Sullivan, P. F. (2013). The genomics of schizophrenia: update and implications. *J. Clin. Invest.* 123, 4557–4563. doi: 10.1172/JCI66031
- Graves, J. A. (2006). Sex chromosome specialization and degeneration in mammals. *Cell* 124, 901–914. doi: 10.1016/j.cell.2006.02.024
- Graydon, M., Cholette, F., and Ng, L.-K. (2009). Inferring ethnicity using 15 autosomal STR loci—Comparisons among populations of similar and distinctly different physical traits. *Forensic Sci. Int. Genet.* 3, 251–254. doi: 10.1016/j.fsigen.2009.03.002
- Guan, L., Ren, C., Li, H., Gao, L., Jia, N., and Guan, H. (2013). [Practicality of rapid prenatal screening for Down syndrome with PCR-short tandem repeat method]. *Chinese J. of Med. Gen.* 30, 277–282. doi: 10.3760/cma.j.issn.1003-9406.2013.03.006
- 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
- Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., et al. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* 48, 22–29. doi: 10.1038/ng.3461
- Hannan, A. J. (2018). Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* 19, 286–298. doi: 10.1038/nrg.2017.115
- Hu, P., Jiao, R., Jin, L., and Xiong, M. (2018). Application of causal inference to genomic analysis: advances in methodology. *Front. Genet.* 9:238. doi: 10.3389/fgene.2018.00238
- Hui, L., Liping, G., Jian, Y., and Laisui, Y. (2014). A new design without control population for identification of gastric cancer-related allele combinations based on interaction of genes. *Gene* 540, 32–36. doi: 10.1016/j.gene.2014.02.033
- Jacewicz, R., Babol-Pokora, K., Berent, J., Pepinski, and Szram, S. (2006a). Are tetranucleotide microsatellites implicated in neuropsychiatric diseases? *Int. Congr. Ser.* 1288, 783–785. doi: 10.1016/j.ics.2005.09.101
- Jacewicz, R., Szram, S., Galecki, P., and Berent, J. (2006b). Will genetic polymorphism of tetranucleotide sequences help in the diagnostics of major psychiatric disorders? *Forensic Sci. Int.* 162, 24–27. doi: 10.1016/j.forsciint.2006.06.024
- Katsanis, S. H., and Wagner, J. K. (2013). Characterization of the standard and recommended CODIS markers. *J. Forensic Sci.* 58(Suppl. 1), S169–S172. doi: 10.1111/j.1556-4029.2012.02253.x
- Kaye, D. H. (2007). Please, Let's Bury the Junk: the CODIS Loci and the Revelation of Private Information. *Nw. UL Rev. Colloquy* 102:70.
- Kidd, K. K. (1993). Associations of disease with genetic markers: déjà vu all over again. *Am. J. Med. Genet.* 48, 71–73. doi: 10.1002/ajmg.1320480202
- Kimpton, C., Fisher, D., Watson, S., Adams, M., Urquhart, A., Lygo, J., et al. (1994). Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. *Int. J. Legal Med.* 106, 302–311. doi: 10.1007/BF01224776
- Klintschar, M., Reichenpfader, B., and Saternus, K. S. (2008). A functional polymorphism in the tyrosine hydroxylase gene indicates a role of noradrenalinergic signaling in sudden infant death syndrome. *J. Pediatr.* 153, 190–193. doi: 10.1016/j.jpeds.2008.02.032
- Kruger, A. N., Brogley, M. A., Huizinga, J. L., Kidd, J. M., de Rooij, D. G., Hu, Y. C., et al. (2019). A neofunctionalized X-linked ampliconic gene family is essential for male fertility and equal sex ratio in mice. *Curr. Biol.* 29, 3699.e5–3706.e5. doi: 10.1016/j.cub.2019.08.057
- Kurumaji, A., Kuroda, T., Yamada, K., Yoshikawa, T., and Toru, M. (2001). An association of the polymorphic repeat of tetranucleotide (TCAT) in the first intron of the human tyrosine hydroxylase gene with schizophrenia in a Japanese sample. *J. Neural Transm.* 108, 489–495. doi: 10.1007/s007020170069
- La Spada, A. R., and Taylor, J. P. (2010). Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.* 11, 247–258. doi: 10.1038/nrg2748
- Lahn, B. T., and Page, D. C. (1997). Functional coherence of the human Y chromosome. *Science* 278, 675–680. doi: 10.1126/science.278.5338.675
- Laird, R., Schneider, P. M., and Gaudieri, S. (2007). Forensic STRs as potential disease markers: a study of VWA and von Willebrand's Disease. *Forensic Sci. Int. Genet.* 1, 253–261. doi: 10.1016/j.fsigen.2007.06.002
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062
- Li, Y.-C., Korol, A. B., Fahima, T., and Nevo, E. (2004). Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.* 21, 991–1007. doi: 10.1093/molbev/msh073
- Liou, J. D., Chu, D. C., Cheng, P. J., Chang, S. D., Sun, C. F., Wu, Y. C., et al. (2004). Human chromosome 21-specific DNA markers are useful in prenatal detection of Down syndrome. *Ann. Clin. Lab. Sci.* 34, 319–323.
- Lowe, A. L., Urquhart, A., Foreman, L. A., and Evett, I. W. (2001). Inferring ethnic origin by means of an STR profile. *Forensic Sci. Int.* 119, 17–22. doi: 10.1016/S0379-0738(00)00387-X
- Mäki, P., Veijola, J., Jones, P. B., Murray, G. K., Koponen, H., Tienari, P., et al. (2005). Predictors of schizophrenia—a review. *Br. Med. Bull.* 73-74, 1–15. doi: 10.1093/bmb/ldh046
- McEwen, B. S. (2002). Sex, stress and the hippocampus: allostasis, allostatic load and the aging process. *Neurobiol. Aging* 23, 921–939. doi: 10.1016/S0197-4580(02)00027-1
- Meiser, J., Weindl, D., and Hiller, K. (2013). Complexity of dopamine metabolism. *Cell Commun. Signal.* 11:34. doi: 10.1186/1478-811X-11-34
- Meloni, R., Albanese, V., Ravassard, P., Treilhou, F., and Mallet, J. (1998). A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. *Hum. Mol. Genet.* 7, 423–428. doi: 10.1093/hmg/7.3.423

- Meraz-Rios, M. A., Majluf-Cruz, A., Santana, C., Noris, G., Camacho-Mejorado, R., Acosta-Saavedra, L. C., et al. (2014). Association of vWA and TPOX polymorphisms with venous thrombosis in Mexican mestizos. *BioMed Res. Int.* 9:697689. doi: 10.1155/2014/697689
- Modai, S., and Shomron, N. (2016). Molecular risk factors for schizophrenia. *Trends. Mol. Med.* 22, 242–253. doi: 10.1016/j.molmed.2016.01.006
- Morimoto, K., Miyatake, R., Nakamura, M., Watanabe, T., Hirao, T., and Suwaki, H. (2002). Delusional disorder: molecular genetic evidence for dopamine psychosis. *Neuropsychopharmacology* 26, 794–801. doi: 10.1016/S0893-133X(01)00421-3
- Ng, J., Papandreou, A., Heales, S. J., and Kurian, M. A. (2015). Monoamine neurotransmitter disorders—clinical advances and future perspectives. *Nat. Rev. Neurol.* 11, 567–584. doi: 10.1038/nrneurol.2015.172
- Orr, H. T., and Zoghbi, H. Y. (2007). Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* 30, 575–621. doi: 10.1146/annurev.neuro.29.051605.113042
- Peloso, G., Grignani, P., Rosso, R., and Previdere, C. (2003). “Forensic evaluation of tetranucleotide STR instability in lung cancer,” in *Progress in Forensic Genetics* 9, eds B. Brinkman and A. Carracedo (Amsterdam: Elsevier Science Bv), 719–721. doi: 10.1016/s0531-5131(02)00500-9
- Persson, M. L., Wasserman, D., Geijer, T., Jonsson, E. G., and Terenius, L. (1997). Tyrosine hydroxylase allelic distribution in suicide attempters. *Psychiatry Res.* 72, 73–80. doi: 10.1016/s0165-1781(97)00068-1
- Persson, M. L., Wasserman, D., Jonsson, E. G., Bergman, H., Terenius, L., Gyllander, A., et al. (2000). Search for the influence of the tyrosine hydroxylase (TCAT)n repeat polymorphism on personality traits. *Psychiatry Res.* 95, 1–8. doi: 10.1016/s0165-1781(00)00160-8
- Qi, X., Yu, Y. J., Ji, N., Ren, S. S., Xu, Y. C., and Liu, H. (2018). Genetic risk analysis for an individual according to the theory of programmed onset, illustrated by lung and liver cancers. *Gene* 673, 107–111. doi: 10.1016/j.gene.2018.06.044
- Quilez, J., Guilmatre, A., Garg, P., Highnam, G., Gymrek, M., Erlich, Y., et al. (2016). Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* 44, 3750–3762. doi: 10.1093/nar/gkw219
- Ramel, C. (1997). Mini- and microsatellites. *Environ. Health Perspect.* 105, 781–789. doi: 10.2307/3433284
- Relethford, J. H. (1997). Hemispheric difference in human skin color. *Am. J. Phys. Anthropol.* 104, 449–457. doi: 10.1002/(sici)1096-8644(199712)104:4<449::aid-ajpa2>3.0.co;2-n
- Rogowski, M., Walenczak, I., Pepinski, W., Skawronska, M., Sieskiewicz, A., and Klatka, J. (2004). Loss of heterozygosity in laryngeal cancer. *Rocz. Akad. Med. Białymst.* 49, 262–264.
- Sander, T., Harms, H., Rommelspacher, H., Hoehe, M., and Schmidt, L. G. (1998). Possible allelic association of a tyrosine hydroxylase polymorphism with vulnerability to alcohol-withdrawal delirium. *Psychiatr. Genet.* 8, 13–17.
- Sarkar, S. P., and Adshear, G. (2010). Whose DNA is it anyway? European court, junk DNA, and the problem with prediction. *J. Am. Acad. Psychiatry Law* 38, 247–250.
- Sawaya, S., Bagshaw, A., Buschiazio, E., Kumar, P., Chowdhury, S., Black, M. A., et al. (2013). Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One* 8:e54710. doi: 10.1371/journal.pone.0054710
- Schooler, J. W. (2014). Metascience could rescue the ‘replication crisis’. *Nature* 515:9. doi: 10.1038/515009a
- Scientific Working Group of DNA Analysis Methods [SWGDM] (2013). *SWGDM Considerations for Claims That the CODIS Core Loci are ‘Associated’ With Medical Conditions/Diseases*. Quantico, VA: SWGDM.
- Serretti, A., Macciardi, F., Verga, M., Cusin, C., Pedrini, S., and Smeraldi, E. (1998). Tyrosine hydroxylase gene associated with depressive symptomatology in mood disorder. *Am. J. Med. Genet.* 81, 127–130. doi: 10.1002/(SICI)1096-8628(19980328)81:2<127::AID-AJMG1>3.0.CO;2-T
- Shi, Y. F., Li, X. Z., Li, Y., Zhang, X. L., Zhang, Y., and Yue, T. F. (2012). [Diagnosis of Downs syndrome using short tandem repeat loci D21S11, D21S1440 and Penta D]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* 29, 443–446. doi: 10.3760/cma.j.issn.1003-9406.2012.04.014
- Studer, J., Bartsch, C., and Haas, C. (2014). Tyrosine hydroxylase TH01 9.3 allele in the occurrence of sudden infant death syndrome in swiss caucasians. *J. Forensic Sci.* 59, 1650–1653. doi: 10.1111/1556-4029.12526
- Sutherland, G., Mellick, G., Newman, J., Double, K. L., Stevens, J., Lee, L., et al. (2008). Haplotype analysis of the IGF2-INS-TH gene cluster in Parkinson’s disease. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 147B, 495–499. doi: 10.1002/ajmg.b.30633
- Szibor, R., Hering, S., and Edelmann, J. (2005). The HumARA genotype is linked to spinal and bulbar muscular dystrophy and some further disease risks and should no longer be used as a DNA marker for forensic purposes. *Int. J. Legal Med.* 119, 179–180. doi: 10.1007/s00414-005-0525-0
- Tautz, D., and Schlotterer, (1994). Simple sequences. *Curr. Opin. Genet. Dev.* 4, 832–837.
- Tochigi, M., Otowa, T., Hibino, H., Kato, C., Otani, T., Umekage, T., et al. (2006). Combined analysis of association between personality traits and three functional polymorphisms in the tyrosine hydroxylase, monoamine oxidase A, and catechol-O-methyltransferase genes. *J. Neurosci. Res.* 84, 180–185. doi: 10.1016/j.neures.2005.11.003
- von Wurmb-Schwark, N., Caliebe, A., Schwark, T., Kleindorp, R., Poetsch, M., Schreiber, S., et al. (2011). Association of TH01 with human longevity revisited. *Eur. J. Hum. Genet.* 19, 924–927. doi: 10.1038/ejhg.2011.43
- Wang, Z. L., Dai, L., Li, S., Qiu, G. Q., and Wu, H. Q. (2012). [Comparison of allelic frequencies of 15 short tandem repeat loci between chronic myeloid leukemia patients and non-related healthy individuals]. *Chin. J. Med. Genet.* 29, 306–308. doi: 10.3760/cma.j.issn.1003-9406.2012.03.013
- Wei, J., Ramchand, C. N., and Hemmings, G. P. (1997). Possible association of catecholamine turnover with the polymorphic (TCAT)n repeat in the first intron of the human tyrosine hydroxylase gene. *Life Sci.* 61, 1341–1347. doi: 10.1016/S0024-3205(97)00679-6
- Wu, Y., Zhang, Q., Liu, B., and Yu, G. (2008). The analysis of the entire HLA, partial non-HLA and HPV for Chinese women with cervical cancer. *J. Med. Virol.* 80, 1808–1813. doi: 10.1002/jmv.21251
- Yang, C., Ba, H., Gao, Z., Zhao, H., Yu, H., and Guo, W. (2013). Case-control study of allele frequencies of 15 short tandem repeat loci in males with impulsive violent behavior. *Shanghai Arch. Psychiatry* 25, 354–363. doi: 10.3969/j.issn.1002-0829.2013.06.004
- Yang, C., Ba, H., and Zhao, H. (2011). Association study between the genetic polymorphism of 15 STR loci and the suicide behavior in Jiangsu province. *J. Psych.* 1:9.
- Yang, C., Huajie, B., Gao, Z., Lin, Z., Zhao, H., Liu, B., et al. (2010). Association study between the genetic polymorphism of 15 STR loci and the crime of rape. *Chin. J. Behav. Med. Brain Sci.* 19, 421–424.
- Yoon, H. R., Park, Y. S., and Kim, Y. K. (2002). Rapid prenatal detection of Down and Edwards syndromes by fluorescent polymerase chain reaction with short tandem repeat markers. *Yonsei Med. J.* 43, 557–566. doi: 10.3349/ymj.2002.43.5.557
- Zhang, L., Rao, F., Wessel, J., Kennedy, B. P., Rana, B. K., Taupenot, L., et al. (2004). Functional allelic heterogeneity and pleiotropy of a repeat polymorphism in tyrosine hydroxylase: prediction of catecholamines and response to stress in twins. *Physiol. Genomics* 19, 277–291. doi: 10.1152/physiolgenomics.00151.2004
- Zhou, S., Wang, H., Wang, Q. K., Wang, P., Wang, F., and Xu, C. (2017). Loss of heterozygosity detected at three short tandem repeat locus commonly used for human DNA identification in a case of paternity testing. *Legal Med.* 24, 7–11. doi: 10.1016/j.legalmed.2016.11.001
- Zhuo, C., Hou, W., Li, G., Mao, F., Li, S., Lin, X., et al. (2019). The genomics of schizophrenia: shortcomings and solutions. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 93, 71–76. doi: 10.1016/j.pnpbp.2019.03.009

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

Copyright © 2020 Wyner, Barash and McNevin. 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.



MicroHapDB: A Portable and Extensible Database of All Published Microhaplotype Marker and Frequency Data

Daniel S. Standage* and Rebecca N. Mitchell

National Bioforensic Analysis Center, National Biodefense Analysis and Countermeasures Center (NBACC), Frederick, MD, United States

OPEN ACCESS

Edited by:

Kenneth K. Kidd,
Yale University, United States

Reviewed by:

Guanglin He,
Sichuan University, China
Yiping Hou,
Sichuan University, China
Dennis McNevin,
University of Technology Sydney,
Australia

*Correspondence:

Daniel S. Standage
daniel.standage@nbacc.dhs.gov

Specialty section:

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

Received: 08 April 2020

Accepted: 30 June 2020

Published: 07 August 2020

Citation:

Standage DS and Mitchell RN (2020)
MicroHapDB: A Portable and
Extensible Database of All Published
Microhaplotype Marker and
Frequency Data.
Front. Genet. 11:781.
doi: 10.3389/fgene.2020.00781

Microhaplotypes are the subject of significant interest in the forensics community as a promising multi-purpose forensic DNA marker for human identification. Microhaplotype markers are composed of multiple SNPs in close proximity, such that a single NGS read can simultaneously genotype the individual SNPs and phase them in aggregate to determine the associated donor haplotype. Abundant throughout the human genome, numerous recent studies have sought to discover and rank microhaplotype markers according to allelic diversity within and among populations. Microhaplotypes provide an appealing alternative to STR markers for human identification and mixture deconvolution, but can also be optimized for ancestry inference or combined with phenotype SNPs for prediction of externally visible characteristics in a multiplex NGS assay. Designing and evaluating panels of microhaplotypes is complicated by the lack of a convenient database of all published data, as well as the lack of population allele frequency data spanning disparate marker collections. We present MicroHapDB, a comprehensive database of published microhaplotype marker and frequency data, as a tool to advance the development of microhaplotype-based human forensics capabilities. We also present population allele frequencies derived from 26 global population samples for all microhaplotype markers published to date, facilitating the design and interpretation of custom multi-source panels. We submit MicroHapDB as a resource for community members engaged in marker discovery, population studies, assay development, and panel and kit design.

Keywords: microhaplotype, forensics, human identification, next generation sequencing, Python, database, bioinformatics

1. INTRODUCTION

Well-studied short tandem repeat (STR) markers have formed the basis of forensic human identification methods since the 1990s. The most common strategy in practice today utilizes several fluorescent dyes to type 20 or more STR markers in a single polymerase chain reaction (PCR) followed by capillary electrophoresis (CE) detection (Butler, 2010). The resulting DNA profiles, combined with STR allele frequency estimates, can then be used to calculate match statistics or evaluate the relative weight of evidence for competing propositions in a likelihood ratio framework

(Butler, 2015; Cowell et al., 2015; Bleka et al., 2016a,b). Statistics obtained via STR typing can provide high confidence given the number of independent markers in an assay and the multiallelic nature of each marker.

Despite impressive recent improvements in DNA sequencing technologies, next-generation sequencing (NGS) assays of single nucleotide polymorphism (SNP) markers have seen slow adoption for forensic human identification. The ability to genotype sufficient numbers of SNPs to achieve suitable statistical power remains beyond the scope of many forensics laboratories. A relevant factor is the forensics community's strong disinclination, on ethical and privacy grounds, to use DNA markers associated with human diseases or conditions, which limits the utilization of many commonly used microarrays and SNP chips. Also, because the majority of SNPs are bi-allelic, less population-level diversity is observed at each marker than at multi-allelic STRs, resulting in reduced discriminatory power when comparing reference and evidentiary samples. While this can be compensated for to some extent with a larger panel (SNPs are incredibly abundant in the human genome), the statistical requirement for markers that are inherited independently complicates panel design and places a practical limit on the resulting panel size.

Microhaplotypes (often abbreviated as *microhaps* or *MHs*) have recently prompted considerable interest in the forensics community as a promising alternative to independent SNPs and STRs for human identification (Kidd et al., 2018; Oldoni et al., 2019). A microhaplotype marker is defined by multiple SNPs¹ residing within a short genomic distance whose state is reported as the allelic combination of all its component SNPs—that is, the haplotype. Here, “short” simply means a few hundred base pairs or fewer, ensuring a low frequency of recombination within the marker, and that a single NGS read or read pair can span all of the marker's component SNPs. This length constraint enables each distinct read to both genotype and phase its target marker; that is, to determine (1) the individual allele of each component SNP, as well as (2) the haplotype. Even if a particular microhap is composed only of biallelic SNPs, the presence of multiple component SNPs makes it possible to observe several haplotypes at the marker, substantially increasing its discriminatory power over independent SNPs.

With a targeted NGS sequencing assay, a sufficient number of reads are collected to confidently genotype each marker, differentiating between true haplotypes and those arising from sequencing error. Microhap markers exhibit none of the stutter artifacts commonly observed in PCR-based STR assays, and the substitution and homopolymer errors common to some NGS platforms are easily resolved with sufficient depth of coverage. The restricted length of microhap markers makes them suitable for typing degraded samples, and the ability of NGS assays to capture additional rare SNPs within the microhaplotype can provide valuable information for mixture detection and analysis.

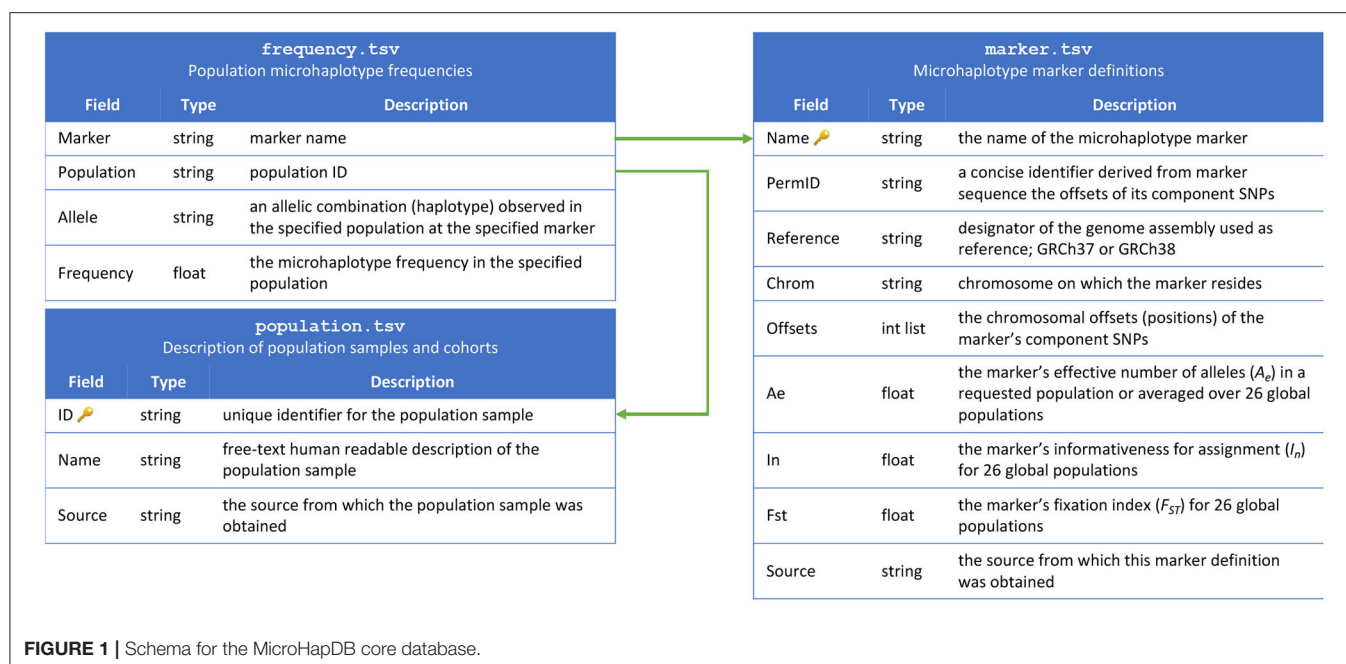
Another notable benefit of microhaps is that they can be selected not only for high *within*-population variation, but also for high *between*-population variation, facilitating prediction of biogeographic ancestry (Oldoni et al., 2017; Chen et al., 2019b; Zhu et al., 2019). It is thus possible to design a comprehensive forensic panel using a combination of microhap and SNP markers that will enable identification, mixture analysis (Bennett et al., 2019; Coble and Bright, 2019), ancestry inference, and prediction of externally visible characteristics (Ruiz et al., 2013; Walsh et al., 2013; Crawford et al., 2017) in a single NGS-based assay.

Microhaps are abundant in the human genome, and thus discovering and ranking them for different purposes is an area of active research interest in the forensics community. In just the last few years, numerous studies presenting new microhap marker collections have been published, together totaling more than 400 markers (Hiroaki et al., 2015; Kidd et al., 2018; van der Gaag et al., 2018; Voskoboinik et al., 2018; Chen et al., 2019a; Staadig and Tillmar, 2019; de la Puente et al., 2020). While two of these studies also include allele frequency data for multiple population samples—including 83 populations² in Kidd et al. (2018) and three populations in van der Gaag et al. (2018)—the others present either no frequency data or data for a single population sample, with little overlap between studies. The absence of a single point of access for published microhaps and the paucity of allele frequency data spanning disparate data sets are obstacles to developing and evaluating custom panels composed of microhap markers selected from different published collections.

To support the development of microhaplotype-based human forensics capabilities, we have compiled a database of all published microhap marker definitions and allele frequencies. MicroHapDB is a portable database that, once installed, can be accessed by the user without an Internet connection. The entire contents of the database are distributed with each copy of MicroHapDB, and instructions for adding private data to a local instance of the database are provided. The same instructions can alternatively be used by MicroHapDB maintainers or interested community contributors to submit new markers and allele frequencies for review and potential inclusion in the public database. MicroHapDB is designed to be user-friendly both for forensic practitioners and researchers, and supports a variety of access methods including browsing, simple or complex text queries, and programmatic database access via a Python application programming interface (API). Finally, to increase the value of the published microhaps in aggregate, we have used 2,504 fully phased genomes from the 1,000 Genomes Project (Auton et al., 2015) to estimate allele frequencies in 26 global populations for 412 microhap markers. MicroHapDB is a valuable resource for researchers, practitioners, and commercial entities engaged in marker discovery, population studies, assay development and validation, and design of custom panels and kits for forensic applications.

¹While the vast majority of published microhaplotype markers are composed exclusively of SNPs, a small number include one or more insertion/deletion polymorphisms (indels).

²As of the latest December 2018 update, there are now 96 populations in the Allele Frequency Database (ALFRED) for which microhaplotype frequency data is available.



2. MATERIALS AND METHODS

2.1. Database Design

The contents of the MicroHapDB database are stored in nine tables, distributed as plain-text tab-delimited files. Three of the tables constitute the “core database,” and include population sample descriptions, marker definitions, and microhaplotype frequencies (Figure 1). The remaining four tables contain ancillary metadata: a cross reference of third-party identifiers to MicroHapDB identifiers; data for a small number of indel variants present in the database; the genomic sequences spanning each marker (to facilitate amplicon design); a mapping of variant identifiers (rsIDs) to marker names; supplementary allelic variation statistics for specific populations; and marker coordinates for the GRCh37 reference genome assembly (GRCh38 is used by default).

2.2. User Interface

MicroHapDB is compatible with Windows and UNIX computers, and has been tested on Windows 10, Mac OS X, and Linux operating systems. In each case, the primary interface for querying MicroHapDB is the terminal or command line. The `microhapdb` command provides three operations corresponding to the three tables in the core database: `microhapdb marker`, `microhapdb population`, and `microhapdb frequency`. Executing any of these commands with no additional arguments will print the entire contents of the specified table to the terminal for browsing. Each command also enables a user to restrict the printed results to data matching a particular identifier, source, or genomic region.

By default, all results are printed in tabular format. Population data can optionally be printed in a “detail” format summarizing the number of markers for which microhaplotype frequency is

available and the total number of microhaplotypes (microhaps) observed in the population sample. Marker data can optionally be printed in FASTA format for use with third-party programs, or in a “detail” format showing the genomic location of the marker and its component SNPs, the core marker sequence (spanning only the microhap's most distal component SNPs), all haplotypes observed at the marker, and a candidate amplicon sequence for the marker (for which the amount of flanking sequence can be configured using the `--delta` and `--min-length` parameters) (see Figure 2).

Once MicroHapDB is installed on a computer, all database list and search operations access only the data files resident on that machine. MicroHapDB doesn't require or permit requests to transfer data to or from databases residing on remote machines.

Installed as a Python package, MicroHapDB also supports programmatic access to the core and ancillary database tables. After invoking `import microhapdb`, users can write custom code to query and analyze marker, population, or frequency data resident in the database tables, pre-loaded into memory as pandas dataframes (McKinney, 2011). Alternatively, users can execute the `microhapdb --files` command from the UNIX shell to show the location of the database table files, which can be imported directly into R, Excel, or any other data analytics environment preferred by the user.

2.3. Implementation and Availability

At its core, MicroHapDB is composed of a small number of tabular plain-text data files containing marker, population, and frequency information for published microhaps. These files enclose the entire contents of the database. In contrast to many databases of genetic variation, each instance of MicroHapDB stores the entire contents of the database locally. MicroHapDB does not communicate with any central database server, and


```

$ microhapdb marker --format=detail mh04CP-002
-----[ MicroHapDB ]-----
mh04CP-002      a.k.a MHDDBM-342c1521, SI664878N

Marker Definition (GRCh38)
Marker extent
- chr4:24304952-24304972 (20 bp)
Target amplicon
- chr4:24304922-24305002 (80 bp)
Constituent variants
- chromosome offsets: 24304952,24304955,24304971
- marker offsets: 0,3,19
- amplicon offsets: 30,33,49
- cross-references: rs35619595,rs34017818,rs6814654
Observed haplotypes
- C,A,A
- C,A,T
- C,G,A
- C,G,T
- G,A,A

--[ Core Marker Sequence ]--
>mh04CP-002
CGCGCCAGGTATGAAGTTAT

--[ Target Amplicon Sequence with Haplotypes ]--
* * *
ATTGCTAAGCATCTACTATGTGGCAAACCCGCGCCAGGTATGAAGTTATTATGGCTGAAGATGGATAAGTCAGACAAAG
.....C..A.....A.....
.....C..A.....T.....
.....C..G.....A.....
.....C..G.....T.....
.....G..A.....A.....
-----

```

FIGURE 2 | The “detail” view for a 3-SNP microhaplotype marker, displayed using the MicroHapDB command-line interface.

network connections are only used to install the database or upgrade to a newer version.

The user interface described in section 2.2 is implemented in a Python package that can be installed and upgraded using the Bioconda software manager (Grüning et al., 2018). Source code for MicroHapDB is published open-access on the GitHub platform at <https://github.com/bioforensics/MicroHapDB/> and is free for commercial and non-commercial use under a permissive open source license. The authors operate the MicroHapDB project under an open governance model that facilitates and encourages contributions from the community.

The software and procedure used by the authors to build the database is also published on the MicroHapDB GitHub repository. During the build process, data from several sources is independently pre-processed and standardized, and then all sources are aggregated and sorted to compile the final database. This strategy, described further in section 2.4, serves several purposes. First, it provides a clear mechanism for the authors or other community members to extend the database in the future as additional marker and frequency data is published in the literature. Second, the same mechanism enables interested users to supplement the public MicroHapDB database with private data in a safe and secure way. By following the guidelines in the database build instructions provided in the MicroHapDB repository, a user can rebuild their local copy of MicroHapDB with additional sources of marker and/or frequency data. Because MicroHapDB doesn’t communicate with any central database, changes made to a user’s local copy of MicroHapDB do not

propagate to GitHub or any other location. Third, it permits careful scrutiny of the entire database construction process by any interested party in case errors in the database contents are ever discovered.

2.4. Data Collection and Pre-processing

MicroHapDB was compiled from seven distinct sources, each of which organized and reported data in a unique format. Extracting the relevant data and cross-referencing with public databases of genomic variation required a combination of manual and automated strategies uniquely designed for each distinct source. The result of this preliminary data acquisition and pre-processing was a collection of seven data sets with consistently formatted population descriptions, marker definitions, and allele frequencies. Once data from each distinct source was collected, cross-referenced with the GRCh37 and GRCh38 human reference genomes, and standardized, the final database compilation was performed by aggregating and sorting all data sources.

Source code and corresponding technical documentation describing data collection, pre-processing, and aggregation of the final database is available at <https://github.com/bioforensics/MicroHapDB/tree/0.6/dbbuild>.

2.5. Estimation of Haplotype Frequencies for 26 Global Populations

MicroHapDB includes data from several distinct sources, but the availability of population frequencies for published microhaps

is inconsistent. For some markers, frequencies are reported for dozens of population samples. Other markers have frequencies reported only for a single cohort, and yet other markers have no frequencies reported whatsoever. Designing and testing panels composed of markers from multiple distinct sources is possible, but prior to the release of MicroHapDB, interpretation of any sample assayed using such a panel would require the development of appropriate frequency data. We used 2,504 fully phased genotypes from a publicly available large-cohort study (Auton et al., 2015) to estimate population frequencies for all published microhaps across a set of 26 global population samples. These frequencies were first published in MicroHapDB version 0.5.

In the most recent version, MicroHapDB 0.6 contains definitions for 417 microhap markers. Five of these markers³ are defined by rare variants not genotyped in the 1,000 Genomes Project Phase 3 data, and were thus excluded from this analysis. For each of the remaining 412 markers, population frequencies for 26 global populations were estimated using the following procedure. First, phased genotype records for each of the marker's component variants were retrieved using the variants' rsIDs. Next, the phased genotypes were aggregated to determine the two haplotypes for each individual at the marker (or the single haplotype observed at X chromosome markers in males). Then, noting the population sample with which each individual was associated, a tally of haplotypes was compiled for each population. Finally, the haplotype tallies for each population sample were normalized by the corresponding number of alleles to compute the final frequency estimates.

2.6. Calculation of Measures of Variation Within and Among Populations

Microhaps are suitable for numerous forensic applications. Two common statistics used for ranking microhaps are the effective number of alleles A_e and the informativeness for assignment I_n (Crow and Kimura, 1970; Rosenberg et al., 2003; Kidd and Speed, 2015; Kidd et al., 2018). The A_e statistic is the reciprocal of a marker's homozygosity. For a marker with N alleles, A_e is computed as

$$A_e = \frac{1}{\sum p_i^2}$$

where p_i is the frequency of allele $i \in N$ and summation is over all alleles. It is a measure of allelic variation within a population, and corresponds to a marker's power for individual identification. For a microhaplotype with N SNPs, the maximal A_e value of 4^N occurs if and only if every possible allelic combination is observed at equal frequencies in the population. In reality, only a subset of possible allelic combinations are generally present in a population, and typically at unequal frequencies, resulting in A_e values that most commonly fall between 1.5 and 4.5 for previously reported microhap markers. The minimal A_e value of

1 occurs when only a single haplotype is observed at the locus in a population.

By contrast, I_n measures the extent of population-specific allelic variation among a set of populations, and corresponds to a microhap's power for predicting an individual's biogeographic ancestry. The I_n statistic for a marker with N alleles across K populations is calculated as

$$I_n = \sum_{j=1}^N \left(-p_j \log p_j + \sum_{i=1}^K \frac{p_{ij}}{K} \log p_{ij} \right)$$

where p_{ij} is the frequency of allele j in population i . This statistic measures the difference in information content when allele frequencies are aggregated across all populations versus when they are collated within populations. The minimal I_n of 0 occurs when all alleles have equal frequencies in all populations, and the maximal value $\log K$ occurs when $N \geq K$ and no allele is found in more than one population (Rosenberg et al., 2003).

A third statistic, the *fixation index* (F_{ST}), is another measure of allelic variation that considers *coancestry*, and is commonly used in forensic analysis to correct for population substructure (Butler, 2015). High F_{ST} values indicate that allele frequencies differ substantially among subpopulations, while low F_{ST} values indicate higher similarity among subpopulations.

As a final post-processing step in the MicroHapDB database build procedure, A_e and I_n statistics were computed for all markers in MicroHapDB⁴. Using population microhaplotype frequencies computed from the 1,000 Genomes Project Phase 3 genotypes (Auton et al., 2015), MicroHapDB scripts computed per-marker A_e values individually for each population. By default, the A_e column of MicroHapDB's `markers` table displays the arithmetic mean of A_e over all 26 populations, but the command line interface and Python API both provide an option for the user to choose a specific population for which to display A_e values. I_n statistics over 26 populations were calculated with the same frequency data using INFOCALC (<https://rosenberglab.stanford.edu/infocalc.html>), and are listed in the `I_n` column of the `markers` table. The same frequency data were also used to calculate F_{ST} statistics using the Weir and Cockerham formulation (Weir and Cockerham, 1984), as implemented in the `scikit-allel` package version 1.21 (Miles et al., 2019). The F_{ST} statistics reported in MicroHapDB were averaged across all alleles for each marker.

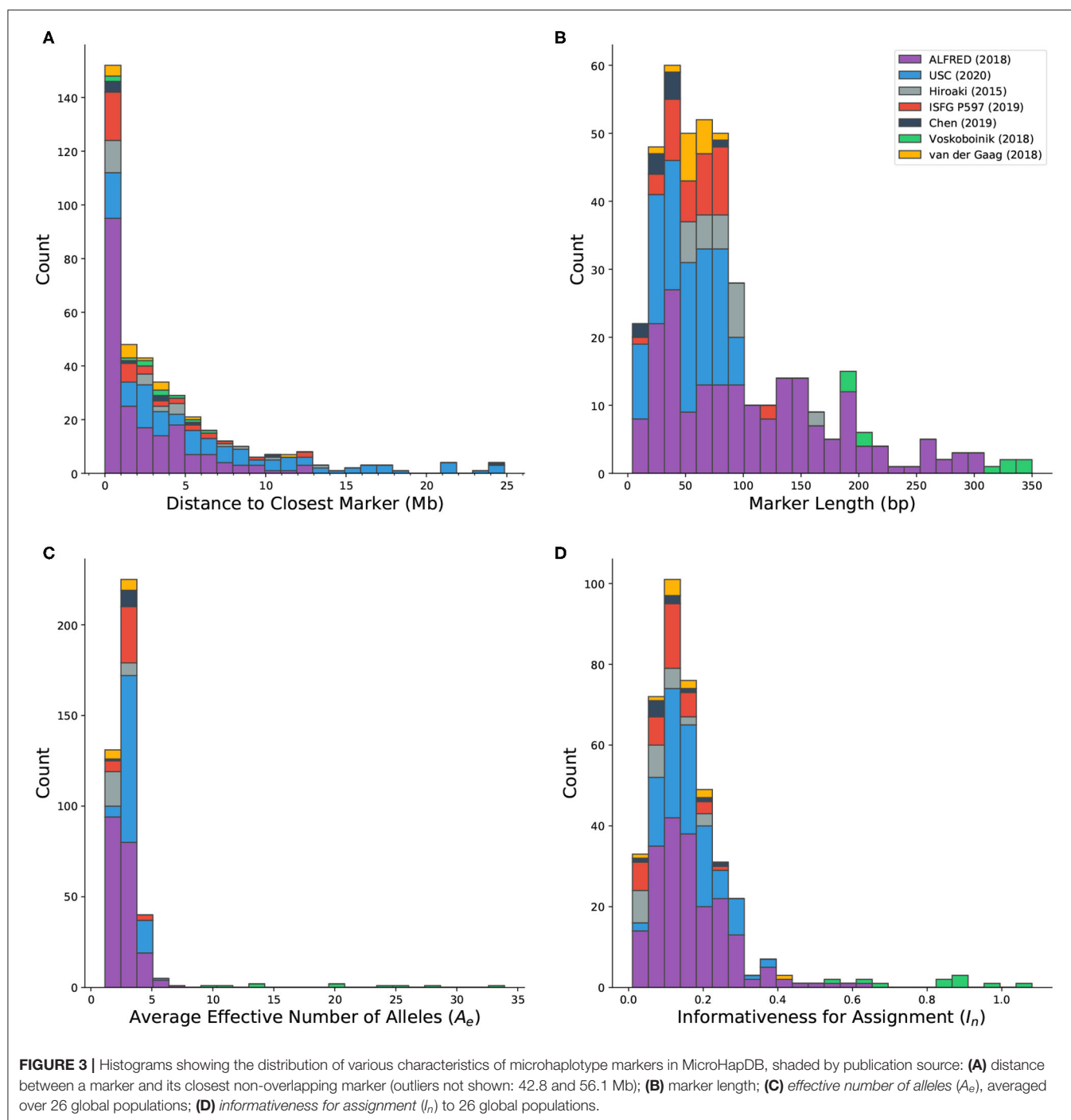
3. RESULTS

3.1. MicroHapDB Aggregates Data for More Than 400 Microhaplotypes

MicroHapDB version 0.6, released in June 2020, includes descriptions of 102 cohorts and population samples from six sources, 417 marker definitions from seven sources, and numerous population frequencies for 5,373 observed haplotypes from six sources, all together comprising a total of 113,995

³mh06PK-24844, mh0XUSC-XqD, mh11PK-63643, mh15PK-75170, and mh22PK-104638.

⁴With the exception of the 5 markers discussed in footnote 3.



records. This database represents a comprehensive collection of all microhaplotype (microhap) data published to date.

The number of single nucleotide polymorphisms (SNPs) used to define microhaps ranges from 2 to 49, with an average of 3.77 SNPs per marker. MicroHapDB includes 115 markers defined by two SNPs, 171 defined by three SNPs, 87 defined by four SNPs, 20 defined by five SNPs, 5 defined by six SNPs, and 19 defined by seven or more SNPs.

Microhap markers are defined on all autosomes as well as the X chromosome. Forty-five marker definitions overlap with

other markers. Most of these (40/45) were defined by Staadig and Tillmar (2019), which includes 11 exact duplicates and 29 probable adjustments to markers from other sources. The distance between each marker and the closest non-overlapping marker ranges between 143 bp and more than 56 Mb. Out of 417 markers in MicroHapDB, 152 (36.5%) reside within 1 Mb of their closest neighbor, and 53 (12.7%) reside within 100 kb of their closest neighbor (**Figure 3A**). Any set of markers separated by a small physical distance is likely in high linkage disequilibrium (LD) and the individual markers would therefore

not be independent. DNA profiles containing information for linked markers further complicates interpretation. This requires either sophisticated statistical modeling to account for the dependencies between markers or adopting an either/or strategy in which one of the loci is discarded when both produce reliable data.

Published microhaps occupy a wide range of lengths, with core marker length (the number of nucleotides spanning the most distal SNPs that define the marker, inclusive) ranging from 4 to 350 bp (**Figure 3B**). The majority of the microhaps in MicroHapDB (307/417; 73.6%) span <100 bp, and a substantial minority (145/417; 34.8%) span <50 bp.

3.2. Microhaplotype Frequencies for 26 Global Populations Enable Interpretation of Multi-Source Panels

Interpretation of any microhap typing result requires the use of appropriate microhaplotype frequency data. Prior to the release of MicroHapDB, availability of frequency data was inconsistent for published microhaps, with some sources providing frequencies for numerous population samples, while other sources providing frequencies for only a single population sample, or no frequency data at all.

MicroHapDB provides a comprehensive set of frequencies for all microhap markers to date. Estimated using 2,504 fully phased genotypes from Phase 3 of the 1,000 Genomes Project (Auton et al., 2015), the MicroHapDB database contains frequencies for 412 microhaplotype markers across 26 global population samples. A total of 113,477 frequencies for 5,373 alleles furnish a broad view of marker variation amongst Africans, admixed Americans, East Asians, Europeans, and South Asians.

3.3. MicroHapDB Provides Three Measures of Allelic Variation Within and Among Populations

The availability of microhaplotype frequency estimates across a standard set of 26 global population samples for all published microhaps provides a consistent means of comparing, ranking, and evaluating microhap markers for different applications. The per-marker effective number of alleles (A_e) was computed independently for each population, and then averaged across all 26 populations (section Materials and Methods). This statistic serves as a measure of within-population allelic diversity observed at a particular marker, and corresponds to the marker's power for individual identification. A previous study (Kidd and Speed, 2015) proposed an A_e threshold of 3, above which a microhaplotype can be considered "exceedingly useful" for forensic purposes. Average A_e values for microhaps in MicroHapDB range from 1.16 to 33.92, with a mean of 3.28 (**Figure 3C**). Most markers in MicroHapDB (238/412, 57.8%) have an average A_e below 3, and only 17 markers (4.1%) have an average A_e above 5.

Marker informativeness for assignment (I_n) was computed for the same 26 global populations (section Materials and Methods). This statistic serves as a measure of variation among populations, and corresponds to the marker's power

for predicting an individual's biogeographic ancestry. I_n values for markers in MicroHapDB fall between 0.01 and 1.08, with a mean of 0.17 (**Figure 3D**). Eight markers have an I_n value >0.682, the highest I_n value previously reported for a microhap (Kidd et al., 2018)—we note however that these I_n values were computed for a different set of populations and are therefore not directly comparable.

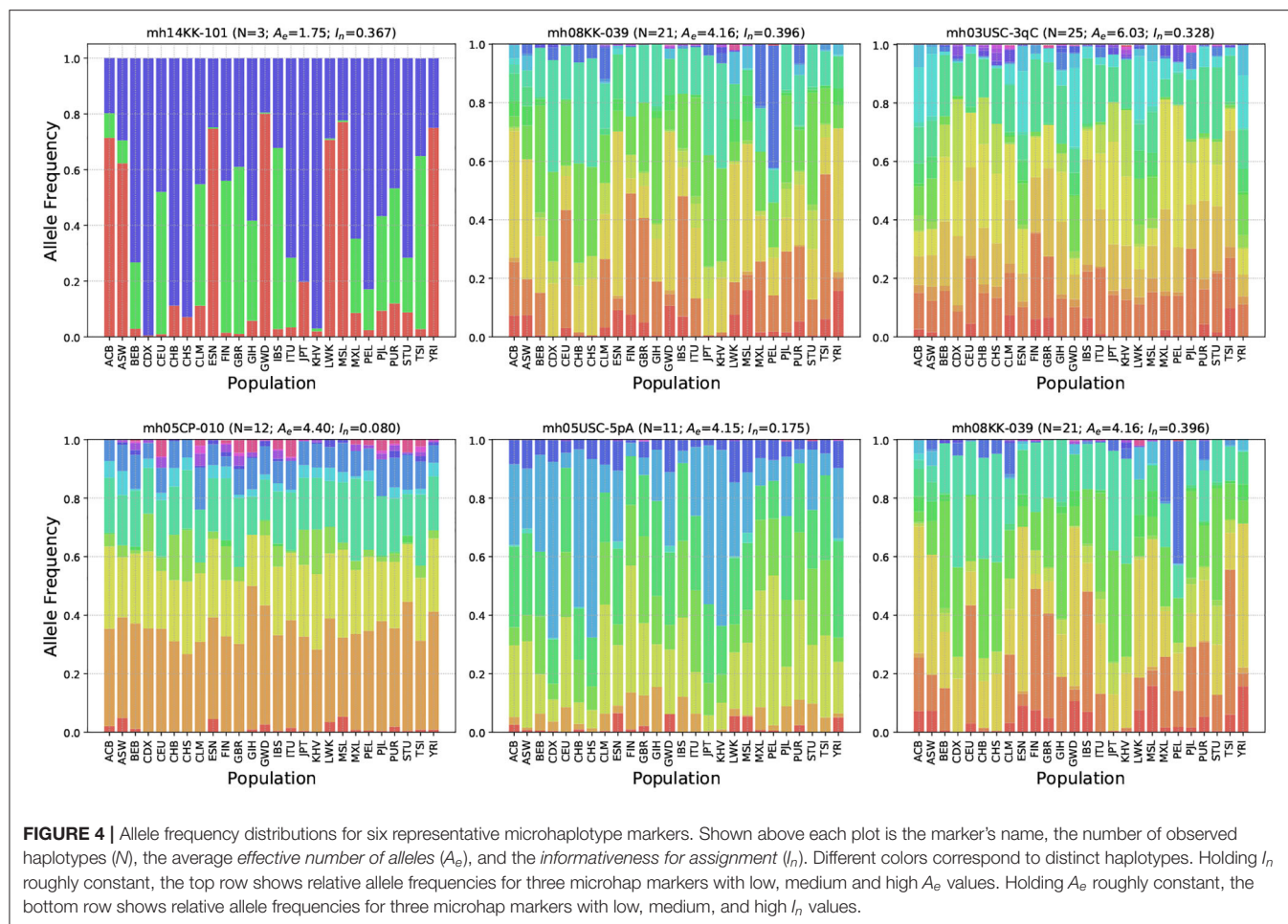
Figure 4 shows allele frequency distributions across the 26 populations for six representative microhap markers. The markers were selected from a range of A_e and I_n values to demonstrate how differences in these statistics are reflected in allele frequencies.

F_{ST} values were also computed for all markers. **Figure 5** shows the correlation between F_{ST} , A_e , and I_n for 391 markers. A weak positive correlation exists between A_e and I_n , while a weak negative correlation exists between A_e and F_{ST} . The latter trend suggests that while it's possible for an increase in allelic diversity to coincide with population-specific patterns of allele frequency, this is not generally the case for the microhap markers here considered. The strongest correlation is between I_n and F_{ST} , reflecting the sensitivity of these two statistics to population-specific allele distributions.

Ten highly polymorphic microhaps reported by Voskoboinik et al. (2018) stand out from microhaps published in other sources in several ways. Originally defined by locus boundaries rather than by explicit lists of SNP identifiers, these 10 microhaps include more SNPs (14–49 per marker) than any other marker in MicroHapDB. They have the highest average A_e values in MicroHapDB by a significant margin, and nine of these markers are included in MicroHapDB's top 10 microhaps ranked by I_n . It is worth noting that these microhaps are also among the longest, reflecting the study's distinct selection criteria and sequencing and evaluation strategy. The longest five markers in this set are also the five longest markers in MicroHapDB, and the remaining five are above the 89th percentile in length with respect to all markers in MicroHapDB.

4. DISCUSSION

In this study, we report the development of a comprehensive database of published microhaplotype (microhap) marker and frequency data. We describe the estimation of microhap frequencies in 26 global population samples, and the use of these frequencies to compute measures of allelic variation, enabling the ranking of microhaps for different forensic applications. This extensive collection of allele frequencies and ranking statistics will facilitate the design and interpretation of forensic panels that include markers from distinct sources without the need for extensive development of frequency data up front. MicroHapDB is a free open-access resource that contains information for all microhaps published as of February 2020. It requires minimal computing resources to install and maintain, and is designed to be easily extended with additional sources of public or private microhap data in the future. We hope MicroHapDB will democratize and accelerate advances in microhap-based forensics capabilities by enabling



researchers and companies to focus solely on marker discovery, or solely on population surveys, or solely on panel and kit design, without the need to invest in development of all of the above.

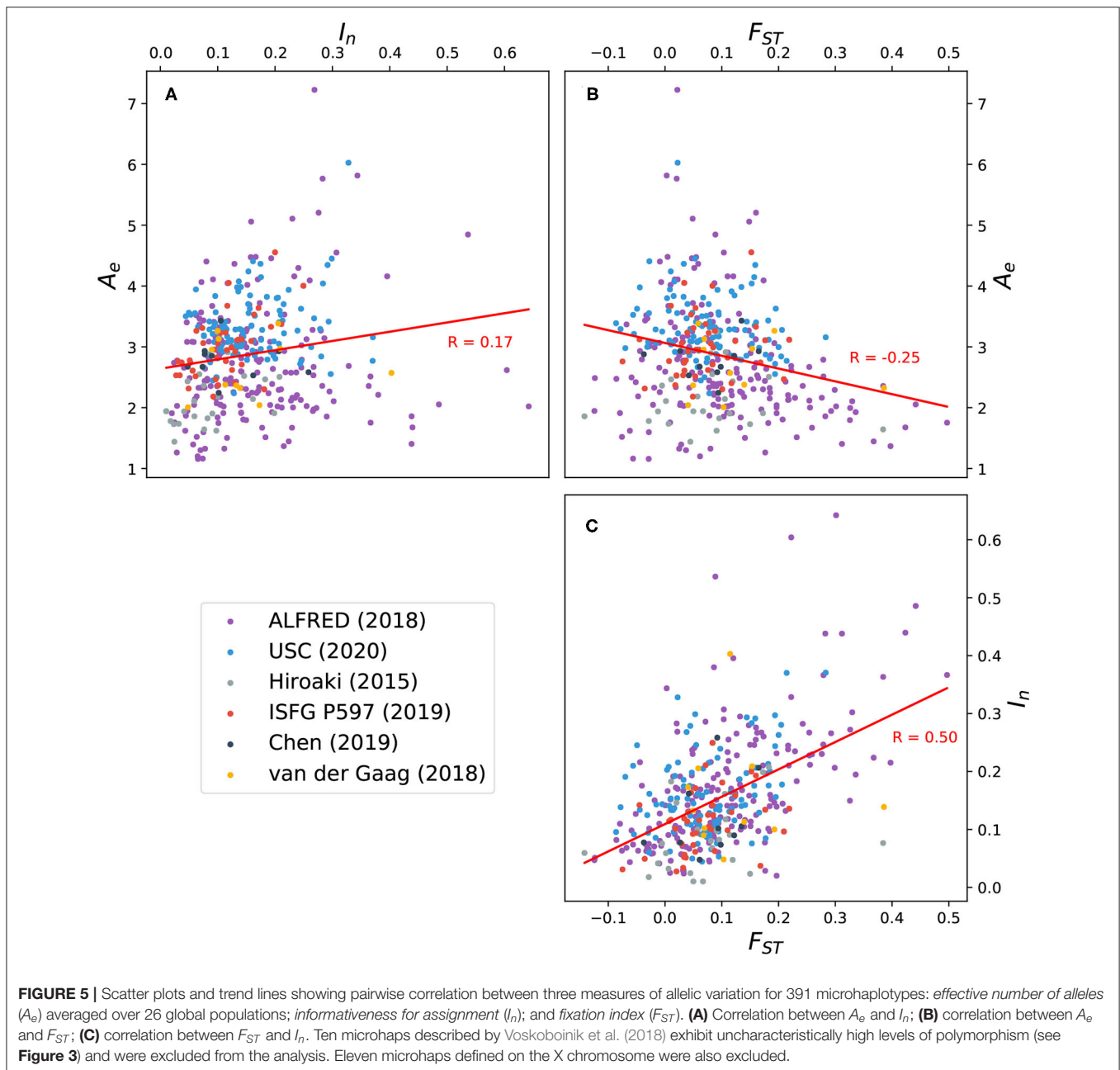
MicroHapDB provides A_e and I_n values for ranking microhaps for different forensic purposes. The genomic coordinates of each marker are also stored in MicroHapDB, enabling convenient calculation of physical distances between markers residing on the same chromosome. However, in addition to normal considerations that must always be addressed when designing a forensic DNA panel (e.g., amplicon sizes, primer kinetics, off-target amplification), correct interpretation of DNA profiles requires researchers to determine the independence of markers in a proposed panel based on the extent of linkage between the markers in the population(s) of interest. The length of candidate markers is also an important consideration depending on the sequencing technology utilized and the priority of recovering profiles from low input or low quality DNA samples.

Unlike conventional short tandem repeat (STR) markers, no comprehensive database analogous to the FBI's Combined DNA Index System (CODIS) databases (<https://www.fbi.gov/services/laboratory/biometric-analysis/codis>) yet exists for microhap

profiles. Constructing, populating, and performing requisite quality control for such a database will require substantial time and investment, which likely could only be pursued in earnest as the forensic community approaches consensus regarding optimal targets and assays. We expect that MicroHapDB can play a useful role in that process.

A major challenge in establishing a database like MicroHapDB, or a shared national database of microhap profiles, or indeed in communicating clearly about microhaps in the scientific literature, is the lack of consistency in nomenclature and in the way that markers are defined. A few papers describing microhap markers have used the nomenclature proposed by Kidd (2016), with marker names such as mh01KK-172, mh01CP-007, and mh06PK-25713. Other papers used a variety of *ad hoc* marker designators, such as 1 and MH02. MicroHapDB has adopted the Kidd nomenclature since its inception in 2018, and for sake of consistency has applied it to microhap collections where it was not previously used (e.g., mh01NH-01 for 1 and mh01AT-02 for MH02).

The question of how microhap markers are *defined* is at least as consequential as how they are *labeled*. A small number of published microhaps have been defined as a specific (but undisclosed) set of single nucleotide polymorphisms (SNPs)



at a genomic locus, the endpoints of which are indicated using coordinates on a reference genome assembly. Other microhaps are defined by a designator that refers to a set of SNPs at a particular locus, but whose specific component SNPs have been adjusted over time to improve the marker's performance. Ambiguous marker definitions of these kinds create substantial challenges for reproducibility and establishing provenance, and should be avoided. de la Puente et al. (2020) propose that microhaps should forgo marker names altogether (e.g., mh01USC-1pA or 1pA) in favor of an explicit list of SNP variants, as designated by dbSNP rsIDs (e.g., rs28503881, rs4648788, rs72634811, rs28689700).

We strongly endorse the sentiment behind this recommendation, although we concede the convenience of concise marker designators, especially when marker definitions are composed of dozens of SNP variants. What is most critical is the need for marker definitions to be *unambiguous* and *invariant over time*.

This discussion highlights the tension that has been provoked by the emergence of NGS technologies in forensics. Conventional assays have required the design of probes for specific SNP targets, which are often genotyped independently and then phased statistically. In contrast, NGS assays permit recovery of the entire sequence at a microhap locus and the simultaneous genotyping

and phasing of all its component SNPs, and indeed any additional intermediate (and often rare) SNPs. We anticipate that as NGS forensic assays become more routine, typing results for microhap assays will include all of the variants occurring in the sequenced genomic segment. This kind of typing result would have full “backwards compatibility” in that it could be used to determine the haplotype of *any* microhap marker explicitly defined at the locus. At the same time, full-coverage sequences of microhap loci will enable significant improvements in, e.g., mixture detection and deconvolution.

The question remains as to whether microhap designators should refer to *markers* (i.e., explicit sets of variants) or to *loci*. We suggest that minor addenda to the nomenclature proposed by Kidd (2016), such as the use of version numbers or other suffixes, would enable support for both. In the mean time, MicroHapDB searches based on genomic coordinates provide a convenient way to resolve spatial relationships between distinct marker definitions.

DATA AVAILABILITY STATEMENT

MicroHapDB version 0.6 has been archived in the Open Science Framework repository and is available at <https://osf.io/gr7h6>. The MicroHapDB database and software can be installed and updated from the Bioconda repository: <https://bioconda.github.io/> for more details. The publicly available datasets analyzed in this study can be found here: <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>.

AUTHOR CONTRIBUTIONS

DS conceived the study, constructed the database, implemented the software interface, and wrote the manuscript. DS and RM collected data, performed quality control, and edited and

approved the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded under Agreement No. HSHQDC-15-C-00064 awarded to Battelle National Biodefense Institute by the Department of Homeland Security Science and Technology Directorate (DHS S&T) for the management and operation of the National Biodefense Analysis and Countermeasures Center, a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security or the U.S. Government. The Department of Homeland Security does not endorse any products or commercial services mentioned in this presentation. In no event shall the DHS, BNBI, or NBACC have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. In addition, no warranty of fitness for a particular purpose, merchantability, accuracy, or adequacy is provided regarding the contents of this document.

ACKNOWLEDGMENTS

We want to thank three reviewers whose thoughtful feedback improved this article. We also express gratitude to Kenneth Kidd, Christopher Phillips, and Lev Voskoboinik for responding to inquiries about published data sets, and for insightful discussions about microhaplotypes. Finally, we thank Rebecca Just, Tim Stockwell, M. J. Rosovitz, and Adam Bazinet for their valuable feedback on drafts of this manuscript and early versions of the database.

REFERENCES

- 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, 68–74. doi: 10.1038/nature15393
- Bennett, L., Oldoni, F., Long, K., Cisana, S., Madella, K., Wootton, S., et al. (2019). Mixture deconvolution by massively parallel sequencing of microhaplotypes. *Int. J. Legal Med.* 133, 719–729. doi: 10.1007/s00414-019-02010-7
- Bleka, Ø., Benschop, C. C., Storvik, G., and Gill, P. (2016a). A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles. *Forens. Sci. Int.* 25, 85–96. doi: 10.1016/j.fsigen.2016.07.016
- Bleka, Ø., Storvik, G., and Gill, P. (2016b). EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. *Forens. Sci. Int.* 21, 35–44. doi: 10.1016/j.fsigen.2015.11.008
- Butler, J. M. (2010). *Fundamentals of Forensic DNA Typing*. San Diego, CA: Academic Press.
- Butler, J. M. (2015). *Advanced Topics in Forensic DNA Typing: Interpretation*. Oxford: Academic Press.
- Chen, P., Deng, C., Li, Z., Pu, Y., Yang, J., Yu, Y., Li, K., et al. (2019a). A microhaplotypes panel for massively parallel sequencing analysis of DNA mixtures. *Forens. Sci. Int.* 40, 140–149. doi: 10.1016/j.fsigen.2019.02.018
- Chen, P., Zhu, W., Tong, F., Pu, Y., Yu, Y., Huang, S., et al. (2019b). Identifying novel microhaplotypes for ancestry inference. *Int. J. Legal Med.* 133, 983–988. doi: 10.1007/s00414-018-1881-x
- Coble, M. D., and Bright, J.-A. (2019). Probabilistic genotyping software: an overview. *Forens. Sci. Int.* 38, 219–224. doi: 10.1016/j.fsigen.2018.11.009
- Cowell, R. G., Graversen, T., Lauritzen, S. L., and Mortera, J. (2015). Analysis of forensic DNA mixtures with artefacts. *J. R. Stat. Soc.* 64, 1–48. doi: 10.1111/rssc.12071
- 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. doi: 10.1126/science.aan8433
- Crow, J. F., and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York, NY: Harper & Row.
- de la Puente, M., Phillips, C., Xavier, C., Amigo, J., Carracedo, A., Parson, W., et al. (2020). Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems. *Forens. Sci. Int.* 45. doi: 10.1016/j.fsigen.2019.102213
- Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., et al. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* 15, 475–476. doi: 10.1038/s41592-018-0046-7
- Hiroaki, N., Koji, F., Tetsushi, K., Kazumasa, S., Hiroaki, N., and Kazuyuki, S. (2015). Approaches for identifying multiple-SNP haplotype blocks for use in human identification. *Legal Med.* 17, 415–420. doi: 10.1016/j.legalmed.2015.06.003

- Kidd, K. K. (2016). Proposed nomenclature for microhaplotypes. *Hum. Genomics* 10:16. doi: 10.1186/s40246-016-0078-y
- Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Wootton, S., and Chang, J. (2018). Selecting microhaplotypes optimized for different purposes. *Electrophoresis* 39, 2815–2823. doi: 10.1002/elps.201800092
- Kidd, K. K., and Speed, W. C. (2015). Criteria for selecting microhaplotypes: mixture detection and deconvolution. *Invest. Genet.* 6:1. doi: 10.1186/s13323-014-0018-3
- McKinney, W. (2011). pandas: a foundational python library for data analysis and statistics. Python for High Performance and Scientific Computing, 14.
- Miles, A., Ralph, P., Rae, S., and Pisupati, R. (2019). *scikit-allel v1.2.1: A Python Package for Exploring and Analysing Genetic Variation Data*.
- Oldoni, F., Hart, R., Long, K., Maddela, K., Cisana, S., Schanfield, M., et al. (2017). Microhaplotypes for ancestry prediction. *Forens. Sci. Int.* 6, e513–e515. doi: 10.1016/j.fsigss.2017.09.209
- Oldoni, F., Kidd, K. K., and Podini, D. (2019). Microhaplotypes in forensic genetics. *Forens. Sci. Int.* 38, 54–69. doi: 10.1016/j.fsign.2018.09.009
- 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, 1402–1422. doi: 10.1086/380416
- Ruiz, Y., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., de Cal, M. C., Cruz, R., et al. (2013). Further development of forensic eye color predictive tests. *Forens. Sci. Int.* 7, 28–40. doi: 10.1016/j.fsign.2012.05.009
- Staadig, A., and Tillmar, A. (2019). “Evaluation of microhaplotypes—a promising new type of forensic marker,” in *The 28th Congress of the International Society for Forensic Genetics* (Prague), P 597.
- van der Gaag, K. J., de Leeuw, R. H., Laros, J. F., den Dunnen, J. T., and de Knijff, P. (2018). Short hypervariable microhaplotypes: a novel set of very short high discriminating power loci without stutter artefacts. *Forens. Sci. Int.* 35, 169–175. doi: 10.1016/j.fsign.2018.05.008
- Voskoboinik, L., Motro, U., and Darvasi, A. (2018). Facilitating complex DNA mixture interpretation by sequencing highly polymorphic haplotypes. *Forens. Sci. Int.* 35, 136–140. doi: 10.1016/j.fsign.2018.05.001
- Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., et al. (2013). The hirisplex system for simultaneous prediction of hair and eye colour from DNA. *Forens. Sci. Int.* 7, 98–115. doi: 10.1016/j.fsign.2012.07.005
- Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370. doi: 10.1111/j.1558-5646.1984.tb05657.x
- Zhu, J., Lv, M., Zhou, N., Chen, D., Jiang, Y., Wang, L., et al. (2019). Genotyping polymorphic microhaplotype markers through the Illumina® MiSeq platform for forensics. *Forens. Sci. Int.* 39, 1–7. doi: 10.1016/j.fsign.2018.11.005

Disclaimer: This manuscript has been authored by Battelle National Biodefense Institute, LLC under Contract No. HSHQDC-15-C-00064 with the U.S. Department of Homeland Security. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

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

Copyright © 2020 Standage and Mitchell. 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 Comparison of Forensic Age Prediction Models Using Data From Four DNA Methylation Technologies

A. Freire-Aradas^{1*}, E. Pośpiech², A. Aliferi³, L. Girón-Santamaría¹, A. Mosquera-Miguel¹, A. Pisarek², A. Ambroa-Conde¹, C. Phillips^{1*}, M. A. Casares de Cal⁴, A. Gómez-Tato⁴, M. Spólnicka⁵, A. Woźniak⁵, J. Álvarez-Dios⁴, D. Ballard³, D. Syndercombe Court³, W. Branicki^{2,5}, Ángel Carracedo^{1,6} and M. V. Lareu¹

¹ Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Galicia, Spain, ² Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland, ³ King's Forensics, Department of Analytical, Environmental and Forensic Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom, ⁴ Faculty of Mathematics, University of Santiago de Compostela, Galicia, Spain, ⁵ Central Forensic Laboratory of the Police, Warsaw, Poland, ⁶ Fundación Pública Galega de Medicina Xenómica – CIBERER-IDIS, Santiago de Compostela, Spain

OPEN ACCESS

Edited by:

Cemal Gurkan,
Turkish Cypriot DNA Laboratory
(TCDL), Cyprus

Reviewed by:

Athina Vidaki,
Erasmus MC, Netherlands
Andrés Pérez-Figueroa,
University of Porto, Portugal

*Correspondence:

A. Freire-Aradas
ana.freire3@hotmail.com
C. Phillips
c.phillips@mac.com

Specialty section:

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

Received: 29 April 2020

Accepted: 27 July 2020

Published: 19 August 2020

Citation:

Freire-Aradas A, Pośpiech E,
Aliferi A, Girón-Santamaría L,
Mosquera-Miguel A, Pisarek A,
Ambroa-Conde A, Phillips C, Casares
de Cal MA, Gómez-Tato A,
Spólnicka M, Woźniak A,
Álvarez-Dios J, Ballard D, Court DS,
Branicki W, Carracedo Á and
Lareu MV (2020) A Comparison
of Forensic Age Prediction Models
Using Data From Four DNA
Methylation Technologies.
Front. Genet. 11:932.
doi: 10.3389/fgene.2020.00932

Individual age estimation can be applied to criminal, legal, and anthropological investigations. DNA methylation has been established as the biomarker of choice for age prediction, since it was observed that specific CpG positions in the genome show systematic changes during an individual's lifetime, with progressive increases or decreases in methylation levels. Subsequently, several forensic age prediction models have been reported, providing average age prediction error ranges of ± 3 –4 years, using a broad spectrum of technologies and underlying statistical analyses. DNA methylation assessment is not categorical but quantitative. Therefore, the detection platform used plays a pivotal role, since quantitative and semi-quantitative technologies could potentially result in differences in detected DNA methylation levels. In the present study, we analyzed as a shared sample pool, 84 blood-based DNA controls ranging from 18 to 99 years old using four different technologies: EpiTYPER®, pyrosequencing, MiSeq, and SNaPshot™. The DNA methylation levels detected for CpG sites from *ELOVL2*, *FHL2*, and *MIR29B2* with each system were compared. A restricted three CpG-site age prediction model was rebuilt for each system, as well as for a combination of technologies, based on previous training datasets, and age predictions were calculated accordingly for all the samples detected with the previous technologies. While the DNA methylation patterns and subsequent age predictions from EpiTYPER®, pyrosequencing, and MiSeq systems are largely comparable for the CpG sites studied, SNaPshot™ gives bigger differences reflected in higher predictive errors. However, these differences can be reduced by applying a z-score data transformation.

Keywords: epigenetics, DNA methylation, age estimation, EpiTYPER®, pyrosequencing, MiSeq, SNaPshot™

INTRODUCTION

DNA methylation is the most widely studied epigenetic mark in the human genome (Jones, 2012). Methylation, which is the incorporation of a methyl group at cytosine-guanine dinucleotide motifs, has been shown to be highly correlated with the human aging process (Bocklandt et al., 2011; Hannum et al., 2013; Horvath, 2013; Johansson et al., 2013), and as a consequence, is currently

considered the most accurate age prediction biomarker. The estimation of an individual's epigenetic age in this way is useful in several areas. From a forensic point of view, the ability to accurately predict the chronological age of the donor of a biological sample can provide relevant information in order to guide police investigation in cases where no suspects or matches in the DNA database are found (Freire-Aradas et al., 2017). From a clinical point of view, biological age estimation may help to determine the life expectancy of the individual (Horvath and Raj, 2018). In order to infer the chronological age, several age prediction models have been developed based on data generated using different DNA methylation technologies, including EpiTYPER® (Xu et al., 2015; Freire-Aradas et al., 2016; Zubakov et al., 2016), pyrosequencing (Weidner et al., 2014; Bekaert et al., 2015; Zbieć-Piekarska et al., 2015), massively parallel sequencing (MPS) (Naue et al., 2017; Vidaki et al., 2017; Aliferi et al., 2018) and SNaPshot™ (Lee et al., 2015; Hong et al., 2017; Jung et al., 2019) systems. As DNA methylation is quantitative in nature, potential differences in DNA methylation levels can be detected by each technology. For this reason, systematic comparisons of methylation detection technologies using a common set of controls become necessary.

A shared characteristic of these aforementioned DNA methylation technologies is their reliance on bisulfite conversion of the analyzed samples. Bisulfite conversion is a pretreatment of the genomic DNA that converts all the unmethylated cytosines to uracil, which after PCR, are replaced with thymine; while the methylated CpGs remain unaltered; converting a methylation difference into a sequence difference (Frommer et al., 1992; Clark et al., 1994). Two main limitations are related to the bisulfite conversion process. First, it degrades the DNA and consequently, a larger amount of genomic DNA than usual must be used. Second, it reduces DNA sequence variability, diminishing the multiplexing capacity of the technologies used. In spite of these constraints, the high quality of the corresponding DNA methylation measurements favors their use in current forensic applications.

EpiTYPER® detects and quantifies DNA methylation based on MALDI-TOF mass spectrometry (Ehrich et al., 2006). EpiTYPER® analyses the cleaved fragments with varied molecular weights depending on the methylation status around each fragment, which are measured and the corresponding DNA methylation levels ascertained. Although single-nucleotide CpG positions are measured, some CpGs very close to each other are detected as a block providing an average methylation value for a set of two to three CpG sites. Pyrosequencing is considered the gold standard method for measuring targeted DNA methylation (Lehmann and Tost, 2015). It is an accurate quantitative sequencing-by-synthesis method based on luminescence, following an enzymatic cascade that uses the production of light after pyrophosphate release when a nucleotide is incorporated onto a growing DNA strand. Despite its high accuracy, it is difficult to analyze multiple markers simultaneously, hindering its use in forensic casework where the quality/quantity of DNA samples is usually restricted. MPS (also called next generation sequencing

or NGS) is a high-throughput technology that sequences multiple target-specific regions in parallel and quantitatively detects DNA methylation levels by the ratio of sequence read coverage amongst targets (Richards et al., 2018). For forensic casework, two main companies provide the equipment necessary to run DNA methylation analysis with MPS technology: Illumina using the MiSeq system, and Thermo Fisher using the Ion S5 detector. MiSeq technology uses sequencing-by-synthesis where a fluorescently labeled reversible terminator is imaged as each dNTP is added, and then cleaved to allow incorporation of the next base. Ion Torrent technology is based on semiconductor sequencing, i.e., each time a nucleotide is incorporated in the growing DNA strand, a proton is released and the consequent variation in pH is measured as a change in electrical conductivity. MPS appears to be a highly accurate technology for DNA methylation analysis while allowing for multiplexing. Nevertheless, the high cost associated with both the equipment and reagents is a constraining factor for some forensic laboratories that require a more cost-effective technology such as single base extension (SBE) that can be easily incorporated into well-established capillary electrophoresis systems. SBE (also called minisequencing or SNaPshot™) is a semi-quantitative technology based on fluorescence (Fondevila et al., 2017). It consists of the annealing of an unlabeled oligonucleotide that matches the sequence immediately adjacent to the target nucleotide site. The subsequent incorporation of a single complementary fluorescently labeled terminator ddNTP produces a sequence strand extended by one nucleotide. While the multiplex capacity of this method is an advantage, the different fluorescence intensities between each of the dyes linked to the ddNTPs used, can potentially bias the methylation values detected.

Up until now, age prediction models have been developed based on data collected using one technology, i.e., if an age prediction model is built based on pyrosequencing data, the subsequent test samples are also analyzed by pyrosequencing, and so on, since some loss of accuracy was previously reported for inter-technology data exchange (Vidaki et al., 2017; Aliferi et al., 2018). This represents a constraint, since each technology requires the re-building of the prediction model with new age reference sample sets. We aimed to compare DNA methylation data from different technologies in order to explore if platform-independent models might be useful for forensic age prediction. The study of Hong et al. (2019) has already introduced this concept by developing a platform-independent model for MPS and SNaPshot™ in saliva samples. In the present study, we cover a further inter-technology comparison for DNA methylation based on MPS and SNaPshot™ technologies, but adding methods based on EpiTYPER® and pyrosequencing. A total of 84 common control DNAs from blood between 18 and 99 years old were analyzed using the four different technologies for three CpG sites in *ELOVL2*, *FHL2*, and *MIR29B2* genes. The corresponding DNA methylation levels were compared, and several age prediction models were subsequently tested in the common samples detected with either the same or different technologies to the system used to build the training set.

MATERIALS AND METHODS

DNA Samples and DNA Methylation Data

A total of 84 blood sample-derived DNA extracts were obtained from healthy European volunteers ranging in age from 18 to 99 years. The samples were used as the testing set (referenced as common controls). All samples were obtained from the ‘Carlos III’ Spanish National DNA Bank, University of Salamanca, and ethical approval was granted from the ethics committee of investigation in Galicia, Spain (CAEI: 2013/543). All DNA samples were quantified by Qubit® dsDNA High Sensitivity (HS) Assay Kit (Thermo Fisher) and subsequently normalized to 10 ng/μL. Additionally, DNA methylation data for a total of 1130 European blood samples ranging in age from 2 to 104 years were selected from previous studies that used EpiTYPER® ($N = 725$) (Freire-Aradas et al., 2016), pyrosequencing ($N = 293$) (Zbieć-Piekarska et al., 2015), and MiSeq ($N = 112$) (Aliferi et al., 2018) for building the training sets. Moreover, a total of 105 European blood samples ranging in age from 18 to 75 years were analyzed using SNaPshot™ in order to build the corresponding training set.

CpG Sites Selection and DNA Methylation Detection

A total of three age-correlated genes were used for comparative purposes: *ELOVL2*, *FHL2*, and *MIR29B2* – loci that have been commonly included in age prediction models analyzing blood-based DNA samples (Garagnani et al., 2012; Bekaert et al., 2015; Zbieć-Piekarska et al., 2015; Freire-Aradas et al., 2016; Park et al., 2016; Zubakov et al., 2016; Hong et al., 2019; Jung et al., 2019). In the present study, these three genes were analyzed by three independent laboratories using four DNA methylation technologies: EpiTYPER®, pyrosequencing, MiSeq, and SNaPshot™. Table 1 describes the overlap between CpG sites and DNA methylation technologies used in this study. All CpGs are single CpG sites, with the exception of *MIR29B2_C1* that consisted of a cluster of three CpG sites, since EpiTYPER® could not give individual results for each site. Therefore, an average of the three corresponding CpG sites was used when comparing the corresponding DNA methylation values with pyrosequencing or MiSeq for this case. Regarding the overlap in Table 1, *ELOVL2*, *FHL2*, and *MIR29B2_C1* were used for comparing EpiTYPER®, pyrosequencing, and MiSeq; whereas SNaPshot™ comparisons were made in a separate analysis.

In this case, *ELOVL2*, *FHL2*, and *MIR29B2_C2* were used for comparing EpiTYPER® and MiSeq systems with the SNaPshot™ system. Analyses were extended to more than one CpG per gene in the case of *MIR29B2* due to a lack of complete overlap between technologies. *ELOVL2* and *FHL2* were represented by the same CpG site in all analyses. All four DNA methylation technologies require a pretreatment with sodium bisulfite. Three bisulfite kits were used according to the methylation detection technology that will be described below. Supplementary Table S1 summarizes additional variable factors between technologies.

Agena Bioscience EpiTYPER® DNA Methylation Analysis

The Agena Bioscience EpiTYPER® system (San Diego, CA, United States) used PCR amplicons of 362 base pairs (bp) for *ELOVL2*, 191 bp for *FHL2* and 344 bp for *MIR29B2*. Samples analyzed using EpiTYPER® were bisulfite converted using the EZ DNA Methylation™ Kit (Zymo Research) using 300 ng of genomic DNA. A detailed description of the EpiTYPER® workflow has been previously reported (Freire-Aradas et al., 2016, 2018). Methylation data were obtained using EpiTYPER® software v.1.2.22 (Agena Bioscience).

Pyrosequencing

Pyrosequencing of the PCR amplicons used for this technology were 308 bp for *ELOVL2*, 167 bp for *FHL2* and 146 bp for *MIR29B2*. Bisulfite conversion was performed using the Qiagen 96-well bisulfite conversion kit (Qiagen, Hilden, Germany) using 1 μg of genomic DNA. Specific procedures for the pyrosequencing workflow were previously outlined (Zbieć-Piekarska et al., 2015). This technology was performed using a PyroMark vacuum prep workstation and a PyroMark Q24 instrument (Qiagen), following the manufacturer's guidelines. The data were automatically analyzed using PyroMark analysis software (Qiagen, Hilden, Germany).

The MiSeq System

Massively parallel sequencing-based detection of methylated DNA was performed using the MiSeq system (Illumina). Samples detected using MiSeq were bisulfite converted using the MethylEdge® Bisulfite Conversion System (Promega Corporation, Fitchburg, WI, United States) using 50 ng of genomic DNA. Detailed information regarding the workflow can be found in Aliferi et al. (2018). The amplicon sizes used were 308 bp for *ELOVL2*, 165 bp for *FHL2* and 210 bp for

TABLE 1 | Overlap between CpG sites from the target genes *ELOVL2*, *FHL2*, and *MIR29B2* in the four evaluated DNA methylation technologies: (A) EpiTYPER®, (B) Pyrosequencing, (C) MiSeq, and (D) SNaPshot™.

Gene	CpG_ID	GRCh38 position	(A) EpiT	(B) Pyros	(C) MiSeq	(D) SNaP
<i>ELOVL2</i>	cg21572722	chr6:11044661	✓	✓	✓	✓
<i>FHL2</i>	cg06639320	chr2:105399282	✓	✓	✓	✓
<i>MIR29B2_C1</i>	–/cg10501210/–	chr1:207823672/75/81	✓	✓	✓	
<i>MIR29B2_C2</i>	–	chr1:207823715	✓		✓	✓

All CpGs are single sites, except *MIR29B2_C1* that is a cluster of three CpG sites. The corresponding DNA methylation values for *MIR29B2_C1* were calculated as the average of the three CpG sites for each technology.

MIR29B2. Analysis of the FASTQ files was conducted with the Burrows-Wheeler Aligner, Sequence Alignment/Map and Genome Analysis Toolkit software following guidelines from Aliferi et al. (2018).

Single Base Extension (SBE)

Single base extension was performed using the SNaPshotTM Multiplex Kit (Thermo Fisher) in replicate analyses. PCR amplicon sizes were 111 bp for *ELOVL2*, 108 bp for *FHL2* and 49 bp for *MIR29B2*. Samples for SNaPshotTM analyses were bisulfite converted using the MethylEdge[®] Bisulfite Conversion System (Promega Corporation, Fitchburg, WI, United States, assay B) using 100 ng of genomic DNA. Specific multiplex protocol details are summarized in **Supplementary Table S2**. Methylation values were calculated based on the peak height ratio [methylated signal/(methylated signal + unmethylated signal)] obtained with GeneMapperID v3.2.1 software, measuring RFU values (relative fluorescence units). The average of the DNA methylation values between replicates were used for SNaPshotTM analyses.

Statistical Analyses

Comparisons of DNA methylation measurement methods were performed using Bland-Altman plots using the *BlandAltmanLeh* R package (Lehnert, 2015). These plots represent the difference between paired measures for the same variable against the corresponding mean (Giavarina, 2015). The upper and lower limits of agreement (LoA) were calculated as the mean of the differences between two measurements ± 1.96 times the standard deviation (SD) between them, accordingly, in order to include 95% of the differences within them (Bland and Altman, 1999). A limit of acceptance, or threshold of ± 0.098 has been established *a priori* ($\pm 1.96 \cdot \text{SD}$, $\text{SD} = 0.05$). Normality was tested using a Shapiro-Wilk normality test. Uniformity, i.e., absence of tendency for DNA methylation differences between methods, was checked using regression analysis (*p*-value for the fitted regression line). Differences in DNA methylation levels between technologies were also explored using analysis of variance (ANOVA).

All age prediction modeling was based on quantile regression (QR) (Freire-Aradas et al., 2016; Smeers et al., 2018). The original datasets for the training sets were composed of a higher number of samples and CpG sites than the data required for the reported analyses; specifically: $N = 725$, 18–104 years, 7 CpGs; $N = 293$, 2–75 years, 5 CpGs and $N = 112$, 11–93 years, 12 CpGs measured by EpiTYPER[®], pyrosequencing, and MiSeq, respectively (Zbieć-Piekarska et al., 2015; Freire-Aradas et al., 2016; Aliferi et al., 2018). Therefore, model re-building was performed in order to harmonize age distribution and sample size, as well as to cover only the three CpG sites under study. The age range was restricted to 18–75 years old for all the training sets. Age range restriction directly led to $N = 100$ for MiSeq. In the case of EpiTYPER[®] and pyrosequencing, besides age range restriction, random selection of a maximum of two individuals per year-of-age led to $N = 116$ for EpiTYPER[®] and $N = 106$ for pyrosequencing. Quantiles 0.1 and 0.9 (q10 and q90) were used for the development of the multivariate QR model

using the *quantreg* R package (Koenker et al., 2019). Random cleavage of the input data for the QR model validation was done using the *cvTools* R package (Alfons, 2015). Validation of the QR model was performed using *k*-fold cross-validation ($k = 10$). The corresponding predictive accuracy was measured with the following performance metrics: median absolute prediction error (MAE); root-mean-square error (RMSE); percent of correct predicted samples with a prediction error of ± 5 years (%CP ± 5) and percent of correct predicted samples within the prediction intervals (%CP \pm PI). Predicted versus chronological age was plotted using the *ggplot2* R package (Wickham and Chang, 2019). Z-score transformation was performed scaling the DNA methylation levels to the corresponding mean and SD. All calculations were performed using R software v3.4.2.

RESULTS

Intra-Technology Variation

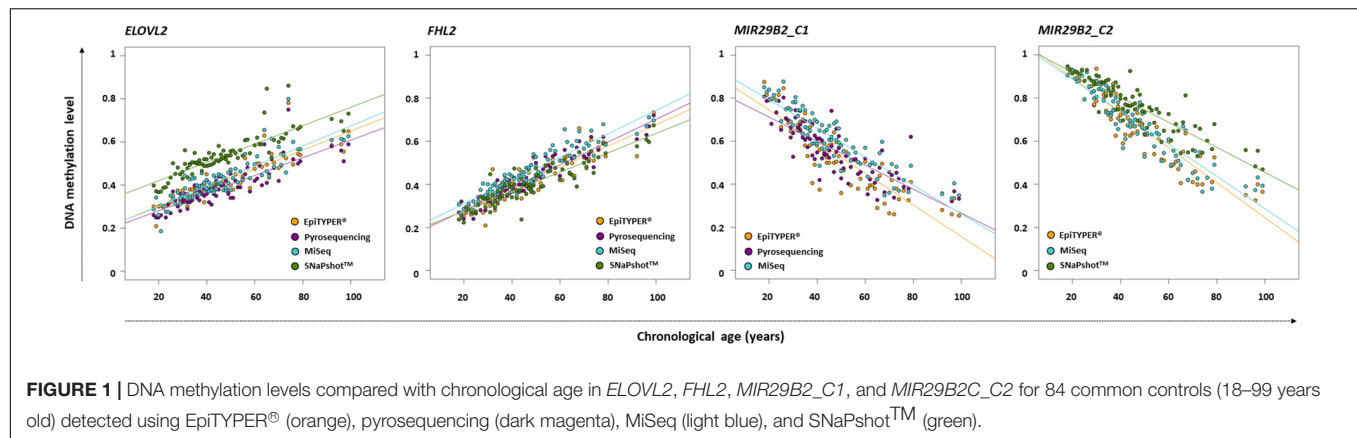
Intra-technology variation was assessed analyzing two replicates for the semi-quantitative technology used in the present study, i.e., SNaPshotTM. Previous work on EpiTYPER[®] (Freire-Aradas et al., 2016), pyrosequencing (Zbieć-Piekarska et al., 2015), and MiSeq (Aliferi et al., 2018) showed absence of technical variation for these DNA methylation technologies and therefore, the corresponding common controls were analyzed using a single replicate.

Supplementary Figure S1 depicts the DNA methylation levels against the chronological age for *ELOVL2*, *FHL2*, and *MIR29B2_C2* for both replicates for the 84 common controls. An absence of statistically significant differences between replicates (*p*-value > 0.01) allowed the study to use the average DNA methylation levels for SNaPshotTM analyses.

Comparison of DNA Methylation Levels for EpiTYPER[®] vs. Pyrosequencing vs. MiSeq vs. SNaPshotTM

DNA methylation analysis for *ELOVL2*, *FHL2*, and *MIR29B2* for the 84 common controls was performed using EpiTYPER[®], pyrosequencing, MiSeq, and SNaPshotTM technologies. The corresponding DNA methylation data (β -values) is shown in **Supplementary Table S3**. **Figure 1** shows the DNA methylation levels against the chronological age for *ELOVL2*, *FHL2*, *MIR29B2_C1*, and *MIR29B2_C2* for the overlapping technologies, while **Table 2** describes the corresponding ANOVA test. The major differences were displayed by *ELOVL2* and *MIR29B2_C2* detected using SNaPshotTM technology. Additionally, moderate statistically significant differences were also found for *MIR29B2_C1* comparing EpiTYPER[®] vs. MiSeq (*p*-value: 0.00303).

Figure 2 depicts the corresponding paired Bland-Altman plots using the previous DNA methylation values. The central dotted gray line represents the mean of the differences; while the discontinuous gray lines represent the upper and lower LoA, including the 95% of differences between one measurement and the other. The red line indicates the theoretical 'no differences



between methods'. Bland-Altman differences presented a normal distribution for *ELOVL2* and *FHL2* for all pairwise comparisons with the exception of those compared to SNaPshot™ technology (Shapiro–Wilk normality test). Regarding *MIR29B2_C1*, an absence of normality for EpiTYPER® vs. pyrosequencing and for pyrosequencing vs. MiSeq was observed (p -value < 0.01). In *MIR29B2_C2*, an absence of normality was found for MiSeq vs. SNaPshot™. Uniformity was found for both *ELOVL2* and *FHL2*, but not for *MIR29B2*. Absence of uniformity was detected for *MIR29B2_C1* for EpiTYPER® vs. pyrosequencing (Figure 2M) and EpiTYPER® vs. MiSeq (Figure 2N), as well as for *MIR29B2_C2* for EpiTYPER®

vs. SNaPshot™ (Figure 2P) and for MiSeq vs. SNaPshot™ (Figure 2R) (p -value < 0.01). In particular, Figure 2M shows a tendency to overestimate DNA methylation levels with EpiTYPER® vs. pyrosequencing at high values, while DNA methylation levels between 0.5 and 0.2 are underestimated using this technology for *MIR29B2_C1* (Figure 1). Regarding *MIR29B2_C2*, when comparing EpiTYPER® vs. SNaPshot™ (Figure 2P) and for MiSeq vs. SNaPshot™ (Figure 2R), similar DNA methylation values are observed at high values that gradually diverge when DNA methylation values between 0.6 and 0.3 are detected. An additional bias was observed for *MIR29B2_C1* for pyrosequencing vs. MiSeq (Figure 2O), that is explained by an underestimation by pyrosequencing or an overestimation by MiSeq of the DNA methylation levels (Figure 1).

If excluding SNaPshot™ comparisons, the mean of the differences between the DNA methylation levels detected using different technologies for *ELOVL2* and *FHL2* for all pairwise comparisons was quite close to zero (average: ± 0.03). SNaPshot™ comparisons for *FHL2* also detected reduced DNA methylation differences between technologies (average: $+0.03$). However, higher deviations were detected for *ELOVL2* (average: -0.12) due to an overestimation of the DNA methylation levels using SNaPshot™ compared to the other three technologies (Figure 1). This explains a raised value for the lower LoA for *ELOVL2* when including SNaPshot™ analyses (average lower LoA: -0.22) – with values that exceed the established threshold (± 0.098). Regarding *MIR29B2* (C1 or C2), for analyses not including SNaPshot™ data, the mean of the differences was reduced (average: -0.04), as the comparison with SNaPshot™ significantly increased these differences (average: -0.085).

Age Prediction for DNA Methylation Data From EpiTYPER®, Pyrosequencing, and MiSeq Using *ELOVL2*, *FHL2*, and *MIR29B2_C1*

The three CpG sites (*ELOVL2*, *FHL2*, and *MIR29B2_C1*) were used for age estimation of the 84 common controls using the reference training sets based on EpiTYPER®, pyrosequencing, and MiSeq. Figure 3 shows the predicted age vs. chronological

TABLE 2 | ANOVA test for evaluation of the differences between the four DNA methylation technologies studied.

	<i>p</i> -value
<i>ELOVL2</i>	
EpiTYPER® vs. pyrosequencing	0.066
EpiTYPER® vs. MiSeq	0.447
Pyrosequencing vs. MiSeq	0.012
EpiTYPER® vs. SNaPshot™	10^{-13}
Pyrosequencing vs. SNaPshot™	2×10^{-16}
MiSeq vs. SNaPshot™	6.1×10^{-11}
<i>FHL2</i>	
EpiTYPER® vs. pyrosequencing	0.636
EpiTYPER® vs. MiSeq	0.0113
Pyrosequencing vs. MiSeq	0.0369
EpiTYPER® vs. SNaPshot™	0.309
Pyrosequencing vs. SNaPshot™	0.128
MiSeq vs. SNaPshot™	0.000257
<i>MIR29B2_C1</i>	
EpiTYPER® vs. pyrosequencing	0.476
EpiTYPER® vs. MiSeq	0.00303
Pyrosequencing vs. MiSeq	0.0111
<i>MIR29B2_C2</i>	
EpiTYPER® vs. MiSeq	0.631
EpiTYPER® vs. SNaPshot™	7.63×10^{-5}
MiSeq vs. SNaPshot™	0.000313

p-values marked in bold are statistically significant (p -value < 0.01).

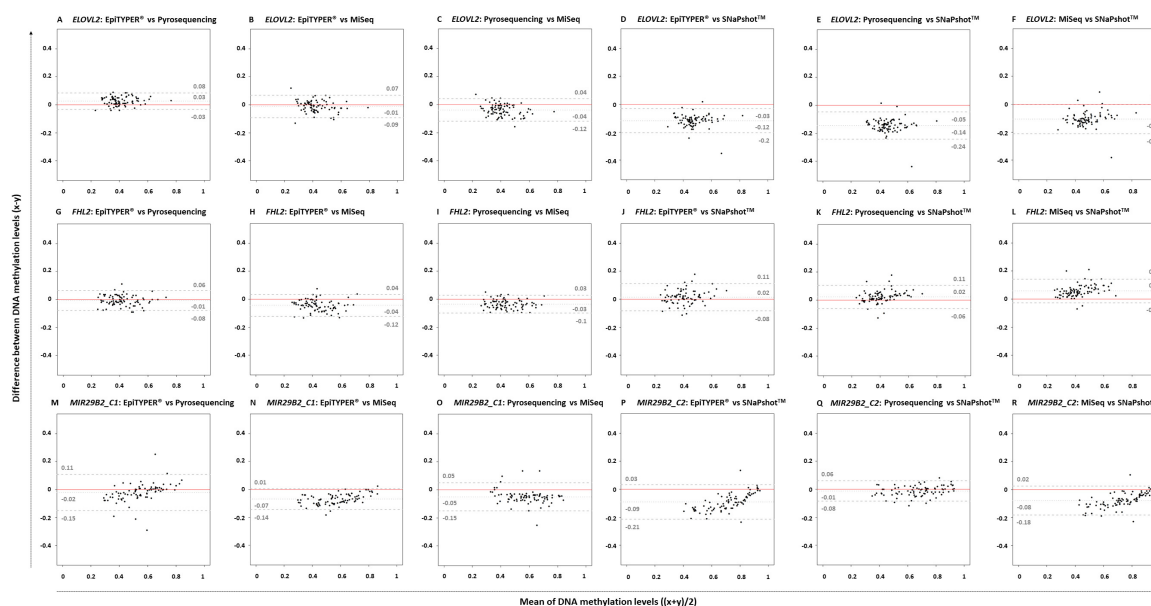


FIGURE 2 | Bland-Altman plots comparing pairs of four DNA methylation technologies: EpiTYPER®, pyrosequencing, MiSeq, and SNaPshot™ in *ELOVL2*, *FHL2*, *MIR29B2_C1*, and *MIR29B2_C2* for 84 common controls (18–99 years old). The plots represent the differences between paired methods ($x-y$) against the mean of both paired methods $[(x+y)/2]$ for the following pairs: (A,G,M) EpiTYPER® (x) vs. pyrosequencing (y), (B,H,N) EpiTYPER® (x) vs. MiSeq (y), (C,I,O) pyrosequencing (x) vs. MiSeq (y), (D,J,P) EpiTYPER® (x) vs. SNaPshot™ (y), (E,K,Q) pyrosequencing (x) vs. SNaPshot™ (y), and (F,L,R) MiSeq (x) vs. SNaPshot™ (y). The central dotted gray line represents the mean of the differences; while the discontinuous gray lines represent the upper and lower LoA. The red line indicates the theoretical no differences between methods.

age for both training and testing sets, and **Table 3**, the corresponding performance metrics.

All four training sets – EpiTYPER®, pyrosequencing, MiSeq and the combined training set derived from the combination of the three technologies (**Figures 3A,E,I,M**, respectively) – provided errors lower than ± 5 years (MAE: from ± 3.14 to ± 4.11) and correct prediction rates higher than 70% (%CP \pm PI: from 74.02 to 77.92%). However, for the aim of the present study, an intra-training set comparison rather than an inter-training set comparison was performed, i.e., we compared testing data from the three DNA methylation technologies analyzed using a uniform training set.

From the data modeled using the EpiTYPER® training set, no statistically significant differences were found for the prediction errors shown by the common controls analyzed with EpiTYPER®, pyrosequencing or MiSeq (p -value > 0.01). However, a general underestimation of age for common controls older than 60 years was detected in pyrosequencing (**Figure 3C**). Samples analyzed with MiSeq provided the best prediction rates (%CP \pm PI: 75%).

The prediction model using the pyrosequencing training set gave no statistically significant differences for the prediction errors in the technologies analyzing common controls (p -value > 0.01). In this case, both pyrosequencing and MiSeq underestimated common control age for the whole age range (**Figures 3G,H**). In spite of this, samples detected with MiSeq provided the best prediction rates (%CP \pm PI: 84.52%).

The prediction model using the MiSeq training set gave no statistically significant differences for the prediction errors in the

technologies analyzing common controls (p -value > 0.01). As with the EpiTYPER® training set, a global underestimation of age for common controls older than 60 years was detected in pyrosequencing (**Figure 3K**). The best prediction rates were again provided by MiSeq detection (%CP \pm PI: 85.71%).

In view of the similarities found for prediction errors and the accurate predictions displayed for common controls, all data from the previous training sets, e.g., EpiTYPER®, pyrosequencing, and MiSeq, were combined in order to create a new enlarged platform-independent training set. As before, no statistically significant differences were found for the common control prediction errors in any technology used (p -value > 0.01). In common with individual training sets, common controls older than 60 years old were underestimated by pyrosequencing (**Figure 3O**) and samples detected with MiSeq gave the best prediction rates (%CP \pm PI: 84.52%).

Age Prediction for DNA Methylation Data From EpiTYPER®, MiSeq, and SNaPshot™ Using *ELOVL2*, *FHL2*, and *MIR29B2_C2*

In a subsequent analysis using a different set of CpG sites (*ELOVL2*, *FHL2*, and *MIR29B2_C2*), age prediction was assessed and results plotted in **Figure 4**, showing the predicted age versus the chronological age using common controls detected with EpiTYPER®, MiSeq, and SNaPshot™. **Table 4** summarizes the corresponding performance metrics.

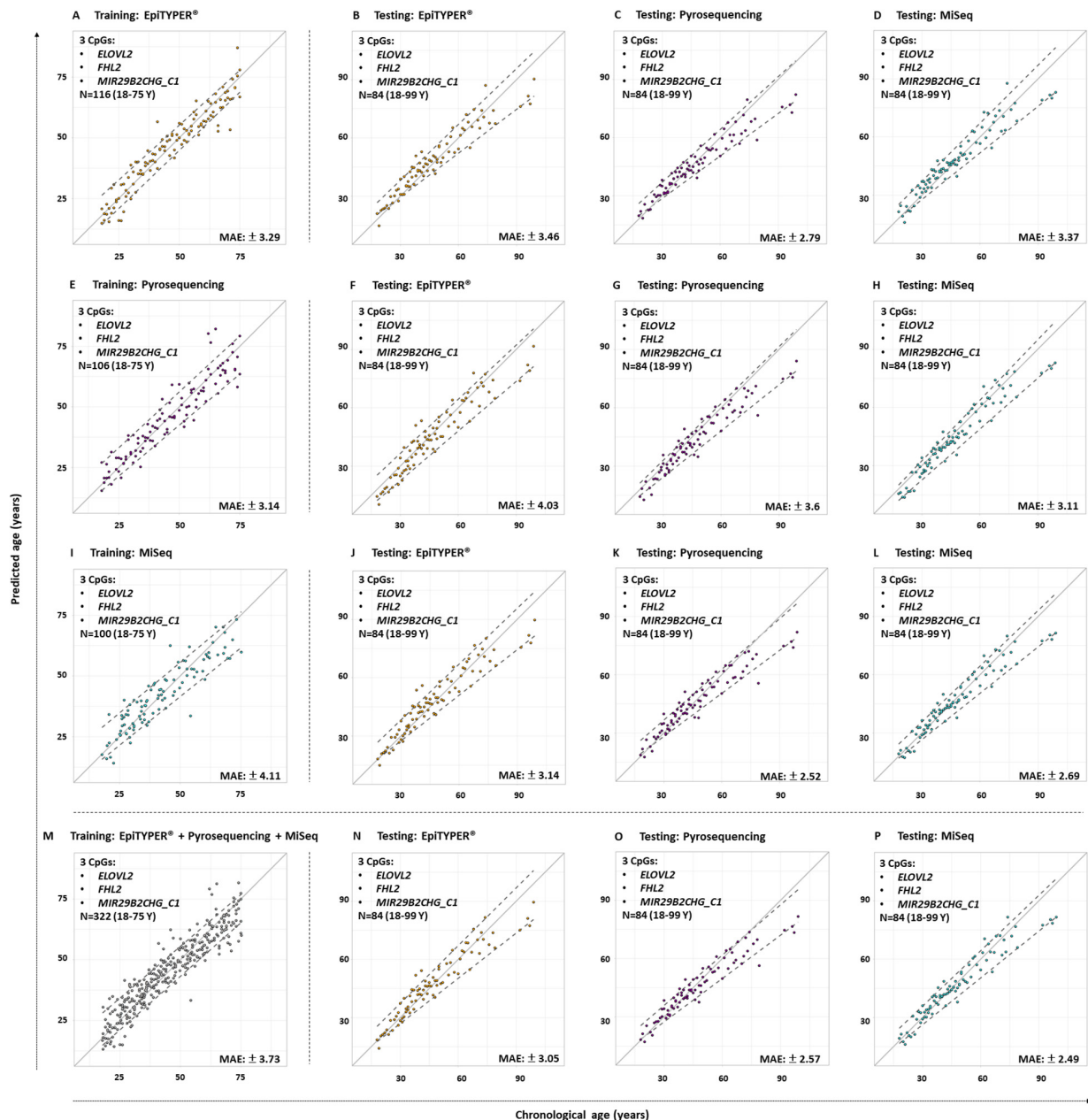


FIGURE 3 | Predicted versus chronological age using four training sets from three DNA methylation technologies using 3-CpG-site models (*ELOVL2*, *FHL2*, and *MIR29B2_C1*) for the 84 common controls (18–99 years old). **(A)** EpiTYPER® training set, **(B–D)** EpiTYPER®, pyrosequencing, and MiSeq testing sets analyzed with the EpiTYPER® training set, **(E)** pyrosequencing training set, **(F–H)** EpiTYPER®, pyrosequencing, and MiSeq testing sets analyzed with the pyrosequencing training set; **(I)** MiSeq training set, **(J–L)** EpiTYPER®, pyrosequencing, and MiSeq testing sets analyzed with the MiSeq training set; **(M–P)** EpiTYPER®, pyrosequencing, and MiSeq test sets analyzed with the combined training set. The continuous gray line represents perfect correlation. The discontinuous gray lines represent the prediction intervals.

All four training sets – EpiTYPER®, MiSeq, SNaPshot™ and the combined training set derived from the combination of the previous three technologies (**Figures 4A,E,I,M**, respectively) – provided errors lower than ± 4 years (MAE: from ± 2.98 to ± 3.83) and correct prediction rates higher than 75% (%CP \pm PI: from 76.59 to 79.12%). **Figures 4B.1–4P.1** represent the corresponding testing sets.

For the prediction model using the EpiTYPER® training set, statistically significant differences were found for the prediction errors in the common controls (p -value < 0.01). These differences were explained by a higher error rate in SNaPshot™ analysis of common controls (MAE: ± 4.71), which was reflected in a decreased correct prediction rate (%CP \pm PI: 44.05%). The best predictions were obtained in EpiTYPER® analysis

TABLE 3 | Age predictive performance metrics for the training and test sets based on the analysis of three CpG sites (*ELOVL2*, *FHL2*, and *MIR29B2_C1*) using EpiTYPER®, pyrosequencing and MiSeq DNA methylation technologies, as well as a combination of all three technologies (Combined).

Technology	Group	MAE			
		(years)	RMSE	%CP ± 5	%CP ± PI
EpiTYPER®	Training (N = 116)	±3.29	5.15	70.53%	74.02%
EpiTYPER®	Testing (N = 84)	±3.46	5.99	67.07%	70.73%
Pyrosequencing	Testing (N = 84)	±2.79	6.48	69.51%	73.17%
MiSeq	Testing (N = 84)	±3.37	5.74	69.05%	75%
Pyrosequencing	Training (N = 106)	±3.14	5.9	67.55%	76.73%
EpiTYPER®	Testing (N = 84)	±4.03	6.36	62.2%	78.05%
Pyrosequencing	Testing (N = 84)	±3.6	6.79	64.63%	81.71%
MiSeq	Testing (N = 84)	±3.11	5.72	72.62%	84.52%
MiSeq	Training (N = 100)	±4.11	6.33	57%	76%
EpiTYPER®	Testing (N = 84)	±3.14	6.08	68.29%	80.49%
Pyrosequencing	Testing (N = 84)	±2.52	6.65	69.51%	85.37%
MiSeq	Testing (N = 84)	±2.69	5.55	76.19%	85.71%
Combined	Training (N = 322)	±3.73	5.86	63.99%	77.92%
EpiTYPER®	Testing (N = 84)	±3.05	5.93	71.95%	80.49%
Pyrosequencing	Testing (N = 84)	±2.57	6.62	74.39%	80.49%
MiSeq	Testing (N = 84)	±2.49	5.52	77.38%	84.52%

MAE: median absolute prediction error, RMSE: root-mean-square error, %CP: percent correct prediction, PI: prediction intervals.

of common controls (%CP ± PI: 75.61%). Although similar errors to those of the initial MiSeq analyses of the training set were found here (MAE: ± 3.31), the correct prediction rate was reduced (%CP ± PI: 58.33%) due to an overestimation of common controls younger than 45 years old (Figure 4C.1).

In spite of not detecting statistically significant differences between the common controls modeled using the MiSeq training set, the correct prediction rate for the SNaPshot™ samples was reduced (%CP ± PI: 51.19%). The best predictions were shown by the MiSeq data (%CP ± PI: 85.71%).

Data modeled using the SNaPshot™ training set presented the highest statistically significant differences (p -value = $2e^{-16}$). Test samples analyzed using either EpiTYPER® or MiSeq displayed high errors (MAE: ±14 and ±11.59, respectively), as well as minimum correct prediction rates (%CP ± PI: 31.71 and 47.62%, respectively).

In spite of these differences, a model using all the previous training sets was combined into a single platform-independent training set (Figure 4M). This combined model harmonized the data derived from all the technologies where prediction errors had no statistically significant differences (p -value > 0.01). Accordingly, similar correct prediction rates were obtained for all common controls detected using EpiTYPER®, MiSeq, and SNaPshot™ (82.93, 78.57, and 77.38%, respectively).

Due to the differences encountered for SNaPshot™ analyses when compared to EpiTYPER® and MiSeq, a z-score transformation was applied in order to check if the corresponding predictions could be improved by data scaling (Table 5). The application of a z-score transformation removed the previously encountered statistical differences. The EpiTYPER® and MiSeq test sets were markedly improved when

modeled with the SNaPshot™ training set (Figures 4J,2,K,2, %CP ± PI: 78.05 and 75% in comparison to the previous 31.71 and 47.62%, respectively). Similarly, the SNaPshot™ test set substantially improved when modeled with EpiTYPER® and MiSeq training sets (Figures 4D,2,H,2, %CP ± PI: 77.38 and 85.71% in comparison to previous values of 44.05 and 51.19%, respectively).

DISCUSSION

Correlation has been proposed as a statistical technique in order to compare technologies (Hong et al., 2019). However, this parameter evaluates the relationship or association between one variable and another, not their differences. In order to compare the differences between two measurement methods in our study, Bland-Altman analysis was applied (Giavarina, 2015). Bland-Altman analysis describes the degree of agreement between two quantitative technologies for the same variable (Bland and Altman, 1999) – in our case, between four DNA methylation detection methods. With this analysis, 95% of the differences between two methods are plotted within the LoA. It is important to consider that the maximum accepted LoA should be established before the analysis, according to analytical or biological criteria. Since some intra-technical variance was already accepted for DNA methylation ($SD \leq 0.05$ for replicates) (Freire-Aradas et al., 2016), inter-technical deviation based on the Bland-Altman's LoA and previous intra-technical variance was explored in the present study ($\pm 1.96 SD = \pm 0.098$). In addition to the LoA, it is also recommended that the differences between technologies are normally distributed, although not essential. However, uniformity is required before data can be used interchangeably.

Four DNA methylation technologies were compared using four CpG sites from *ELOVL2*, *FHL2*, and *MIR29B2*. Normality was analyzed for all the pairwise comparisons. Differences were normally distributed for *ELOVL2* and *FHL2* for all pairwise comparisons, except for *ELOVL2* in EpiTYPER®/pyrosequencing/MiSeq vs. SNaPshot™, and *MIR29B2* presented partial normal distribution. However, absence of normality can be handled in our study since subjects were not chosen randomly, but to give a wide distribution of the factor measured. The critical parameter that is required to exchange data among technologies without affecting the outcome is uniformity (Bland and Altman, 1999). Uniformity can be observed as the absence of a tendency for the differences to change between methods, i.e., the extent of the differences is uniformly maintained independently of the magnitude of the variable. In our study, uniformity is measured as the variance in the differences across the range of DNA methylation levels, and this should be maintained. The *ELOVL2* and *FHL2* CpG sites in the present work displayed complete uniformity for all comparisons made. The same cannot be said for *MIR29B2*. This marker had an evident tendency to show changes in differences in several comparisons (Figures 2M,N,P,R; p -value < 0.1). Some of them are explained by similar DNA methylation detection at high DNA methylation levels that gradually diverge – increasing

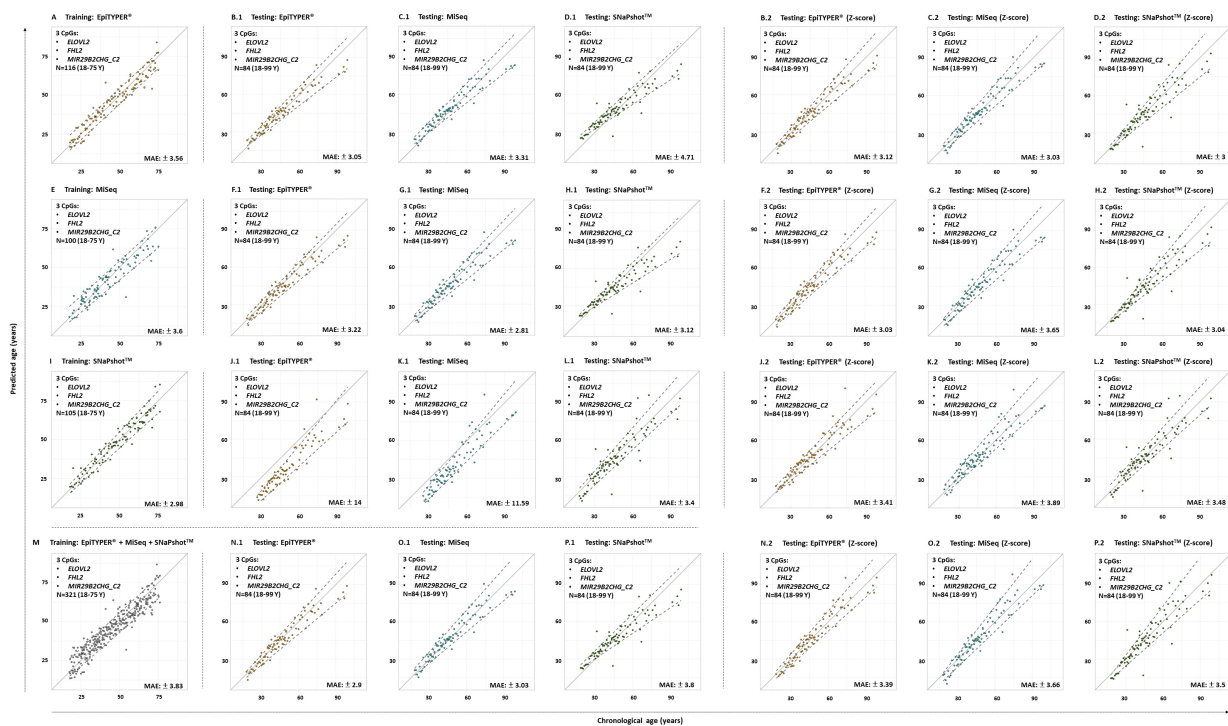


FIGURE 4 | Predicted versus chronological age using four training sets from three DNA methylation technologies and a 3-CpG-site model (*ELOVL2*, *FHL2*, and *MIR29B2_C2*) for the 84 common controls (18–99 years old). **(A)** EpiTYPER[®] training set, **(B–D)** EpiTYPER[®], MiSeq, and SNaPshot[™] testing sets analyzed with the EpiTYPER[®] training set, **(E)** MiSeq training set, **(F–H)** EpiTYPER[®], MiSeq, and SNaPshot[™] testing sets analyzed with the MiSeq training set, **(I)** SNaPshot[™] training set, **(J–L)** EpiTYPER[®], MiSeq and SNaPshot[™] testing sets analyzed with the SNaPshot[™] training set, **(M)** Combined training set, **(N–P)** EpiTYPER[®], MiSeq and SNaPshot[™] testing sets analyzed with the combined training set. Panels **(B.1–P.1)** correspond to untransformed data, while panels **(B.2–P.2)** correspond to z-score transformed data. The continuous gray line represents the perfect correlation. The discontinuous gray lines represent the prediction intervals.

the value of the differences when DNA methylation levels between 0.6 and 0.3 are detected (**Figure 1**, *MIR29B2_C2*). However, *MIR29B2_C1* for MiSeq vs. pyrosequencing behaves in the opposite way; i.e., bigger differences are found at high DNA methylation levels, that progressively decrease when DNA methylation values at about 0.3 are detected (**Figure 1**).

With regard to the DNA methylation levels, the highest levels of similarity were displayed by the quantitative technologies of EpiTYPER[®], pyrosequencing and MiSeq, especially for *ELOVL2* and *FHL2*. Nevertheless, some differences among these technologies were found in *MIR29B2*, although not statistically significant (p -value > 0.01) (**Figures 2M–R**) as they slightly exceeded the LoA from the established threshold of ± 0.098 (average lower LoA: -0.15). However, it can be concluded from our findings that when differences are uniformly within an established LoA, then methodologies can be used interchangeably (Bland and Altman, 1999). In order to test if the differences observed in the DNA methylation values are critical or not, the corresponding age predictions were performed using four re-configured age prediction models constructed using DNA methylation data detected with EpiTYPER[®] (Freire-Aradas et al., 2016), pyrosequencing (Zbieć-Piekarska et al., 2015), and MiSeq (Aliferi et al., 2018). In this way, despite differences encountered in *MIR29B2*, the prediction accuracy of the corresponding age prediction models when comparing

EpiTYPER[®], pyrosequencing, and MiSeq was not affected (**Figure 3** and **Table 3**). Nevertheless, it is important to note the underestimation of predicted age using pyrosequencing (**Figures 3C,G,K,O**) to test common controls older than 60 years.

In addition to this, previous replication experiments had indicated reproducibility for DNA methylation levels detected using EpiTYPER[®] (Bocklandt et al., 2011), MPS (Pabinger et al., 2016), and pyrosequencing (Bocklandt et al., 2011) compared with data from Infinium BeadChip arrays. This is an important factor to consider, since discovery studies are predominantly based on Infinium arrays, and identified age-associated CpG sites are subsequently validated using targeted DNA methylation technologies.

Different patterns are obtained when SNaPshot[™] analyses are included. SNaPshot[™] is a semi-quantitative method and this is reflected in the differences detected for estimated methylation levels. Due to a lack of overlap among the *MIR29B2* CpG sites between technologies, an independent comparison was performed in order to include SNaPshot[™] analyses. The major differences were found for *ELOVL2* and *MIR29B2_C2* when comparing either EpiTYPER[®] or MiSeq with SNaPshot[™] (**Figures 2D–F,P–R**). The main difference between both markers is the lack of uniformity for *MIR29B2_C2*, with a tendency to generate more differences when detecting DNA methylation levels between 0.6 and 0.3 (**Figure 1**). However, differences for

TABLE 4 | Age predictive performance metrics for the training and test sets based on the analysis of three CpG sites (*ELOVL2*, *FHL2*, and *MIR29B2_C2*) using EpiTYPER®, MiSeq, and SNaPshot™ DNA methylation technologies, as well as a combination of all three technologies (Combined).

Technology	Group	MAE (years)	RMSE	%CP ± 5	%CP ± PI
EpiTYPER®	Training (N = 116)	±3.56	4.88	72.2%	76.59%
EpiTYPER®	Testing (N = 84)	±3.05	5.32	71.95%	75.61%
MiSeq	Testing (N = 84)	±3.31	5.77	64.29%	58.33%
SNaPshot™	Testing (N = 84)	±4.71	7.78	53.57%	44.05%
MiSeq	Training (N = 100)	±3.6	6	65%	77%
EpiTYPER®	Testing (N = 84)	±3.22	5.73	69.51%	74.39%
MiSeq	Testing (N = 84)	±2.81	5.83	77.38%	85.71%
SNaPshot™	Testing (N = 84)	±3.12	7.96	67.86%	51.19%
SNaPshot™	Training (N = 105)	±2.98	4.43	76.09%	77%
EpiTYPER®	Testing (N = 84)	±14	14.93	4.88%	31.71%
MiSeq	Testing (N = 84)	±11.59	12.49	14.29%	47.62%
SNaPshot™	Testing (N = 84)	±3.4	8.2	58.33%	60.71%
Combined	Training (N = 321)	±3.83	5.56	64.45%	79.12%
EpiTYPER®	Testing (N = 84)	±2.9	5.27	68.29%	82.93%
MiSeq	Testing (N = 84)	±3.03	5.52	73.81%	78.57%
SNaPshot™	Testing (N = 84)	±3.8	7.43	63.1%	77.38%

MAE: median absolute prediction error, RMSE: root-mean-square error, %CP: percent correct prediction, PI: prediction intervals.

TABLE 5 | Age predictive performance metrics based on a z-score transformation for the training and test sets analyzing three CpG sites (*ELOVL2*, *FHL2*, and *MIR29B2_C2*) and using EpiTYPER®, MiSeq, and SNaPshot™ DNA methylation technologies, plus the combination of all three technologies (Combined).

Technology	Group	MAE (years)	RMSE	%CP ± 5	%CP ± PI
EpiTYPER®	Training (N = 116)	±3.56	4.88	72.2%	76.59%
EpiTYPER®	Testing (N = 84)	±3.12	5.08	75.61%	80.49%
MiSeq	Testing (N = 84)	±3.03	5.41	75%	80.95%
SNaPshot™	Testing (N = 84)	±3	7.09	70.24%	77.38%
MiSeq	Training (N = 100)	±3.6	6	65%	77%
EpiTYPER®	Testing (N = 84)	±3.03	5.4	70.24%	90.24%
MiSeq	Testing (N = 84)	±3.65	5.75	63.1%	88.1%
SNaPshot™	Testing (N = 84)	±3.04	7.35	76.09%	85.71%
SNaPshot™	Training (N = 105)	±2.98	4.43	76.09%	77%
EpiTYPER®	Testing (N = 84)	±3.41	6.15	70.73%	78.05%
MiSeq	Testing (N = 84)	±3.89	6.61	66.67%	75%
SNaPshot™	Testing (N = 84)	±3.48	7.81	71.43%	77.38%
Combined	Training (N = 321)	±3.83	5.56	64.45%	79.12%
EpiTYPER®	Testing (N = 84)	±3.39	5.49	69.51%	85.37%
MiSeq	Testing (N = 84)	±3.66	6.09	64.29%	80.95%
SNaPshot™	Testing (N = 84)	±3.5	7.67	65.48%	82.14%

MAE: median absolute prediction error, RMSE: root-mean-square error, %CP: percent correct prediction, PI: prediction intervals.

ELOVL2, although present, are almost proportional between methods (Figure 1). On the other hand, SNaPshot™ analysis of *FHL2* provided more similarities when compared with EpiTYPER® or MiSeq (mean of the differences: +0.02 and +0.06, respectively). Differences encountered for SNaPshot™ could be

explained by differences in the intensity of the fluorochromes; as previously reported by Hong et al. (2019). In spite of the differences in *ELOVL2* and *MIR29B2_C2*; it is important to note that all three markers, including *FHL2*, were genotyped using CT dyes and detected using an ABI3130. The CT dyes used in SNaPshot™ are characterized by more closely matched intensities in terms of fluorescence, and theoretically should provide unbiased DNA methylation values, more similar to those obtained using quantitative technologies than sites detected with the AG SNaPshot™ dyes. Our results agree this assumption only for *FHL2*, so additional factors are likely to be affecting the results for *ELOVL2* and *MIR29B2_C2*. The effect of such differences is reflected in the age prediction accuracies (Table 4). Either using the EpiTYPER® or MiSeq training set, with the worst predictions obtained for data analyzed using SNaPshot™. In view of these results, SNaPshot™ data cannot be used with prediction models based on EpiTYPER® or MiSeq technology. However, it has been observed that if expanding the training set to data from the three technologies, i.e., EpiTYPER®, MiSeq, and SNaPshot™ (combined training set), SNaPshot™ common controls are correctly predicted at a similar rate to those detected with EpiTYPER® or MiSeq.

Although the training sets used in the prediction models from the different DNA methylation technologies were harmonized in terms of sample size, age distribution and the underlying statistical model, factors potentially affecting technical variation such as bisulfite conversion, DNA input, amplicon length or PCR cycles should be taken into account (Supplementary Table S1). One of the main factors affecting methylation results is the efficiency of bisulfite conversion. The acid pH and high temperatures accompanying this molecular process lead to DNA fragmentation. It has been observed that different DNA degradation rates can be encountered if using different bisulfite conversion kits (Kint et al., 2018). Since fragmentation usually leads to sequences smaller than 500 bp (Grunau et al., 2001), a reduced amplification of longer amplicons could occur, although the exact effects remain unknown. Differences in DNA input could also affect results, although this should be linked to the manufacturer's recommendations as well as to the levels of technical optimization achieved for each methodology. Variation in DNA methylation levels could also be explained by biological variation. Although in order to minimize this effect, common test samples were represented by a single individual per year (18–99 years old), biological variation cannot be discounted since differences in white blood cell composition could alter DNA methylation levels.

CONCLUDING REMARKS

To the best of our knowledge, this is the first study covering the broadest possible comparison between DNA methylation technologies currently applied to forensic age prediction. Interchangeability of methylation data was found to be a suitable strategy when differences in the DNA methylation levels from different technologies do not exceed the uniformity threshold established by this study of ±0.098 (±1.96.SD, SD = 0.05), and

maintain this uniformity across the range of DNA methylation values detected. If the differences slightly exceed the threshold, it should be confirmed that these variations are not relevant for age estimation. Although the CpG sites for *ELOVL2*, *FHL2*, and *MIR292B* covered by the present study provide high accuracy for age prediction for most of the comparisons performed, in *MIR292B* the LoA is exceeded, and a lack of uniformity is consistently observed. Therefore, DNA methylation data for *MIR292B* should not be used independently of the technology applied. These deviations could be explained by internal technical problems, which we have observed for this gene (additional methylation studies with publications in preparation). In *ELOVL2* and *FHL2*, similar patterns of DNA methylation for EpiTYPER®, pyrosequencing, and MiSeq were observed, and subsequently data from these techniques can be used in platform-independent age prediction models. However, our results are linked to specific CpG positions – so no general extrapolations can be assumed. If additional CpG sites from *ELOVL2* and *FHL2* are considered for inclusion in technology-free age prediction models, the necessary validation tests should be made.

SNaPshot™ is a semi-quantitative technology based on fluorescence using dyes with different signal intensities. This introduces a bias in the DNA methylation values detected that explains the differences found for *ELOVL2* and *MIR292B_C2* in SNaPshot™ compared with EpiTYPER® or MiSeq, which subsequently decrease the accuracy of corresponding age predictions.

If differences are encountered between technologies; two viable corrective measures could be applied, as proposed by previous studies: (a) a z-score transformation in order to solve batch effects (Feng et al., 2018) or; (b) the addition of an extra covariate in the model indicating the type of technology used, then introducing a correction for the method (Hong et al., 2019). When a z-score transformation was tested in the present study it markedly improved the results when SNaPshot™ data was included in the analyses. If applying a platform-independent model, there is a risk of losing age prediction accuracy if the underlying sample size does not match the sample size of the original age prediction model. Further work to increase the number of samples tested among technologies will be necessary to detect if prediction accuracies are affected by different sample sizes. On the other hand, a re-analysis of the corresponding training set using the technology of interest would be also be a viable approach.

REFERENCES

- Alfons, A. (2015). *Package “cvTools”: Cross-Validation Tools for Regression Models*.
 Aliferi, A., Ballard, D., Gallidabino, M. D., Thurtle, H., Barron, L., and Syndercombe Court, D. (2018). DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. *Foren. Sci. Int. Genet.* 37, 215–226. doi: 10.1016/j.fsigen.2018.09.003
 Bekaert, B., Kamalandua, A., Zapico, S. C., Van De Voorde, W., and Decorte, R. (2015). Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. *Epigenetics* 10, 922–930. doi: 10.1080/15592294.2015.1080413
 Bland, J., and Altman, D. (1999). Measuring agreement in method comparison studies. *Stat. Methods Med. Res.* 8, 135–160. doi: 10.1191/096228099673819272

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/Supplementary Material.

ETHICS STATEMENT

Ethical approval was granted from the Ethics Committee of investigation in Galicia, Spain (CAEI: 2013/543).

AUTHOR CONTRIBUTIONS

AF-A, CP, EP, AA, WB, DB, DC, ÁC, and ML devised the experiments. AF-A, EP, AA, LG-S, AM-M, AP, AA-C, MS, and AW performed the laboratory work. AF-A, EP, AA, MC, AG-T, JÁ-D, DB, and WB compiled and analyzed the data. AF-A and CP wrote the manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

AF-A was supported by a postdoctorate grant funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (Modalidade B, ED481B 2018/010).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00932/full#supplementary-material>

FIGURE S1 | DNA methylation levels against the chronological age for *ELOVL2*, *FHL2*, and *MIR292B_C2* for both SNaPshot™ replicates for the 84 common controls (18–99 years old).

TABLE S1 | Summary of variable factors between DNA methylation technologies.

TABLE S2 | Specific multiplex protocol details for SNaPshot™.

TABLE S3 | DNA methylation data (β -values) for *ELOVL2*, *FHL2*, *MIR292B_C1*, and *MIR292B_C2* for the 84 common controls (18–99 years old) analyzed using EpiTYPER®, pyrosequencing, MiSeq, and SNaPshot™.

- Bocklandt, S., Lin, W., Sehl, M., Sánchez, F., Sinsheimer, J., Horvath, S., et al. (2011). Epigenetic predictor of age. *PLoS One* 6:e14821. doi: 10.1371/journal.pone.0014821
 Clark, S. J., Harrison, J., Paul, C. L., and Frommer, M. (1994). High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* 22, 2990–2997. doi: 10.1093/nar/22.15.2990
 Ehrlich, M., Correll, D., and Van Den Boom, D. (2006). Introduction to epityper for quantitative DNA methylation analysis using the MassARRAY® system. *Seq. Appl. Note* 8, 1–8. doi: 10.1016/b978-0-12-420194-1.00001-4
 Feng, L., Peng, F., Li, S., Jiang, L., Sun, H., Ji, A., et al. (2018). Systematic feature selection improves accuracy of methylation-based forensic age estimation in Han Chinese males. *Foren. Sci. Int. Genet.* 35, 38–45. doi: 10.1016/j.fsigen.2018.03.009

- Fondevila, M., Børsting, C., Phillips, C., de la Puente, M. C. E., Carracedo, A., Morling, N., et al. (2017). Forensic SNP genotyping with SNaPshot: technical considerations for the development and optimization of multiplexed SNP assays. *Foren. Sci. Rev.* 29, 57–76.
- Freire-Aradas, A., Phillips, C., Girón-Santamaría, L., Mosquera-Miguel, A., Gómez-Tato, A., Casares de Cal, M. Á, et al. (2018). Tracking age-correlated DNA methylation markers in the young. *Foren. Sci. Int. Genet.* 36, 50–59. doi: 10.1016/j.fsigen.2018.06.011
- Freire-Aradas, A., Phillips, C., and Lareu, M. V. (2017). Forensic individual age estimation with DNA: from initial approaches to methylation tests. *Forens. Sci. Rev.* 29, 121–144.
- Freire-Aradas, A., Phillips, C., Mosquera-Miguel, A., Girón-Santamaría, L., Gómez-Tato, A., Casares De Cal, M., et al. (2016). Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the athena bioscience EpiTYPER system. *Forens. Sci. Int. Genet.* 24, 65–74. doi: 10.1016/j.fsigen.2016.06.005
- Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., et al. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U.S.A.* 89, 1827–1831. doi: 10.1073/pnas.89.5.1827
- Garagnani, P., Bacalini, M. G., Pirazzini, C., Gori, D., Giuliani, C., Mari, D., et al. (2012). Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell* 11, 1132–1134. doi: 10.1111/acel.12005
- Giavarina, D. (2015). Understanding bland altman analysis. *Biochem. Med.* 5, 141–151. doi: 10.11613/bm.2015.015
- Grunau, C., Clark, S., and Rosenthal, A. (2001). Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res.* 29:e65.
- Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sada, S., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* 49, 359–367. doi: 10.1016/j.molcel.2012.10.016
- Hong, S. R., Jung, S. E., Lee, E. H., Shin, K. J., Yang, W. I., and Lee, H. Y. (2017). DNA methylation-based age prediction from saliva: high age predictability by combination of 7 CpG markers. *Foren. Sci. Int. Genet.* 29, 118–125. doi: 10.1016/j.fsigen.2017.04.006
- Hong, S. R., Shin, K. J., Jung, S. E., Lee, E. H., and Lee, H. Y. (2019). Platform-independent models for age prediction using DNA methylation data. *Foren. Sci. Int. Genet.* 38, 39–47. doi: 10.1016/j.fsigen.2018.10.005
- Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol.* 14:R115.
- Horvath, S., and Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* 19, 371–384. doi: 10.1038/s41576-018-0004-3
- Johansson, Å, Enroth, S., and Gyllenstein, U. (2013). Continuous aging of the human DNA methylome throughout the human lifespan. *PLoS One* 8:e67378. doi: 10.1371/journal.pone.0067378
- Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* 13, 484–492. doi: 10.1038/nrg3230
- Jung, S. E., Lim, S. M., Hong, S. R., Lee, E. H., Shin, K. J., and Lee, H. Y. (2019). DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples. *Foren. Sci. Int. Genet.* 38, 1–8. doi: 10.1016/j.fsigen.2018.09.010
- Kint, S., De Spiegelaere, W., De Kesel, J., Vandekerckhove, L., and Van Crielinge, W. (2018). Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. *PLoS One* 13:e0199091. doi: 10.1371/journal.pone.00199091
- Koenker, R., Portnoy, S., Ng, P. T., Zeileis, A., Grosjean, P., Moler, C., et al. (2019). *Quantile Regression, Package 'quantreg'*.
- Lee, H. Y., Jung, S. E., Oh, Y. N., Choi, A., Yang, W. I., and Shin, K. J. (2015). Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study. *Foren. Sci. Int. Genet.* 19, 28–34. doi: 10.1016/j.fsigen.2015.05.014
- Lehmann, U., and Tost, J. (2015). *Pyrosequencing, Methods and Protocols*. Berlin: Springer.
- Lehnert, B. (2015). *Plots (Slightly Extended) Bland-Altman Plots, Package Bland-Altman*.
- Naue, J., Hoefsloot, H. C. J., Mook, O. R. F., Rijlaarsdam-Hoekstra, L., van der Zwalm, M. C. H., Henneman, P., et al. (2017). Chronological age prediction based on DNA methylation: massive parallel sequencing and random forest regression. *Foren. Sci. Int. Genet.* 31, 19–28. doi: 10.1016/j.fsigen.2017.07.015
- Pabinger, S., Ernst, K., Pulverer, W., Kallmeyer, R., Valdes, A. M., Metrustry, S., et al. (2016). Analysis and visualization tool for targeted amplicon bisulfite sequencing on ion torrent sequencers. *PLoS One* 11:e0160227. doi: 10.1371/journal.pone.0160227
- Park, J. L., Kim, J. H., Seo, E., Bae, D. H., Kim, S. Y., Lee, H. C., et al. (2016). Identification and evaluation of age-correlated DNA methylation markers for forensic use. *Forens. Sci. Int. Genet.* 23, 64–70. doi: 10.1016/j.fsigen.2016.03.005
- Richards, R., Patel, J., Stevenson, K., and Harbison, S. A. (2018). Evaluation of massively parallel sequencing for forensic DNA methylation profiling. *Electrophoresis* 39, 2798–2805. doi: 10.1002/elps.201800086
- Smeers, I., Decorte, R., Van de Voorde, W., and Bekaert, B. (2018). Evaluation of three statistical prediction models for forensic age prediction based on DNA methylation. *Foren. Sci. Int. Genet.* 34, 128–133. doi: 10.1016/j.fsigen.2018.02.008
- Vidak, A., Ballard, D., Aliferi, A., Miller, T. H., Barron, L. P., and Syndercombe Court, D. (2017). DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Foren. Sci. Int. Genet.* 28, 225–236. doi: 10.1016/j.fsigen.2017.02.009
- Weidner, C. I., Lin, Q., Koch, C. M., Eisele, L., Beier, F., Ziegler, P., et al. (2014). Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* 15:R24.
- Wickham, H., and Chang, W. (2019). *Create Elegant Data Visualisations Using the Grammar of Graphics, Package ggplot2*.
- Xu, C., Qu, H., Wang, G., Xie, B., Shi, Y., Yang, Y., et al. (2015). A novel strategy for forensic age prediction by DNA methylation and support vector regression model. *Sci. Rep.* 5:17788.
- Zbieć-Piekarska, R., Spólnicka, M., Kupiec, T., Parys-Proszek, A., Makowska, Z., Paleczka, A., et al. (2015). Development of a forensically useful age prediction method based on DNA methylation analysis. *Foren. Sci. Int. Genet.* 17, 173–179. doi: 10.1016/j.fsigen.2015.05.001
- Zubakov, D., Liu, F., Kokmeijer, I., Choi, Y., van Meurs, J. B. J., van IJcken, W. F. J., et al. (2016). Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length. *Foren. Sci. Int. Genet.* 24, 33–43. doi: 10.1016/j.fsigen.2016.05.014

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

Copyright © 2020 Freire-Aradas, Pośpiech, Aliferi, Girón-Santamaría, Mosquera-Miguel, Pisarek, Ambroa-Conde, Phillips, Casares de Cal, Gómez-Tato, Spólnicka, Woźniak, Álvarez-Dios, Ballard, Court, Branicki, Carracedo and Lareu. 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.



Evaluation of the Precision of Ancestry Inferences in South American Admixed Populations

Vania Pereira¹, Roberta Santangelo¹, Claus Børsting¹, Torben Tvedebrink², Ana Paula F. Almeida³, Elizeu F. Carvalho³, Niels Morling¹ and Leonor Gusmão^{3,4*}

¹ Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, ² Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark, ³ DNA Diagnostic Laboratory, State University of Rio de Janeiro, Rio de Janeiro, Brazil, ⁴ Instituto de Investigação e Inovação em Saúde, i3S, Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal

OPEN ACCESS

Edited by:

Kenneth K. Kidd,
Yale University, United States

Reviewed by:

Rick Kittles,
City of Hope National Medical Center,
United States
Francesco Montinaro,
University of Tartu, Estonia

*Correspondence:

Leonor Gusmão
leonorbgsmao@gmail.com

Specialty section:

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

Received: 31 May 2020

Accepted: 31 July 2020

Published: 21 August 2020

Citation:

Pereira V, Santangelo R, Børsting C, Tvedebrink T, Almeida APF, Carvalho EF, Morling N and Gusmão L (2020) Evaluation of the Precision of Ancestry Inferences in South American Admixed Populations. *Front. Genet.* 11:966. doi: 10.3389/fgene.2020.00966

Ancestry informative markers (AIMs) are used in forensic genetics to infer biogeographical ancestry (BGA) of individuals and may also have a prominent role in future police and identification investigations. In the last few years, many studies have been published reporting new AIM sets. These sets include markers (usually around 100 or less) selected with different purposes and different population resolutions. Regardless of the ability of these sets to separate populations from different continents or regions, the uncertainty associated with the estimates provided by these panels and their capacity to accurately report the different ancestral contributions in individuals of admixed populations has rarely been investigated. This issue is addressed in this study by evaluating different AIM sets. Ancestry inference was carried out in admixed South American populations, both at population and individual levels. The results of ancestry inferences using AIM sets with different numbers of markers among admixed reference populations were compared. To evaluate the performance of the different ancestry panels at the individual level, expected and observed estimates among families and their offspring were compared, considering that (1) the apportionment of ancestry in the offspring should be closer to the average ancestry of the parents, and (2) full siblings should present similar ancestry values. The results obtained illustrate the importance of having a good balance/compromise between not only the number of markers and their ability to differentiate ancestral populations, but also a balanced differentiation among reference groups, to obtain more precise values of genetic ancestry. This work also highlights the importance of estimating errors associated with the use of a limited number of markers. We demonstrate that although these errors have a moderate effect at the population level, they may have an important impact at the individual level. Considering that many AIM-sets are being described for inferences at the individual level and not at the population level, e.g., in association studies or the determination of a suspect's BGA, the results of this work point to the need of a more careful evaluation of the uncertainty associated with the ancestry estimates in admixed populations, when small AIM-sets are used.

Keywords: population stratification, ancestry informative marker, Brazil, biogeographical ancestry, population assignment

INTRODUCTION

Patterns of human genetic variation have been thoroughly investigated to unveil past events and disclose historical affinities among populations. Although most of the genetic variation can be observed within populations, a significant fraction can still be used to distinguish human populations, particularly from different continents. For that purpose, markers in a wide range of evolutionary rates and modes of inheritance have been used, showing clear differences between populations from Eurasia, sub-Saharan Africa, East Asia, America, and Oceania, even for small numbers of randomly selected markers.

In the last few years, many sets of Ancestry Informative Markers (AIMs) including SNPs and indels have been described to address individual ancestry or to detect diversity patterns between and within continental populations (Rosenberg et al., 2002; Nassir et al., 2009; Galanter et al., 2012; Pereira et al., 2012; Kidd et al., 2014; Phillips et al., 2014; Moriot et al., 2018; Cheung et al., 2019). To better capture the genetic differences among groups, these AIMs were selected to have large discrepancies in allele frequencies between populations. However, carefully selected markers are required to distinguish close population groups or to characterize continental fringe populations, which are often difficult to distinguish due to gene flow (Bulbul et al., 2016; Li et al., 2016; Yuasa et al., 2018; Pereira et al., 2019; Phillips et al., 2019).

The interest in studying AIMs is growing, and nowadays many DNA testing companies are offering online information on ancestry or genetic history to the average public in a fast and easy way. In forensic genetics, besides tracing back individual genealogies, AIMs can have a prominent role during the investigation phase of missing person cases and in the identification of crime perpetrators. AIMs are also used in clinical genetics, in case/control association studies, to avoid spurious associations due to population substructure (Marchini et al., 2004; Tian et al., 2008; Price et al., 2010).

The same AIM sets developed for human population genetics have also been used to investigate forensic cases. In this context, however, these sets are not usually utilized to question the continental ancestry of a sample contributor, but rather, the most likely population of origin of the DNA profile, i.e., the Biogeographical Ancestry (BGA) of a sample donor (Phillips et al., 2007; Rajeevan et al., 2012, 2020; Tvedebrink et al., 2017, 2018; Mogensen et al., 2020). However, inferring the most likely population of origin of an individual does not always provide direct information about its ancestry profile (and vice versa), namely in recently admixed populations. A set of markers that separates main population groups will not necessarily be the most adequate for determination of the apportionment of ancestry at an individual level, which requires several loci with large allele frequency differences among source populations.

Frequently used metrics proposed for AIM selection rely on the maximization of genetic distances or allele frequency differentials with the minimal number of markers (Pfaffelhuber et al., 2020). However, large genetic distances are usually associated with strong drift and/or selective pressure and, therefore, ancestry inferences or determination of the population

of origin using few markers can be highly influenced by the correct definition of contributing or reference populations. The AIMs in use have always some degree of error associated when performing ancestry assignments, and one of the major challenges has been to select markers that minimize that error rate, increasing the accuracy of the studies or inferences.

In this work, we assessed ancestry with different sets of markers (46 indels developed for capillary electrophoresis and 165 SNPs included in the Precision ID Ancestry panel for massively parallel sequencing). Parent-offspring data from 65 families with mixed parentage were used. Since full-siblings have the same apportionment of common ancestry inherited from their parents, the most informative loci will be those presenting the smallest degree of deviation between the observed and expected ancestry proportions. Data on the genetic profiles of unrelated individuals from the Rio de Janeiro population considering the 210 AIMs are also reported.

We aimed to further investigate the factors that could cause differences in ancestry estimation and their impact when addressing ancestry at individual and population levels. Ultimately, these parameters can be used to understand how to achieve more accurate estimations, namely in populations harboring a trihybrid admixture from European, African, and Native American groups, which is typical for most South American populations.

MATERIALS AND METHODS

Samples, Extraction of DNA and Quantification

Blood samples on FTA cards (Whatman Inc., Clifton, NJ, United States) were collected from 65 Brazilian families (with confirmed kinship) composed by mother, father, and two children (260 individuals in total), as well as from 84 unrelated Brazilian individuals. Informed consent was obtained from all participants included in the study. The project was approved by an ethical committee of the State University of Rio de Janeiro (CAAE: 0067.0.228.000-09).

DNA was extracted with the standard phenol-chloroform method. DNA extract concentration was measured using the InnoQuant HY kit (InnoGenomics) according to the manufacturer's protocol or using the Qubit dsDNA High Sensitivity assay and the Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, United States) following the manufacturer's instructions.

Analysis of Ancestry Markers

The apportionment of the ancestry of each individual was investigated with different sets of AIMs. One set consisted of 46 indels selected to assess European, African, Asian, and Native American ancestries. The indels were amplified by PCR, and analyzed by capillary electrophoresis, according to Pereira et al. (2012). The individuals were also analyzed for 165 SNPs included in the Precision ID Ancestry panel (Thermo Fisher Scientific, Waltham, MA, United States) following the protocol recommended by the manufacturer. The DNA was sequenced

using either the Ion PGMTM or the Ion S5TM platforms (Thermo Fisher Scientific). For the Ion PGMTM, each run contained 25 libraries (50 pM) loaded on an Ion 318TM chip v2 (Thermo Fisher Scientific). For the Ion S5TM, 96 libraries (35–50 pM) were loaded on Ion 530TM chips in each run (Thermo Fisher Scientific).

Data Analysis

Allele calls for the 46 indels were considered for >50 RFUs for heterozygote individuals, and >100 RFUs for homozygote genotypes. For the 165 AIMs, allele calls were carried out following the same criteria as described in Santangelo et al. (2017).

The Precision ID Ancestry panel combines two published assays of 55 (Kidd et al., 2014) and 123 AIMs (Kosoy et al., 2009; Nassir et al., 2009), with 13 overlapping SNPs. Therefore, for ancestry inference analysis, the following five datasets were considered: 46 indels, 55 SNPs, 122 SNPs, 164 SNPs, and 210 markers (46 indels + 164 SNPs). The SNP rs10954737 was not included in the analyses, as it was not typed in all the African (AFR), European (EUR), and Native American (NAM) reference populations (hence, the analysis considered 164 SNPs instead of 165 SNPs).

Reference population data used in the analyses were available for all panels and consisted of 100 AFR, 100 EUR, and 47 NAM individuals. Data for 46 indels were retrieved from the 1000 Genomes database or were previously generated for HGDP-CEPH samples (Pereira et al., 2012). Genotypes for the same individuals for the 164 AIMs were kindly collected and provided by the Kidd Lab from publicly available data.

Allele frequencies, Hardy-Weinberg Equilibrium (HWE), genetic diversities, and pairwise F_{ST} genetic distances were calculated using the Arlequin v3.5.2.2 software (Excoffier and Lischer, 2010). HWE analysis was carried out using 1,000,000 Markov Chain Monte Carlo (MCMC) steps and 1,000,000 dememorization steps. Correction for multiple testing was done according to Bonferroni (1936). Statistical significance among genetic diversities was assessed with the t -test.

Ancestry Inference

The distribution of NAM, EUR, and AFR genetic ancestry in each individual was estimated using the STRUCTURE v.2.3.4.21 software (Pritchard et al., 2000; Falush et al., 2003). The analysis was carried out using a burn-in period of 100,000 iterations, followed by 100,000 repetitions for the MCMC. The “admixture” and the “correlated allele frequencies” models were considered. Population information was used to assist clustering. Three assumed clusters (K) were considered in the analyses, and five independent runs were performed to verify the consistency of the results. The cluster membership coefficients of the five runs were combined using CLUMPP v.1.1.222 (Jakobsson and Rosenberg, 2007).

The apportionment of ancestry in each individual was plotted using the package “*plotrix*” developed for R software (R Core Team, 2013). Statistical significance among average ancestry estimates was assessed with the z -score.

The combined individual ancestry values provided by CLUMPP were used to calculate all the parameters reported in this manuscript (average ancestry levels for the different datasets, absolute differences in ancestry among siblings and parents, and levels of variance reported in each component for all AIM sets).

Population Assignment of Individuals

The assignment of individuals to a population of origin was assessed using the GenoGeographer software (Tvedebrink et al., 2018; Mogensen et al., 2020). In this analysis, the z -score was computed for each individual, considering AFR, EUR, NAM, and Rio de Janeiro as reference populations. The test considers the variance of the allele frequencies in the reference populations (Chakraborty et al., 1993), and the respective p -values are used to assess the most likely population of origin of the profile. The analyses were performed using a leave-one-out approach, excluding the individual tested from the reference dataset.

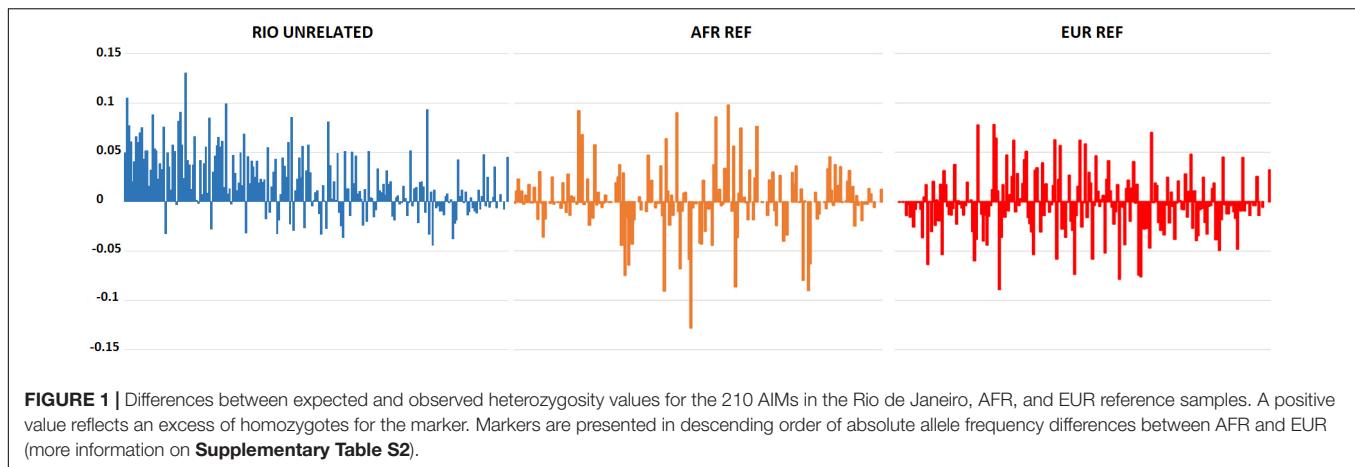
RESULTS

Genetic Profile of the Rio de Janeiro Population for 210 AIMs

Data from 214 unrelated individuals (130 unrelated parents from 65 families, plus 84 additional unrelated individuals), living in Rio de Janeiro (Brazil), were used to calculate population descriptive statistics for 210 ancestry markers (164 SNPs + 46 indels). Three populations were used as reference – AFR, EUR, and NAM. **Supplementary Table S1** contains detailed information on allele frequencies for these markers in Rio de Janeiro compared to the reference populations.

Three loci – rs1800414, rs3811801, and rs671 – were monomorphic in the sample from Rio de Janeiro. This result is in accordance with previous studies showing that these loci are only polymorphic in East Asian populations (Kidd et al., 2014; Pereira et al., 2017; Santangelo et al., 2017). As expected for an admixed population with NAM, AFR, and EUR ancestry, the remaining 207 markers were polymorphic in the Rio de Janeiro dataset. For the reference populations included in this study, the number of monomorphic loci was higher: 34 loci were monomorphic in the AFR reference population, 8 in the EUR sample, and 9 in the NAM group (**Supplementary Table S1**).

Hardy-Weinberg Equilibrium was assessed for the 207 polymorphic markers in the Brazilian population. After correction for multiple tests, only rs6451722 presented a statistically significant deviation (p -value: 0.0002). This deviation was associated with an excess of observed homozygotes (63% compared to 50% expected under HWE), pointing to some degree of population stratification in Rio de Janeiro. Indeed, although statistically non-significant when applying the Bonferroni correction, 72% of the polymorphic loci showed lower observed heterozygosity values than expected in a population in HWE. The excess of homozygotes was higher for loci with greater differences in the allele frequencies between AFR and EUR populations, which are the main contributors to the current population of Rio de Janeiro (**Figure 1**). This general tendency



for an excess of homozygosity was not observed in the AFR and EUR reference population samples.

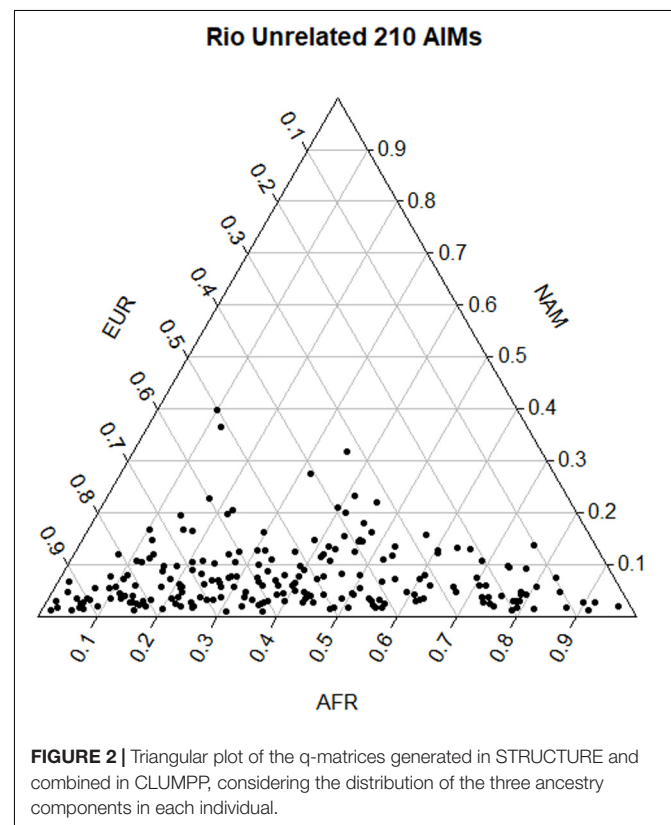
The average genetic diversity was higher in the Brazilian sample (0.376 ± 0.179) than in any of the three continental references (AFR: 0.202 ± 0.098 ; EUR: 0.264 ± 0.127 ; NAM: 0.294 ± 0.142), reflecting the trihybrid origin of the Rio de Janeiro population. Differences in the genetic diversity values between Rio de Janeiro sample and all reference samples were statistically significant (t -test: p -value < 0.016).

Pairwise F_{ST} values among populations showed a smaller differentiation between the Brazilian dataset and the EUR reference ($F_{ST} = 0.113$) than with AFR ($F_{ST} = 0.212$) and NAM ($F_{ST} = 0.314$), which is in accordance with the distribution of ancestry proportions in the Brazilian sample. STRUCTURE results showed that the EUR component was the one with the highest contribution (54.0%), followed by the AFR (38.5%), and the NAM (7.5%) components. The apportionment of ancestry in each individual is plotted in **Figure 2**. The wide dispersal of individuals across the plot (mostly along the AFR and EUR axes) is consistent with a great intrapopulation variation, compatible with recent admixture events and/or a certain degree of population substructure.

A previous study using the same 46 ancestry informative indels as in this work, but in 280 individuals from Rio de Janeiro, reported slightly different ancestry proportions (Manta et al., 2013; **Figure 3B**). Furthermore, the difference between the NAM proportions in both studies (**Figures 3A,B**) was statistically significant (z-score p -value = 0.0271). Since the 46 indels are a subset of the 210 AIMs analyzed here, we recalculated the average ancestry values for our sample of 214 individuals based on the 46 indels alone (**Figure 3C**).

Comparing the results obtained in the two population samples from Rio de Janeiro for the 46 indels (**Figures 3B,C**), higher AFR and NAM contributions were detected in this study. Although differences in these two components were not high enough to be statistically significant (AFR z-score p -value = 0.09102; NAM z-score p -value = 0.4902), a statistically significant decrease of 9.6% (z-score p -value = 0.03486) was found in the EUR component. This variation observed for the same AIM panel could be related to the sampling in both studies.

Manta et al. (2013) investigated randomly selected unrelated individuals born in the metropolitan region of Rio de Janeiro. In the current study, the samples were collected from paternity cases from Rio de Janeiro that also include surrounding areas outside the metropolis. Variation in the ancestry contributions across Rio de Janeiro has been reported by others (Almeida et al., 2017). A previous study that evaluated ancestry inference when using different sampling cohorts from the Rio de Janeiro population reported increased AFR and decreased EUR contributions outside the metropolitan region (Almeida et al., 2017). This sampling effect is also observed in this work.



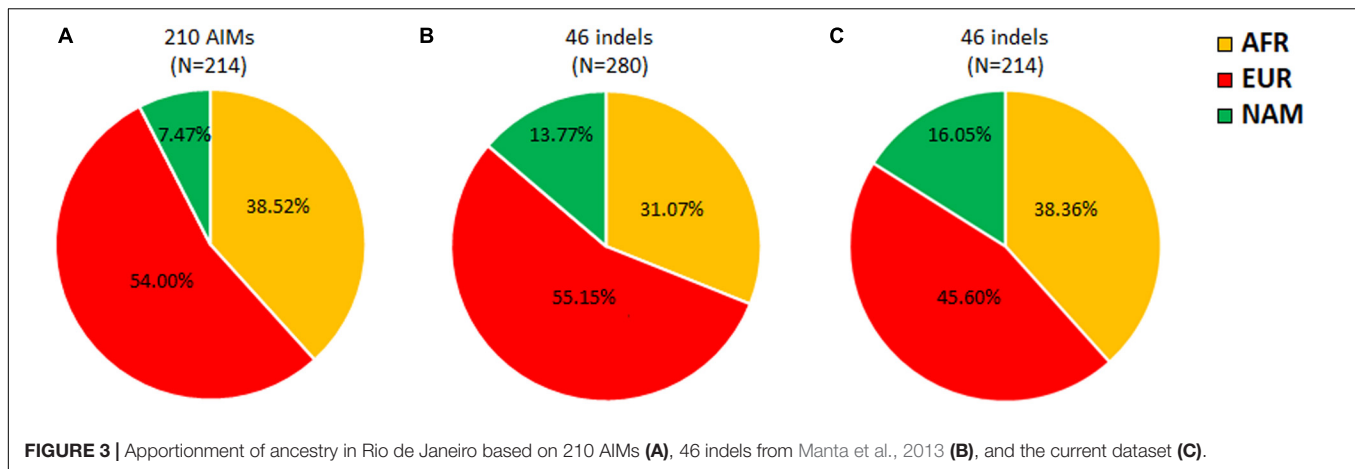


FIGURE 3 | Apportionment of ancestry in Rio de Janeiro based on 210 AIMs (A), 46 indels from Manta et al., 2013 (B), and the current dataset (C).

Differences in the ancestry components were not only observed between the two studies but also when comparing the same individuals analyzed in this work for the 210 and 46 markers (Figures 3A,C). Compared to the complete AIM set, the 46 indels reported increased NAM and decreased EUR ancestry proportions. The difference in the NAM component was statistically significant (z -score p -value = 0.00578).

In the following sections, we intended to investigate the factors that may influence these differences in ancestry estimation and their impact when addressing ancestry at the population and individual levels. Ultimately, we aimed to disclose and compare the effect of the parameters that most influence ancestry determination. This will help to understand how to achieve more accurate estimates, particularly in populations harboring a trihybrid admixture from EUR, AFR, and NAM groups, like the Brazilian population.

Factors Influencing Ancestry Estimations at the Population Level

Although ancestry estimates can be deduced from full genomes or genome-wide studies, the overall ancestry at both population and individual levels is most often calculated based on a certain number of genetic markers showing low discrepancies to the genome-wide results (e.g., Galanter et al., 2012; Santos et al., 2016). Since just a limited portion of the entire genome is analyzed, the accuracy of the results relies on the type and number of selected markers. Loci with low variation among the source populations will tend to give poor ancestry estimates. In these cases, an overestimation of the less represented ancestry components at the expense of those most represented in the population is expected, as seen previously for Rio de Janeiro (Figures 3A,C).

Similarly, even if the markers are highly informative, a balanced discriminatory power between reference populations is also required. As shown in Galanter et al. (2012), a lower mean locus-specific branch length for European ancestry results in an underestimation of this component in MXL and PUR subjects. The same was observed for the AFR ancestry in that study. Besides these factors, the number of markers may also play a

role when addressing ancestry, since a low number of autosomal loci, even if unlinked, may lead to stochastic variations in the representativeness of the different ancestors.

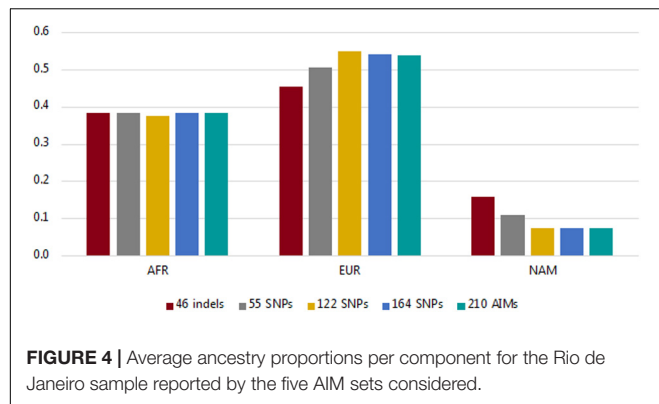
To explore this issue further, we compared the ancestry estimates obtained when using different AIM sets in several American admixed populations.

Ancestry Estimates in Rio de Janeiro Using Different Panels

Average values of ancestry among the unrelated samples from Rio de Janeiro were calculated after dividing the data into several datasets. The Precision ID Ancestry panel combines two ancestry sets: the 55 SNPs selected by the Kidd lab (Kidd et al., 2014), and 123 out of the 128 SNPs selected by the Seldin lab (Kosoy et al., 2009; Nassir et al., 2009). The strategies for marker selection of the panels were slightly different. The 55 SNP panel was made to contain few markers to identify the BGA of an unknown sample. The SNPs are representative of diverse geographical regions, and the selection process included pairwise comparisons of reference populations to select those markers with the largest allele frequency differences. The final set was balanced between population groups so that the geographic regions could be distinguishable with the same level of robustness (Kidd et al., 2014). The strategy for the development of the 128 SNP panel from the Seldin lab was to include markers with large allele frequency differences among European, Sub-Saharan African, American, and East Asian groups (Kosoy et al., 2009; Nassir et al., 2009).

Using the information from the unrelated individuals ($N = 214$), we compared the average ancestry proportions per component reported by the five panels (46 indels, 55 SNPs, 122 SNPs, 164 SNPs, and the total dataset of 210 AIMs) to evaluate the level of variation among them (Figure 4; more information on the average, range, and variance of the ancestry values reported for each panel is presented in Supplementary Table S3).

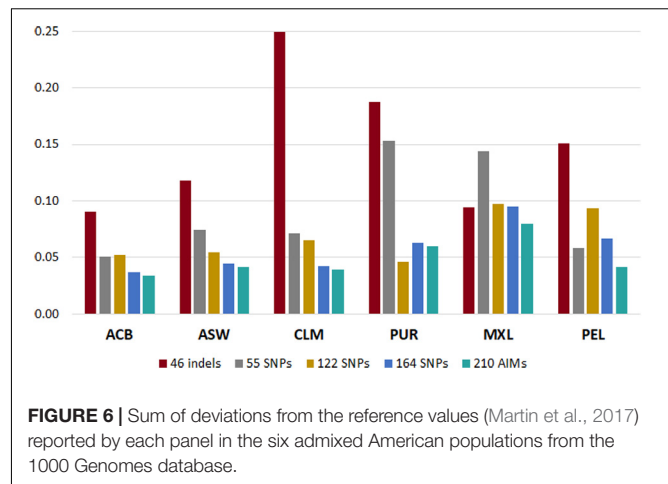
The values for the AFR component were similar for all sets of markers (values varied from 0.3755 for 122 SNPs to 0.3852 for the 210 AIMs). The variation was higher for the EUR and NAM components, which represent the highest and lowest ancestry proportions, respectively (discussed in more detail below).



A previous study that compared ancestry inferences in admixed samples from Brazil and Colombia using 30 ancestry and 30 identity indel-markers showed that the proportions of each component in a trihybrid population always tended to be equally divided for human identity markers that were not optimal for discrimination of ancestry (Aquino et al., 2015). Therefore, when the true ancestry proportions were not captured by the selected markers, for $K = 3$ there was a tendency to overestimate ancestry to values closer to 33%, and vice-versa: values above 33% tended to be underestimated.

In the studied population, the AFR ancestry component is close to 33% (Figure 4). As stated, smaller ancestry differences might not be captured at this level, regardless of the panel used, which may be why no significant difference was observed in the AFR estimates with the different AIM sets.

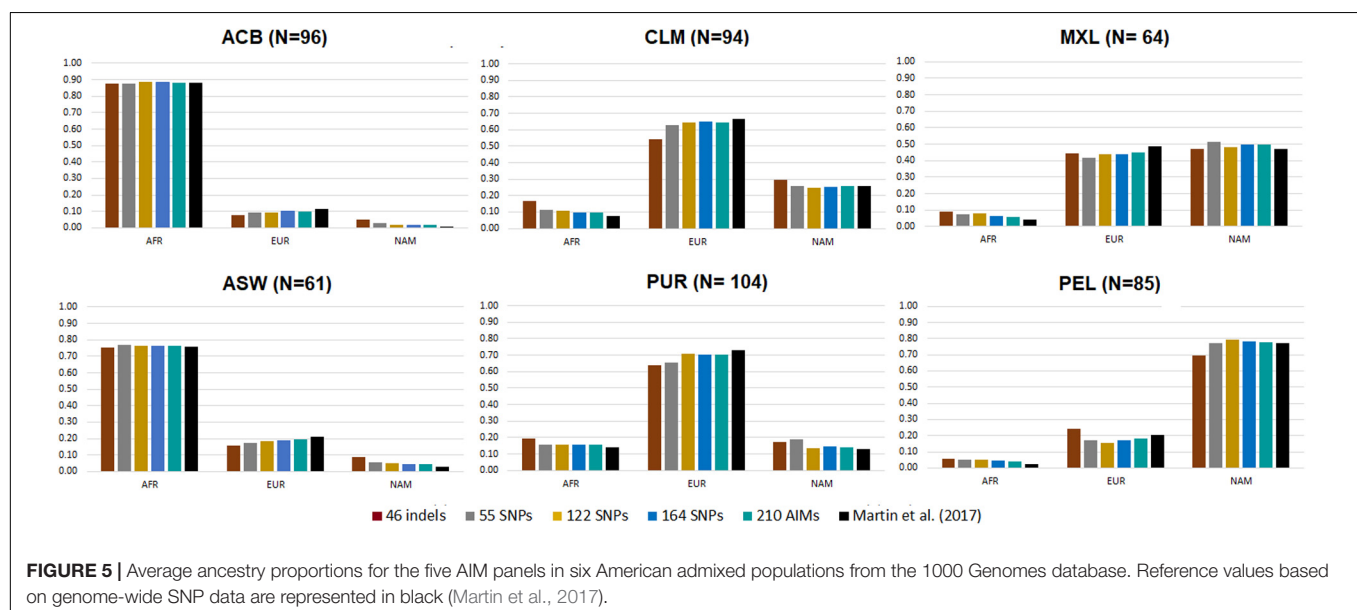
More variation was observed for the EUR and NAM components (Figure 4 and Supplementary Table S3). The EUR component was smaller when the samples were analyzed for the 46 indels, and conversely, this was the panel reporting the highest value of the NAM component (Figure 4 and



Supplementary Table S3). As stated previously, if the panels provide low levels of population differentiation, a tendency to underestimate the major ancestry component (in this case, EUR) and to overestimate the minor component (NAM) would be expected, as seen for the 46 indel panel. Although to a lesser extent, this tendency was also observed in the 55 SNP set.

Ancestry Estimates in American Admixed Populations From the 1000 Genomes

African, European, and Native American ancestry components were estimated for the previously defined AIM sets using data from six American admixed populations included in the 1000 Genomes Project (phase 3): African Caribbean in Barbados (ACB); Americans of African ancestry in Southwest United States (ASW); Colombians from Medellin, Colombia (CLM); Mexican Ancestry from Los Angeles, United States (MXL); Peruvians from Lima, Peru (PEL); and Puerto Ricans from Puerto Rico (PUR). The results for each panel of AIMs were compared to



the ancestry estimates based on common genome-wide SNPs (Martin et al., 2017; **Figure 5**; more information on the average, range, and variance of the ancestry values reported for each panel is presented in **Supplementary Table S4**). The triangular plots of the individual q-matrices generated in STRUCTURE for the six populations based on 210 AIMs are presented in **Supplementary Figure S1**.

The six admixed populations could be divided into four groups: Populations with mainly AFR ancestry (ACB and ASW), populations with mainly EUR ancestry (CLM and PUR), a population with mainly NAM ancestry (PEL), and a population with similar proportions of EUR and NAM ancestries (MXL).

Considering the estimates for genome-wide data as reference, the ancestry values reported by the 46 indels overestimated the minor components and underestimated the major ancestry components for all American populations, except for populations with high AFR ancestry (ACB and ASW), similarly to the observation in the population from Rio de Janeiro (**Figure 4**). For the other AIM panels, there was a small underestimation of the EUR component compared to the values obtained with the genome-wide SNPs (EUR reference values varying from 11.7 to 73.2%; AFR varying from 2.5 to 88%). In contrast, the NAM component was overestimated (NAM reference values between 0.3 and 77.3%).

However, it is worth noting the relatively low variation among all sets in most populations. Most estimates fell within the interval defined by one standard deviation of the average reference values (Martin et al., 2017; **Supplementary Table S4**). Few cases were the exception, namely: the AFR component in MXL, for 46 indels, 55 SNPs, and 122 SNPs; the EUR component in PUR, for the 46 indels; the NAM component in PUR, for 46 indels and 55 SNPs; and the NAM component in ACB, for all sets.

To understand which panel presented more variation compared to the reference values based on the genome-wide SNPs (Martin et al., 2017), we calculated the sum of all the absolute deviations from the reference values for the five panels in each population (**Figure 6**). Different trends could be seen for each panel, depending on the ancestry profile of the population.

In all populations, the 210 AIM set had the smallest accumulated error for the three continental components. The 46 indels performed worst in most populations, but it presented smaller deviations than the 55, 122, and 164 SNP panels in the MXL population. This population had similar proportions of EUR and NAM ancestries (**Figure 5**). For the populations with lowest NAM ancestry, the combination of 46 indels and 164 SNPs did not substantially improve the accuracy of the estimates compared to the 164 SNP panel alone.

However, for populations with high proportions of NAM ancestry (MXL – 40.6% and PEL – 77.6%), the inclusion of the 46 indels improved the estimates obtained with the 164 SNPs of the Precision ID Ancestry panel.

The type of errors seen for the 46 indels can be explained by the low number of markers and/or low F_{ST} values among the three populations. As for the remaining panels, the systematic biases were more likely due to an unbalanced genetic differentiation among populations, with EUR-NAM showing the lowest F_{ST} value (discussed in more detail below).

Number of Markers and the Genetic Differentiation of the Reference Populations

From the results obtained for the different AIM-sets, it can be seen that the number of loci and their capacity to differentiate source populations influence the accuracy of the ancestry estimations. With a higher number of loci, the variations associated with the estimations were smaller, as seen for example in the inferences provided by the 122, 164, and 210 AIM panels (**Figures 4, 5**). Apart from the variation in the number of loci, the five panels presented different pairwise F_{ST} values among the three reference populations (**Figure 7A**).

This leads to the question of whether results of ancestry inferences are more dependent on the number of markers included in an AIMs panel than the combined population differentiation these markers provide, or if they are equally dependent on both?

To address this issue, we returned to the global set of 210 AIMs, and defined three additional AIM panels based on different selection strategies (more details on **Supplementary Information 1**):

(a) two new panels with 46 and 55 AIMs (named 46 panel B and 55 panel B), where we aimed to maintain the same number of markers but selected those that would have the highest and most balanced pairwise F_{ST} s among all population groups. The distances among EUR-NAM were given preference since they had the smallest distances in the original panels;

(b) a new panel with a small number of markers (40 AIMs), but the emphasis was now on the selection of the combination of markers that produced smaller differences between the F_{ST} s among the reference groups (i.e., same levels of differentiation between AFR-EUR, AFR-NAM, and EUR-NAM).

Figure 7 presents the pairwise F_{ST} s among reference populations obtained with each panel (**Figure 7A**), the average ancestry proportions reported for the 214 unrelated individuals (**Figure 7C**), and their absolute differences compared to the estimates provided by the 210 AIMs (**Figure 7B**).

Looking at the pairwise F_{ST} values (**Figure 7A**), we observed that the number of markers is not the only factor responsible for the differences previously reported in the ancestry estimations (**Figure 7C**). Panels with the same number of markers presented different magnitudes of F_{ST} . Compared to the 46 indels, the 46 panel B had greater F_{ST} s among the three population groups, and they were similar to the F_{ST} s for larger panels. In **Figure 7B**, the new 46 panel B has much smaller deviations from the ancestry values obtained with the total set of 210 AIMs, and it appears to perform better than the 55 SNP panel, which has less balanced pairwise F_{ST} s (**Figure 7A**).

For the two panels with 55 markers, smaller F_{ST} values were obtained for AFR-EUR and AFR-NAM. For EUR-NAM, which was the genetic distance that was prioritized upon selection of these markers, the F_{ST} was slightly higher. The 55 panel B also showed smaller differences compared to the 55 SNP set (**Figure 7B**), probably due to more balanced pairwise F_{ST} s among the source populations.

Compared to the 46 indel and 55 SNP sets, the average ancestry values obtained with the 40 AIMs were overall closer to those reported for the 210 panel (**Figure 7C**); differences

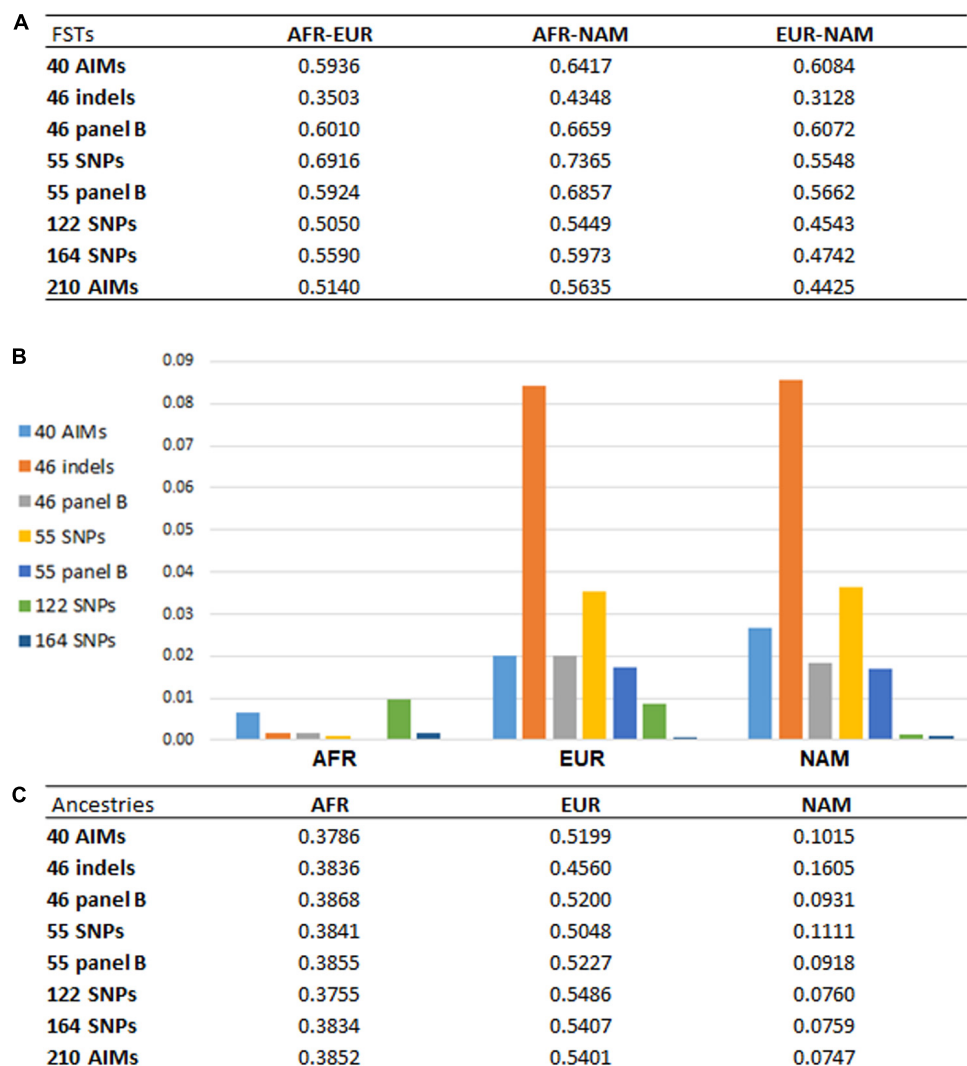


FIGURE 7 | (A) pairwise F_{ST} values among reference populations (AFR, $N = 100$; EUR, $N = 100$; NAM, $N = 47$) based on the different AIM panels; **(B)** absolute values of the differences between the average ancestries reported for each panel compared to the 210 AIMs; **(C)** average ancestry values per component and per panel for unrelated individuals.

ranged from 0.0077 in the AFR component to 0.0277 in the NAM component. As expected from the selection criteria, the F_{ST} values based on the 40 AIMs were higher than those obtained for other less balanced panels, or panels with a higher number of markers (**Figure 7A**). However, when the number of markers included in the panel increases to 122 or 164, the ancestry estimates were closer to those obtained for the full set, with no significant variation observed between 164 AIMs and the total set of 210 AIMs. A similar trend was observed when comparing the performance of the three newly selected sets in the six American populations from the 1000 Genomes project (**Supplementary Figure S2**). However, the errors associated to each panel showed a variation that depends on the ancestry profile of the populations.

The results highlight that a balanced population differentiation among the reference groups also plays an important role in the accuracy of the ancestry estimations,

especially for small sets of 40–55 SNPs. Large AIM sets (e.g., the 164 AIMs), result in smaller variation in the ancestry estimates even if these panels had slightly lower and less-balanced F_{ST} s.

Factors Influencing Ancestry Estimations at the Individual Level

As illustrated above, differences in ancestry estimates are expected when using different groups of AIMs. These differences can be due to the poor performance of the markers to differentiate ancestry components. In this case, there will be a directional bias in the estimations, and some ancestry components will tend to be overestimated at both the individual and population level. However, if the differences are not related to the marker performance, but with the (low) number of markers used, it is expected that the differences in population genetic statistics will

be random. These variations will have a much smaller effect in larger population samples than at the individual level.

Individual Ancestry Estimates Using Different Panels

To investigate differences in ancestry estimates at the individual level, we plotted the pairwise comparisons between the panels of 46 indels, 55 SNPs, and 122 SNPs.

As can be seen in **Figure 8**, there are large differences at the individual level when the results of the three panels are compared. The results are in accordance with the ancestry estimates at the population level, with the NAM component showing the worst results. The highest levels of correlation and agreement among comparisons were found between the 55 SNP and 122 SNP panels for the AFR and EUR components ($r = 0.942$ and $r = 0.931$, respectively). The correlation was lower for the NAM component in all pairwise comparisons ($r \leq 0.585$).

The results for the six American admixed populations of the 1000 Genomes Project showed a similar trend, with the 55 SNP and 122 SNP panels presenting the highest levels of correlation and agreement among comparisons. However, the component with the largest differences varied among populations. The most extreme disagreement among the three panels was obtained for the NAM component in ACB and for the AFR component in MXL (**Supplementary Figures S3–S8**).

Comparison of Average Ancestry Proportions in Parents vs. Offspring

Families with parents and their offsprings are excellent proxies to study the variation of ancestry estimations at the individual level. To this end, data from 65 families (mother, father, and two offsprings) with confirmed kinship were investigated ($N = 260$). The estimates reported by the five panels (46 indels, 55 SNPs, 122 SNPs, 164 SNPs, and 210 AIMs) were compared once again considering that: (1) the apportionment of ancestry in the offspring should be close to the average ancestry of the parents, and (2) full siblings should present very similar ancestry values for a set of unlinked markers.

In this context, the most informative group of loci will be the one presenting the smallest difference in ancestry between the siblings and their parents.

We looked at the variation in the average ancestry proportions provided by STRUCTURE for the datasets defined by the parents (mothers and fathers, M + F) and the offspring (O1, O2, and O1 + O2). The average values and their absolute differences are presented in **Figure 9**.

In theory, the average ancestry proportions in these four groups should be similar, but differences were observed between the estimates. The largest difference between datasets was 2% for the EUR component estimated by the 122 AIMs, and for the NAM, when using 46 indels. The variation in the ancestry proportions was observed regardless of the number of markers included in the panel. The set of 55 SNPs had the lowest variation (all values were below 0.078%).

The analyses were based on a limited number of loci and a random variation was expected that depended on the numbers of markers and samples analyzed. However, there was no clear correlation between the number of markers and the differences in

the variation observed between parents and offspring subsamples. We can, therefore, conclude that if there is a drift effect at the individual level, this is not reflected at the population level for the number of samples analyzed here. A directional bias could also influence the differences observed within each panel. For instance, an approximation of ancestry components to a certain value will result in a smaller difference among individuals.

To investigate the expected variation of ancestry estimates at the individual level, we compared the average ancestry of the parents and offsprings for each component (**Figure 10**).

A high positive correlation was obtained for all AIM sets, and the values were closer to $r = 1$ when the number of markers was increased. For the AFR component, the highest correlation was observed for the 164 SNPs ($r = 0.989$). For the EUR component, the 164 SNPs and total AIM set of 210 markers presented the highest values ($r = 0.984$). For the NAM component, the highest value ($r = 0.927$) was reported for the 122 SNPs.

To evaluate the agreement between the observed ancestries for each offspring and the expected values given by the average ancestry of the parents, we calculated the absolute differences of these values, shown in **Figure 11**. The differences decreased when a higher number of markers was used.

Although the differences between the average ancestry estimated in parents vs. offspring were lowest for the 55 SNPs at the population level (**Figure 9B**), the full set of 210 AIMs produced the smallest variation at the individual level. Moreover, the addition of the 46 Indels to the 164 SNPs had the highest effect in the NAM component, which is in accordance with what was observed for the estimates obtained at the population level (**Figure 5** – Section “Ancestry Estimates in American Admixed Populations From the 1000 Genomes”).

Comparison of Ancestry Estimates Among Sibling Pairs

A further comparison was performed between siblings (**Figure 12**) based on the assumption that siblings should have identical ancestry components from the three continental sources when accessed by a large enough number of well balanced AIMs.

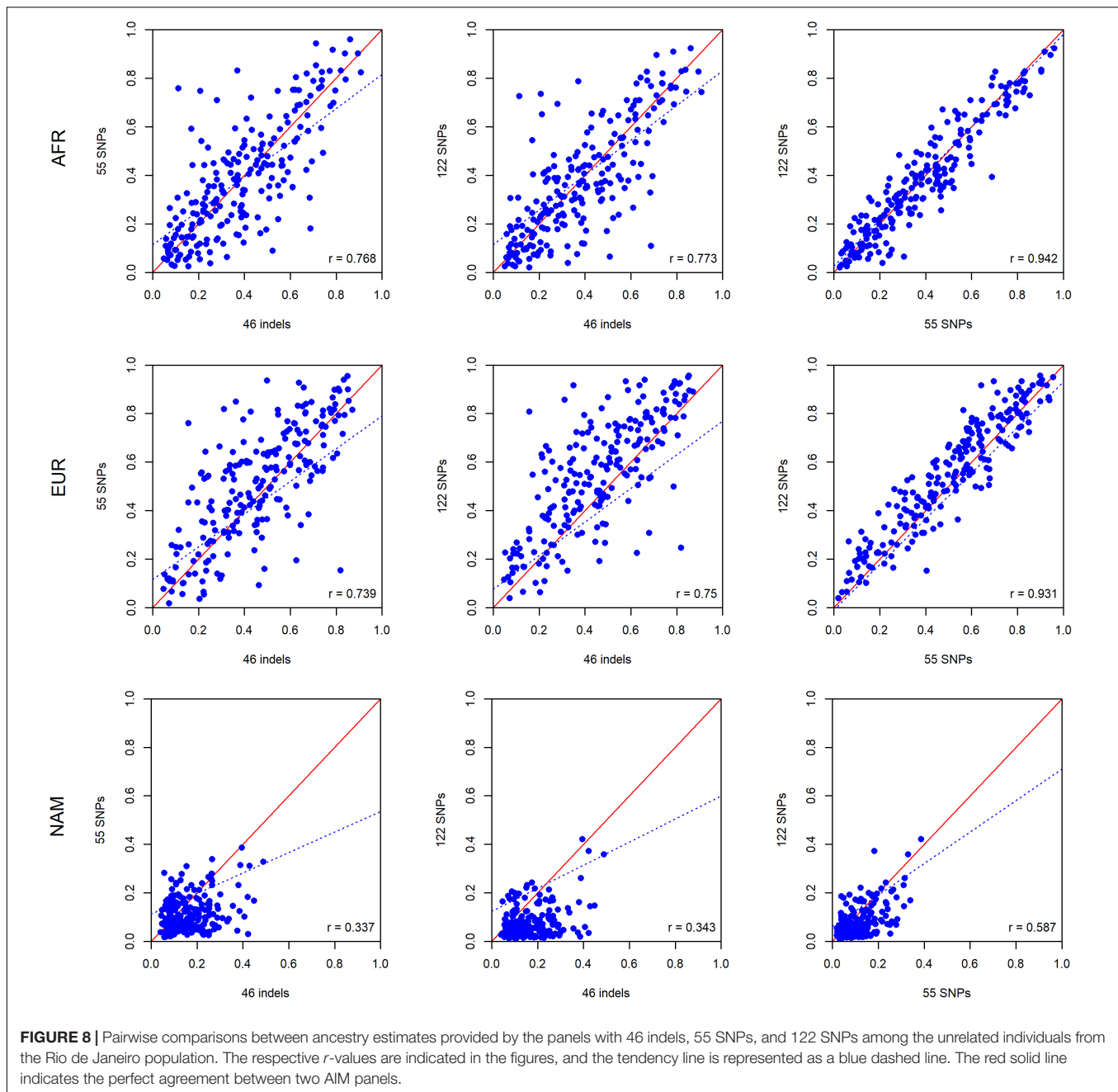
As for the comparisons between parents and offsprings, a high correlation was also observed between the ancestry proportions of siblings. The smallest r -value was 0.531 for the 55 SNPs in NAM; all other correlations were above 0.761 (**Figure 12**). There was an overall tendency of increased correlation with an increase in the number of markers.

The concordance between the ancestries of the siblings was measured by calculating the absolute differences observed (**Figure 13**). Again, smaller differences among siblings were obtained with increasing numbers of markers. The 210 AIM panel had the smallest deviation in ancestry estimations among siblings (**Figure 13**).

For all AIM sets, both correlation and concordance were higher between parents vs. offspring than between siblings.

Inferences on Biogeographical Ancestry (BGA)

The five previously defined AIM sets were used for prediction of the biogeographical origin of the profiles from Rio de Janeiro,



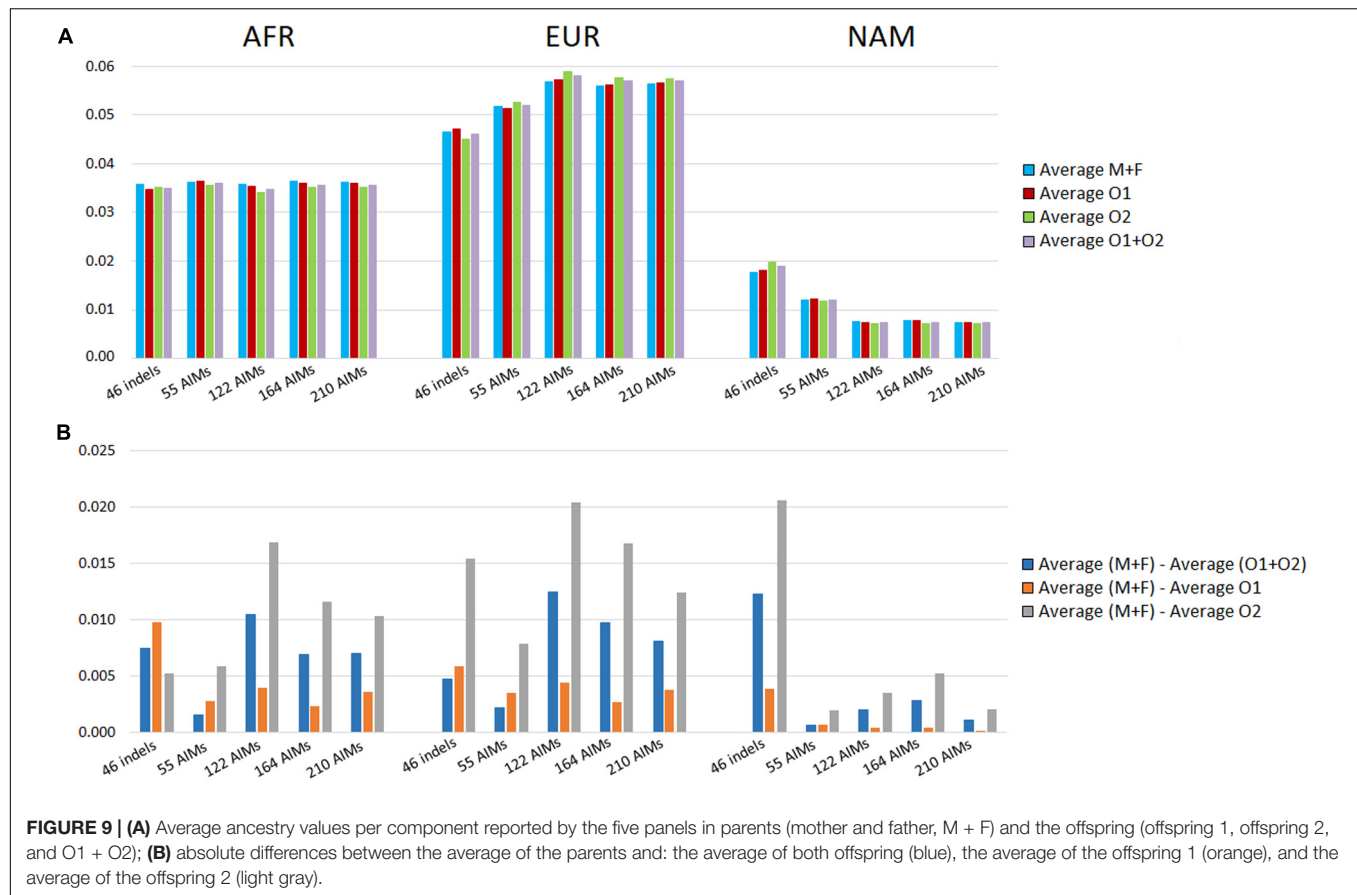
considering four reference populations: AFR, EUR, NAM, and Rio de Janeiro. A z -score test was applied to the 214 unrelated individuals and to each offspring of the 65 sibling pairs from Rio de Janeiro, to assess whether one (or more) of the four reference populations was accepted as a potential population of origin of each AIM profile. This test was performed using the approach described in Tvedebrink et al. (2018).

Biogeographical Ancestry Inferences in Rio de Janeiro Using Different Panels

The accuracy of BGA inferences for the five AIM sets was estimated considering AFR, EUR, NAM, and Rio de Janeiro as

the potential source populations. To this end, for all AIM sets, we evaluated the proportion of individuals that were classified as “Rejected” (none of the four reference populations was defined as a possible population of origin of the profile; z -score > 1.64 , p -value < 0.05) or “Accepted” (at least one of the four reference populations was defined as a possible population of origin of the profile; z -score ≤ 1.64 , p -value ≥ 0.05).

Among the cases defined as “Accepted,” it was also calculated (1) the proportion of “concordant” assignments (individuals accepted in the true population of origin or, when accepted in more than one population, a significant higher likelihood was obtained for the true population of origin), (2) “discordant”



(individuals accepted in a population that was not the true population of origin, or accepted in the true population but with a significantly lower likelihood than in another), and (3) “ambiguous” (individuals accepted in more than one population with non-significantly different likelihoods).

The results in **Table 1** show that there is a relatively high rate of rejection, depending on the population and the panel considered. The highest values were found for the Rio de Janeiro samples. In this population, the percentage of samples rejected increased for larger panels, reaching 31% for the 210 AIMs. Except for the AFR, the 46 Indels showed discordant results that reach 21% in Rio de Janeiro. Although with high percentage of rejection, larger panels show higher percentage of concordant profiles. However, even for the AIM sets with high concordance, there is still 9% of individuals being assigned to the wrong population.

The final proportion of all cases that were accepted in the true population with significant higher likelihood was only 63% for the largest panel (135 individuals out of the 214). Taking together both sensitivity and specificity (concordant results), the 55 SNPs presented the highest rate of assignment of individuals in the true population of origin (71%).

The discordant assignments were mainly due to the low discrimination between EUR and Rio de Janeiro. Comparing the z-scores obtained with the different panels (**Supplementary Figure S9**) it is possible to see an overlap of the z-scores for the

EUR and Rio de Janeiro samples when considering EUR as the population of origin.

Comparison of Biogeographical Ancestry Estimates Between Sibling Pairs

To compare the results among the two siblings, we investigated (1) how many sibling pairs had both siblings accepted or both rejected in the true population of origin, (2) how many had one sibling accepted into a reference population and the other sibling was rejected, and (3) how many sibling pairs had both offspring rejected or both accepted as belonging to any of the three reference populations considered rather than the true one (**Supplementary Table S5**).

For the 65 sibling pairs, the number of pairs rejected in all reference populations varied from one for the 122 SNPs to six for the 210 AIMs. The number of sibling pairs where one individual was accepted and one was rejected as belonging to any of the tested populations varied from seven (for the 46 indels) to 15 cases (for the 122 SNPs) (see **Supplementary Figure S10**). The 46 indels showed the highest sensitivity (percentage of sibling pairs that were not rejected in the true population), and the lowest sensitivity was obtained for the 210 AIMs (**Supplementary Table S5** and **Supplementary Figures S9, S10**). **Supplementary Table S5** presents the percentage of cases where individuals were accepted in their true population of origin and rejected in other reference populations, which indicates the specificity of each

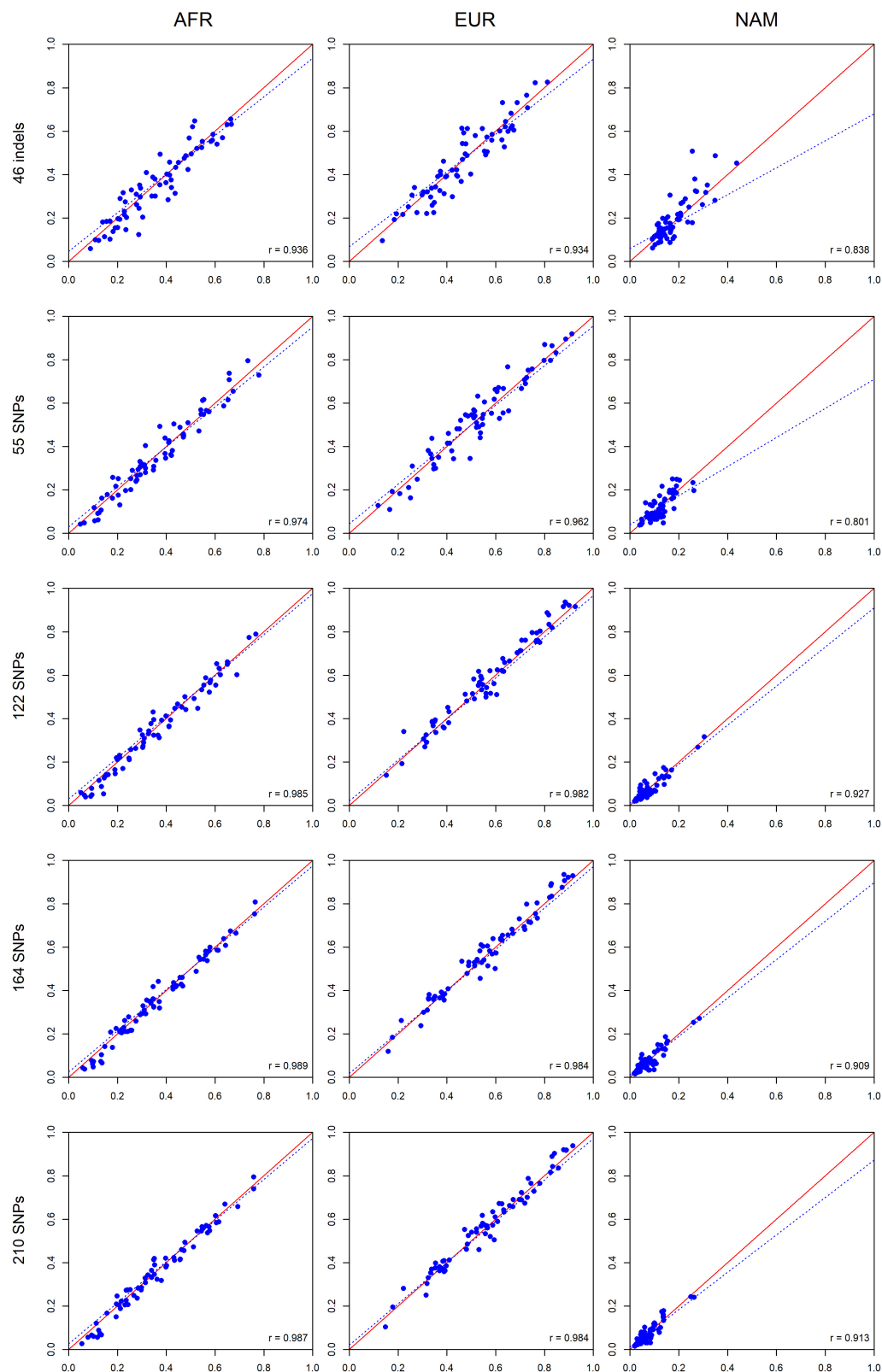
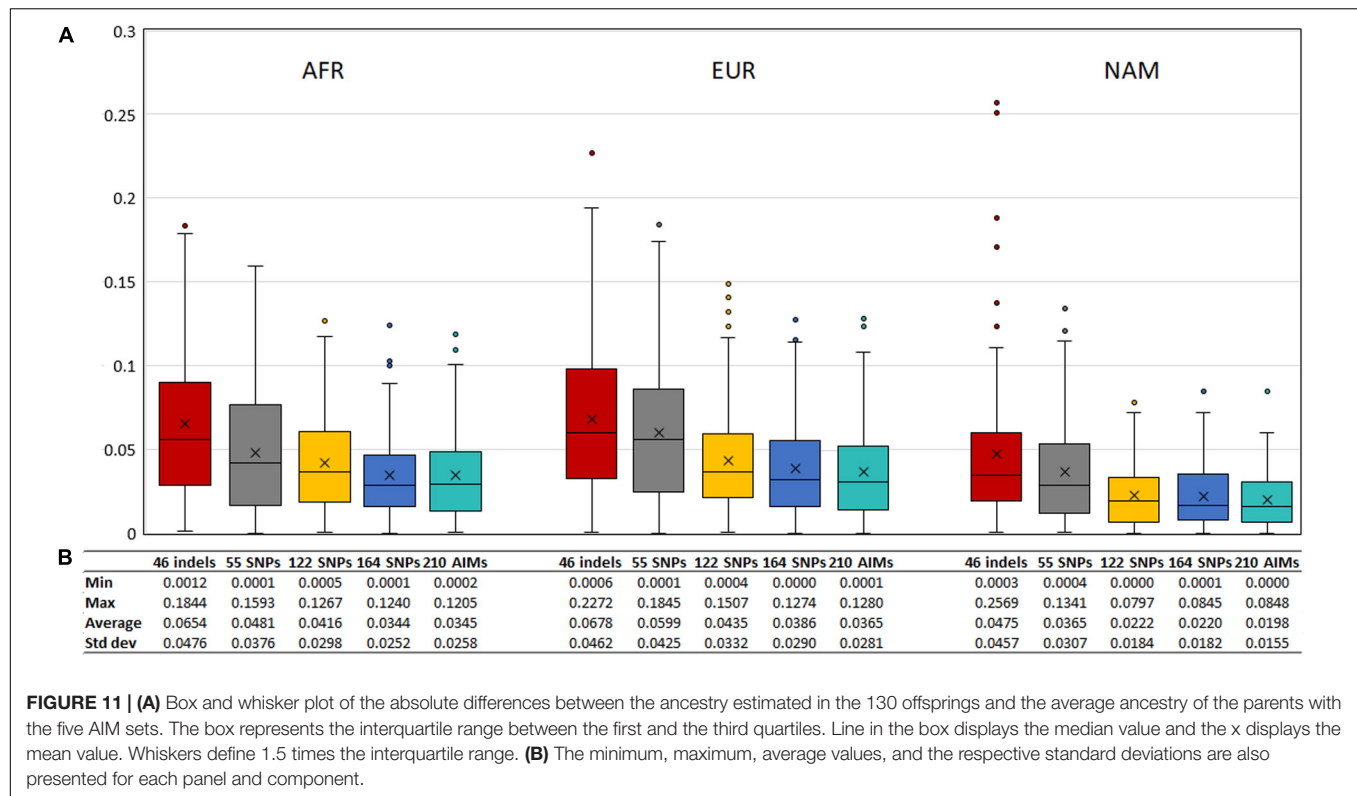


FIGURE 10 | Pairwise comparisons between the average ancestry estimates of the parents and the average estimates of their offsprings based on five AIM sets. The respective r -values are indicated in the figures, and the tendency line is represented as a blue dashed line. The red solid line indicates the perfect agreement between two AIM panels.



panel. In all cases, profiles were rejected as belonging to the NAM reference population. In two cases, one of the two siblings was also accepted in the AFR population. The highest proportion of ambiguities corresponded to profiles where individuals were both accepted in EUR and the Rio de Janeiro populations. The panel showing the lowest specificity was the 46 indels, and the 55 SNP panel was the one with the highest specificity (Supplementary Table S5). Taking together both sensitivity and specificity, the best results were obtained for the 55 SNPs, with 79.23% of acceptance in Rio de Janeiro and exclusion from other populations. The frequency of rejection of the true population plus ambiguous assignment was the highest for the 46 indels (44.62%) and varied from approximately 20 to 30% for the remaining panels.

The results obtained for the sibling pairs with different acceptance output in the true population (one accepted and one rejected), for at least one marker set, are described in Supplementary Figure S11. The highest agreement between siblings (both rejected or accepted) was obtained for the 46 indels. The number of sibling pairs with a different outcome was 13 for the 55 SNPs, 164 SNPs, and 210 AIMS, increasing to 15 for 122 SNPs. Except in one case (F87), the acceptance/rejection result varied among panels.

The z-scores calculated for the 65 sibling pairs considering in the four reference populations (Figure 14) showed that despite their low sensibility and specificity, larger panels resulted in higher rejection values when considering AFR, EUR, and NAM as possible populations of origin. A good agreement can also be seen in the z-scores between siblings.

DISCUSSION

Genetic Profile of the Rio de Janeiro Population

Several studies have pointed to a high variation in the genetic background of Brazilian populations, that present different proportions of EUR, AFR, and NAM admixture. This characteristic is shared by most populations in South American countries (Salzano and Sans, 2014; Homburger et al., 2015; Chacón-Duque et al., 2018). The results from the analysis of 210 AIMS in the 214 unrelated individuals from Rio de Janeiro indicated that the population was predominantly European (54.0%) with admixture of African (38.5%) and Native American genetic heritage (7.5%), which was in accordance with the expectations based on previous studies on Brazilian populations (e.g., Pena et al., 2011; Manta et al., 2013; Salzano and Sans, 2014; Moura et al., 2015).

A comparison with another sample from Rio de Janeiro (Manta et al., 2013) showed differences in ancestry estimates. These differences can be explained by different sampling strategies in association with population stratification. Locus by locus analysis did not reveal statistically significant deviations to the HWE, except for one locus. Nevertheless, an overall excess of homozygotes was observed, particularly for loci showing large differences in allele frequency between the two main source populations (AFR and EUR). This excess of homozygosity is also supportive of population stratification in Rio de Janeiro.

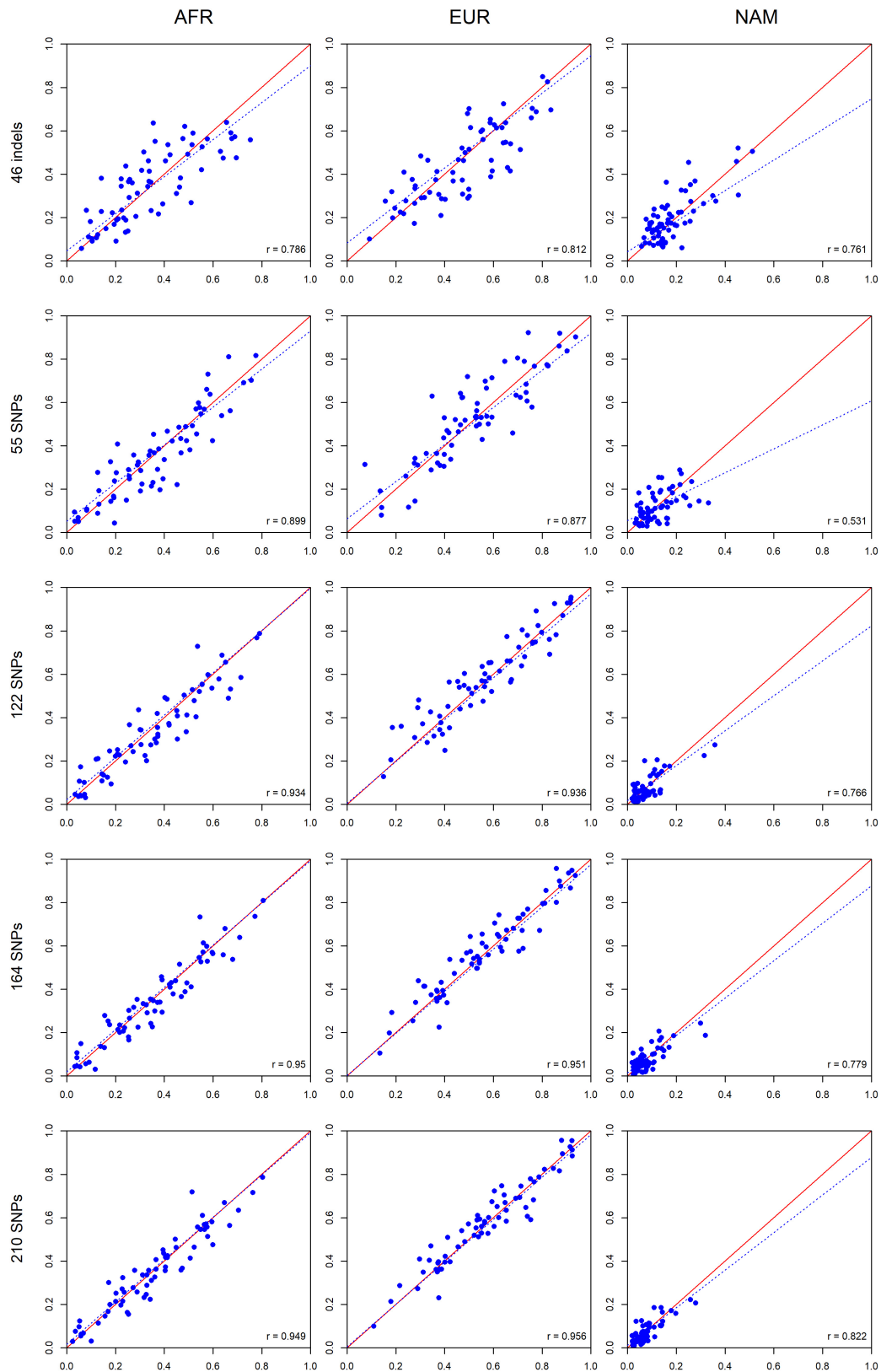


FIGURE 12 | Comparisons between the ancestry estimates for siblings pairs based on the five AIM panel. The respective r -values are indicated in the figures, and the tendency line is represented as a blue dashed line. The red solid line indicates the perfect agreement between two AIM panels.

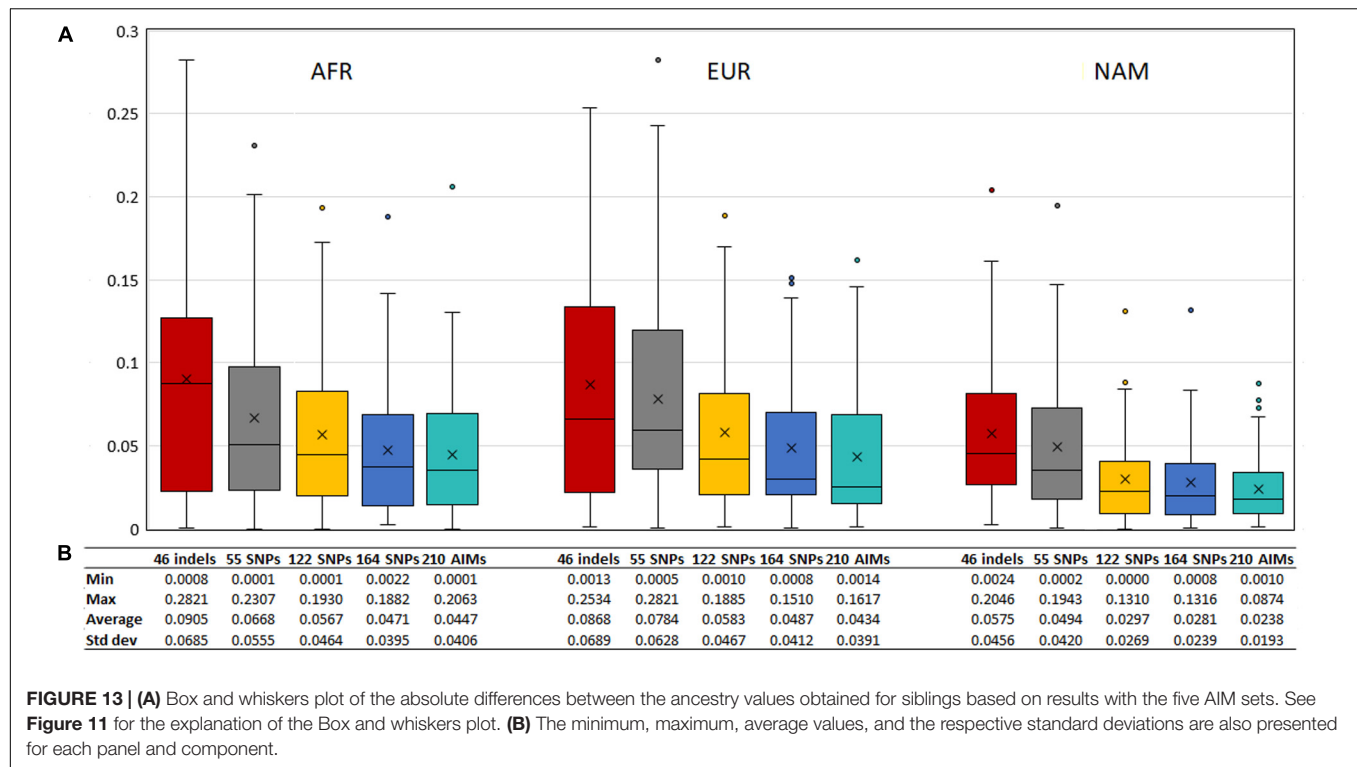


TABLE 1 | Results of BGA inferences for five AIM panels, considering AFR, EUR, NAM, and Rio de Janeiro populations.

	Panel	Accepted			Rejected
		Ambiguous	Concordant	Discordant	
AFR	46 indels	0 (0.00%)	93 (100.00%)	0 (0.00%)	7 (7.00%)
	55 SNPs	0 (0.00%)	89 (100.00%)	0 (0.00%)	11 (11.00%)
	122 SNPs	0 (0.00%)	92 (100.00%)	0 (0.00%)	8 (8.00%)
	164 SNPs	0 (0.00%)	91 (100.00%)	0 (0.00%)	9 (9.00%)
	210 AIMs	0 (0.00%)	91 (100.00%)	0 (0.00%)	9 (9.00%)
EUR	46 indels	2 (2.02%)	94 (94.95%)	3 (3.03%)	1 (1.00%)
	55 SNPs	0 (0.00%)	90 (100.00%)	0 (0.00%)	10 (10.00%)
	122 SNPs	0 (0.00%)	97 (100.00%)	0 (0.00%)	3 (3.00%)
	164 SNPs	0 (0.00%)	96 (100.00%)	0 (0.00%)	4 (4.00%)
	210 AIMs	0 (0.00%)	97 (100.00%)	0 (0.00%)	3 (3.00%)
NAM	46 indels	0 (0.00%)	41 (97.62%)	1 (2.38%)	5 (10.64%)
	55 SNPs	0 (0.00%)	39 (100.00%)	0 (0.00%)	8 (17.02%)
	122 SNPs	0 (0.00%)	36 (100.00%)	0 (0.00%)	11 (23.40%)
	164 SNPs	0 (0.00%)	34 (100.00%)	0 (0.00%)	13 (27.66%)
	210 AIMs	0 (0.00%)	33 (100.00%)	0 (0.00%)	14 (29.79%)
Rio de Janeiro	46 indels	8 (4.60%)	129 (74.14%)	37 (21.26%)	40 (18.69%)
	55 SNPs	0 (0.00%)	151 (90.96%)	15 (9.04%)	48 (22.43%)
	122 SNPs	1 (0.63%)	134 (84.81%)	23 (14.56%)	56 (26.17%)
	164 SNPs	0 (0.00%)	137 (91.95%)	12 (8.05%)	65 (30.37%)
	210 AIMs	0 (0.00%)	135 (91.22%)	13 (8.78%)	66 (30.84%)

"Rejected" – none of the four reference populations was defined as a possible population of origin of the profile; z -score > 1.64 , p -value < 0.05 ; "Accepted" – at least one of the four reference populations was defined as a possible population of origin of the profile; z -score ≤ 1.64 , p -value ≥ 0.05 ; "Ambiguous" – accepted in more than one population with non-significantly different likelihoods; "Concordant" – accepted in the true population of origin or, when accepted in more than one population, a significant higher likelihood was obtained for the true population of origin; "Discordant" – accepted in a population that was not the true population of origin, or accepted in the true population but with a significantly lower likelihood than in another.

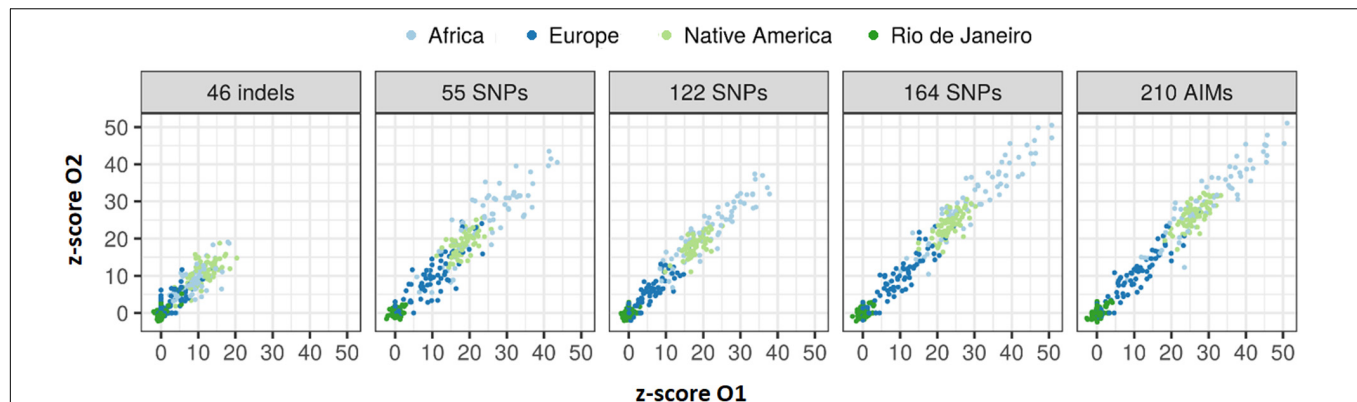


FIGURE 14 | Values of z-score for the 65 sibling pairs when tested against Rio de Janeiro, AFR, EUR, and NAM population samples.

In forensic genetics, it is important to consider the population stratification in the definition of allele frequency databases, and sub-structuring levels are also relevant for adjusting match probabilities (Curran et al., 2002; Buckleton et al., 2016; Hessab et al., 2018). In contrast to many North American populations, there are admixture gradients within populations in South America, which makes it difficult to define ethnic subgroups except for some Native and Afro-descendant communities that have maintained a certain degree of cultural identity and geographical isolation.

The 210 AIMs are essentially biallelic and a smaller number of individuals is usually necessary for accurate allele frequency estimations compared to multiallelic STRs. However, large sample sizes are required to detect HWE deviations and linkage disequilibrium that are more likely to occur in recently admixture and/or stratified populations (Kling et al., 2015).

No deviations from HWE have so far been reported for the commonly used STRs in admixed Brazilian populations (Rodrigues et al., 2007; de Assis Poiars et al., 2010; Alves et al., 2014; Hessab et al., 2015; Moyses et al., 2017). This may be attributed to the relatively small sample sizes since small deviations can only be detected in large samples. Furthermore, STRs selected for forensic identification have a high intrapopulation diversity and low intercontinental variability. Therefore, they are less efficient for the detection of HWE deviations in admixed populations than AIMs. In most forensic genetic publications, the authors employ Bonferroni adjustments whenever HWE p -values surpass the predefined significance level (usually 5%). However, no further consideration concerning the result itself or the sample size is usually made, which neglects the possibility of population stratification (Ye et al., 2020).

Lineage markers may also be useful to detect intrapopulation substructure since they present strong geographical differentiation. The presence of gametic associations between autosomal, mtDNA, and Y-chromosomal markers can be due to recent admixture and population stratification (Vullo et al., 2015). A study carried out in the Brazilian population of Rio de Janeiro showed a gametic association between autosomal AIMs and mtDNA haplogroups. This association between unlinked markers supports our hypotheses regarding the

presence of population substructure in Rio de Janeiro (Simão et al., 2018). In summary, the results obtained in this study highlight the importance of having large sample sizes to investigate population substructure in admixed populations. Although statistically significant deviations to HWE could only be detected for a single marker when applying Bonferroni correction, the results indicated the need of studying a larger sample from Rio de Janeiro to investigate an overall excess of observed homozygosity.

Ancestry Estimations in South American Admixed Populations

In the last few years, many studies have been published reporting new AIM sets to determine the proportion of intercontinental individual admixture and to infer BGA. Selected sets of different types of AIMs have been proposed based on their ability to determine population clustering patterns (Soundararajan et al., 2016; Kidd et al., 2017). In most cases, these panels were based on their ability to correctly assign the origin of individuals from African, Eastern Asian, European, Oceanian, and Native American populations. Less often, a higher resolution was pursued within one of these five groups (e.g., Li et al., 2016; Bulbul et al., 2018; Jung et al., 2019; Verdugo et al., 2020). Regardless of the ability of these sets to separate populations from different continents or geographic regions, the uncertainty associated with the estimates provided by these panels and their capacity to accurately report the different ancestral contributions in individuals of admixed populations has rarely been investigated.

This work aimed to compare the results of different groups of AIMs currently in use in the forensic field and their ability to determine the admixture proportions of a population, the profile of an individual's ancestry, and the assignment of its population of origin in admixed populations from South America.

At the population level, all AIM sets reported similar population profiles in terms of the relative proportions of AFR, EUR, and NAM components in the seven admixed American populations. However, the absolute ancestry values were quite variable. Comparisons made for panels with different numbers of markers and different ability to differentiate the three main

reference populations showed that the differences obtained were a function of these two variables. Depending on the profile of the population, it was observed that the performance of the studied AIM sets was related to the differentiation levels between reference populations as well as the equilibrium between these values. Therefore, obtaining reliable ancestry estimates in Admixed American populations not only depends on the selection of markers with high differentiation capacity but also on a balance of the differentiation values between the source populations (Galanter et al., 2012; Kidd et al., 2014; Phillips, 2015). The present study showed that the populations with the highest NAM ancestry were those, whose estimates had increased associated error. For these populations, this study also showed that more accurate estimates can be obtained when analyzing the 46 indels from Pereira et al. (2012) and the 164 SNPs of the Precision ID Ancestry panel together.

The discrepancies observed among panels at the individual level were higher than those at the population level. Particularly for the NAM component, the large differences observed in all populations regardless of the panel point to low accuracies of the estimates. These differences were also observed between the ancestries of parents vs. offspring, as well as between full siblings from the Rio de Janeiro population. The correlation and agreement between the ancestry estimates increased with the number of markers analyzed.

The high correlation and agreement between parents vs. offspring showed that this can be a good strategy for the evaluation of the performance of different panels. Although the admixture-enabled selection was described in the same Latin American populations that we studied from the 1000 genomes (Norris et al., 2020), this phenomenon was restricted to coding genes and not expected for the markers included in most of the sets selected for forensic use.

In forensic genetics, AIMs can be useful for BGA inference, as an investigative lead in the absence of a suspect (Phillips, 2015; Mogensen et al., 2020). To this end, it is, however, necessary that the relevant population is included in the investigated database. To evaluate this, Tvedebrink et al. (2018) derived a measure of agreement (z-score) that indicates whether a profile may come from a population that is represented within those being assessed. The results of the z-score analysis in 65 sibling pairs from Rio de Janeiro resulted in a large number of AIM profiles that were outliers in the true population. There was also a high number of ambiguous results, most of which were profiles that could belong to Rio de Janeiro and European populations. Moreover, increasing the number of AIMs did not increase the sensibility, although the specificity was higher. It is worth noting that no other South American populations were included, which would most certainly reduce the specificity even more. These results point out the complexity of BGA inference in highly admixed populations as those from South America and the large variation in the admixture proportions present in the population from Rio de Janeiro.

In a recent study, Pfaffelhuber et al. (2020) found high misclassification errors for the continental origin when Admixed American populations are included in the analysis of BGA. These authors concluded that, even for the AIM sets with the

best performance in BGA inferences, when Admixed American populations were considered the misclassification was too large (30%) for forensic applications.

In summary, we illustrated the differences that can be expected when inferring ancestry or the populational origin of genetic profiles from South American admixed populations. Similar differences are expected to be present in other AIM sets with comparable characteristics in terms of the number of markers and genetic differentiation among source populations. Ancestry estimates are not only influenced by the number of markers included in the panel, but it is also essential to assess the level of differentiation that these markers provide among the reference populations. As seen in this work, there is a fine balance in the interplay of these factors.

The analysis of ancestry estimates at the population and individual levels helped to disclose what aspects to consider when selecting markers for an ancestry inference panel. Nevertheless, ancestry analyses will always present some degree of error when performing individual and population assignments. The focus should be to identify strategies for marker selection that minimize the error rate and increase the accuracy of the ancestry inference. Notwithstanding, the results obtained showed that even when the differences in estimates at the population level were minimized through the selection of a balanced group of markers or the use of the combined set, the errors at the individual level remained too high, demonstrating the need for a much higher number of markers for this purpose.

In the future, it would be interesting to perform investigations considering panels with higher resolution and also explore admixed populations with different number of source contributors to compare how the number of parental populations influences the ancestry results for different AIM panels.

Although it was not the scope in this work, an aspect to consider when inferring ancestry is the impact of the selection of appropriate reference populations. The admixture patterns in South America present differential contributions of several African and European populations from different regions along the continent. As an example, recent studies have attested that the presence of Northern Europeans is more restricted to the South, whereas Western European admixture events are more generalized (Montinaro et al., 2015; Gouveia et al., 2020).

The panels evaluated in this work have been designed to maximize differences between continents and are commonly used to ascertain main continental ancestry contributions. Indeed, previous studies reported absence of fine resolution within Sub-Saharan African, European, and East Asian groups (Al-Asfi et al., 2018; Lee et al., 2018; Nakanishi et al., 2018; Mogensen et al., 2020). Finer-scale admixture patterns within the South American continent have most recently been addressed with genome wide studies based on high density SNP data (Montinaro et al., 2015; Chacón-Duque et al., 2018; Ongaro et al., 2019; Gouveia et al., 2020). These studies have attested the complexity of the admixture dynamics of South America.

For the purposes of direct comparison of different datasets and other literature data, we have considered 100 Yorubans, 100 Central and British Europeans, and 47 Native Americans from several groups as references for all the populations studied.

We used all available data for Native Americans and selected a random subset of 100 Africans and Europeans, to avoid large differences in the effective size between reference datasets. These individuals (and the reduced sample size of each reference group) are not necessarily the most appropriate references when looking particularly at the history of the Rio de Janeiro population. However, this work aimed to investigate how ancestry inferences fluctuate according to the number of loci used, the balance of the AIM panels, and the differentiation these AIMS provide. As such, the number and populations used for reference data will have minor impact on the conclusions of the study. Nevertheless, we should highlight that when assessing ancestry patterns for population and forensic genetic studies, it is important to consider the specific history of each population, and select a collection of reference individuals that is representative and better reflects those events.

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 Comitê de Ética em Pesquisa, Universidade do Estado do Rio de Janeiro. The patients/participants provided their written informed consent to participate in this study.

REFERENCES

- Al-Asfi, M., McNevein, D., Mehta, B., Power, D., Gahan, M. E., and Daniel, R. (2018). Assessment of the Precision ID Ancestry panel. *Int. J. Leg. Med.* 132, 1581–1594. doi: 10.1007/s00414-018-1785-9
- Almeida, A. P. F., Simão, F., Aquino, J. G., Carvalho, E. F., and Gusmão, L. (2017). Contrasting admixture estimates in Rio de Janeiro obtained by different sampling strategies. *For. Sci. Int.* 6, e89–e91. doi: 10.1016/j.fsigs.2017.09.046
- Alves, H. B., Leite, F. P., Sotomaior, V. S., Rueda, F. F., Silva, R., and Moura-Neto, R. S. (2014). STR data for 15 autosomal STR markers from Parana (Southern Brazil). *Int. J. Leg. Med.* 128, 269–270. doi: 10.1007/s00414-013-0859-y
- Aquino, J. G., Jannuzzi, J., Carvalho, E. F., and Gusmão, L. (2015). Assessing the suitability of different sets of indels in ancestry estimation. *For. Sci. Int.* 5, e34–e36. doi: 10.1016/j.fsigen.2015.09.014
- Bonferroni, C. E. (1936). *Teoria Statistica Delle Classi e Calcolo Delle Probabilità*. Firenze: Istituto Superiore di Scienze Economiche e Commerciali di Firenze.
- Buckleton, J., Curran, J., Goudet, J., Taylor, D., Thiery, A., and Weir, B. S. (2016). Population-specific FST values for forensic STR markers: a worldwide survey. *For. Sci. Int. Genet.* 23, 91–100. doi: 10.1016/j.fsigen.2016.03.004
- Bulbul, O., Cherni, L., Khodjet-El-Khil, H., Rajeevan, H., and Kidd, K. K. (2016). Evaluating a subset of ancestry informative SNPs for discriminating among Southwest Asian and circum-Mediterranean populations. *For. Sci. Int. Genet.* 23, 153–158. doi: 10.1016/j.fsigen.2016.04.010
- Bulbul, O., Speed, W. C., Gurkan, C., Soundararajan, U., Rajeevan, H., Pakstis, A. J., et al. (2018). Improving ancestry distinctions among Southwest Asian populations. *For. Sci. Int. Genet.* 35, 14–20. doi: 10.1016/j.fsigen.2018.03.010

AUTHOR CONTRIBUTIONS

VP and LG conceived and supervised the study, and wrote the first draft of the manuscript. RS and AA were responsible for samples collection, DNA extraction, and genotyping. VP, TT, and LG performed the statistical analysis of the data. CB and NM helped with data interpretation and manuscript drafting. All authors critically revised and approved the final manuscript.

FUNDING

LG was supported by the Ellen og Aage Andersen's Foundation, Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (ref. 306342/2019-7), and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro – FAPERJ (CNE-2018).

ACKNOWLEDGMENTS

The authors would like to thank Rui Pereira for providing and retrieving the data for the 46 indels, and Nadia Jochumsen for laboratory technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00966/full#supplementary-material>

- Chacón-Duque, J. C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuña-Alonzo, V., Barquera, R., et al. (2018). Latin Americans show widespread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* 9:5388. doi: 10.1038/s41467-018-07748-z
- Chakraborty, R., Srinivasan, M. R., and Daiger, S. P. (1993). Evaluation of standard error and confidence interval of estimated multilocus genotype probabilities, and their implications in DNA forensics. *Am. J. Hum. Genet.* 52, 60–70.
- Cheung, E. Y. Y., Phillips, C., Eduardoff, M., Lareu, M. V., and McNevein, D. (2019). Performance of ancestry-informative SNP and microhaplotype markers. *For. Sci. Int. Genet.* 43:102141. doi: 10.1016/j.fsigen.2019.102141
- Curran, J. M., Buckleton, J. S., Triggs, C. M., and Weir, B. S. (2002). Assessing uncertainty in DNA evidence caused by sampling effects. *Sci. Justice* 42, 29–37. doi: 10.1016/S1355-0306(02)71794-71792
- de Assis Poiares, L., de Sa Osorio, P., Spanhol, F. A., Coltre, S. C., Rodenbusch, R., Gusmao, L., et al. (2010). Allele frequencies of 15 STRs in a representative sample of the Brazilian population. *For. Sci. Int. Genet.* 4, e61–e63. doi: 10.1016/j.fsigen.2009.05.006
- Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x
- Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587.
- Galanter, J. M., Fernandez-Lopez, J. C., Gignoux, C. R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., et al. (2012). Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet.* 8:e1002554. doi: 10.1371/journal.pgen.1002554

- Gouveia, M. H., Borda, V., Leal, T. P., Moreira, R. G., Bergen, A. W., Kehdy, F. S. G., et al. (2020). Origins, admixture dynamics, and homogenization of the African gene pool in the Americas. *Mol. Biol. Evol.* 37, 1647–1656. doi: 10.1093/molbev/msaa033
- Hessab, T., Aranha, R. S., Moura-Neto, R. S., Balding, D. J., and Schrago, C. G. (2018). Evaluating DNA evidence in a genetically complex population. *For. Sci. Int. Genet.* 36, 141–147. doi: 10.1016/j.fsigen.2018.06.019
- Hessab, T., Carvalho, R. M., Souza, M., Martha, S. F., Garrido, R. G., Freitas, N. F., et al. (2015). Genetic data on 17 STR autosomal loci for a sample population of the State of Rio de Janeiro, Brazil. *For. Sci. Int. Genet.* 14, e4–e5. doi: 10.1016/j.fsigen.2014.10.001
- Homburger, J. R., Moreno-Estrada, A., Gignoux, C. R., Nelson, D., Sanchez, E., Ortiz-Tello, P., et al. (2015). Genomic insights into the ancestry and demographic history of South America. *PLoS Genet.* 11:e1005602. doi: 10.1371/journal.pgen.1005602
- Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233
- Jung, J. Y., Kang, P. W., Kim, E., Chacon, D., Beck, D., and McNevin, D. (2019). Ancestry informative markers (AIMs) for Korean and other East Asian and South East Asian populations. *Int. J. Leg. Med.* 133, 1711–1719. doi: 10.1007/s00414-019-02129-2127
- Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., et al. (2014). Progress toward an efficient panel of SNPs for ancestry inference. *For. Sci. Int. Genet.* 10, 23–32. doi: 10.1016/j.fsigen.2014.01.002
- Kidd, K. K., Speed, W. C., Pakstis, A. J., Podini, D. S., Lagacé, R., Chang, J., et al. (2017). Evaluating 130 microhaplotypes across a global set of 83 populations. *For. Sci. Int. Genet.* 29, 29–37. doi: 10.1016/j.fsigen.2017.03.014
- Kling, D., Dell'Amico, B., and Tillmar, A. O. (2015). FamLinkX - implementation of a general model for likelihood computations for X-chromosomal marker data. *For. Sci. Int. Genet.* 17, 1–7. doi: 10.1016/j.fsigen.2015.02.007
- Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., et al. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutat.* 30, 69–78. doi: 10.1002/humu.20822
- Lee, J. H., Cho, S., Kim, M. Y., Shin, D. H., Rakha, A., Shinde, V., et al. (2018). Genetic resolution of applied biosystems™ precision ID Ancestry panel for seven Asian populations. *Leg. Med.* 34, 41–47. doi: 10.1016/j.legalmed.2018.08.007
- Li, C. X., Pakstis, A. J., Jiang, L., Wei, Y. L., Sun, Q. F., Wu, H., et al. (2016). A panel of 74 AISNPs: improved ancestry inference within Eastern Asia. *For. Sci. Int. Genet.* 23, 101–110. doi: 10.1016/j.fsigen.2016.04.002
- Manta, F. S., Pereira, R., Caiafa, A., Silva, D. A., Gusmao, L., and Carvalho, E. F. (2013). Analysis of genetic ancestry in the admixed Brazilian population from Rio de Janeiro using 46 autosomal ancestry-informative indel markers. *Ann. Hum. Biol.* 40, 94–98. doi: 10.3109/03014460.2012.742138
- Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nat. Genet.* 36, 512–517. doi: 10.1038/ng1337
- Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Gravel, S., Daly, M. J., et al. (2017). Population genetic history and polygenic risk biases in 1000 Genomes populations. *Am. J. Hum. Genet.* 100, 635–649. doi: 10.1101/070797
- Mogensen, H. S., Tvedebrink, T., Borsting, C., Pereira, V., and Morling, N. (2020). Ancestry prediction efficiency of the software GenoGeographer using a z-score method and the ancestry informative markers in the Precision ID Ancestry Panel. *For. Sci. Int. Genet.* 44:102154. doi: 10.1016/j.fsigen.2019.102154
- Montinaro, F., Busby, G. B. J., Pascali, V. L., Myers, S., Hellenthal, G., and Capelli, C. (2015). Unravelling the hidden ancestry of American admixed populations. *Nat. Commun.* 6:6596.
- Moriot, A., Santos, C., Freire-Aradas, A., Phillips, C., and Hall, D. (2018). Inferring biogeographic ancestry with compound markers of slow and fast evolving polymorphisms. *Eur. J. Hum. Genet.* 26, 1697–1707. doi: 10.1038/s41431-018-0215-212
- Moura, R. R., Coelho, A. V., Balbino Vde, Q., Crovella, S., and Brandao, L. A. (2015). Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. *Am. J. Hum. Biol.* 27, 674–680. doi: 10.1002/ajhb.22714
- Moyses, C. B., Tsutsumida, W. M., Raimann, P. E., da Motta, C. H., Nogueira, T. L., Dos Santos, O. C., et al. (2017). Population data of the 21 autosomal STRs included in the GlobalFiler(R) kits in population samples from five Brazilian regions. *For. Sci. Int. Genet.* 26, e28–e30. doi: 10.1016/j.fsigen.2016.10.017
- Nakanishi, H., Pereira, V., Borsting, C., Yamamoto, T., Tvedebrink, T., Hara, M., et al. (2018). Analysis of mainland Japanese and Okinawan Japanese populations using the precision ID Ancestry Panel. *For. Sci. Int. Genet.* 33, 106–109. doi: 10.1016/j.fsigen.2017.12.004
- Nassir, R., Kosoy, R., Tian, C., White, P. A., Butler, L. M., Silva, G., et al. (2009). An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genet.* 10:39. doi: 10.1186/1471-2156-10-39
- Norris, E. T., Rishishwar, L., Chande, A. T., Conley, A. B., Ye, K., Valderrama-Aguirre, A., et al. (2020). Admixture-enabled selection for rapid adaptive evolution in the Americas. *Genome Biol.* 21:29. doi: 10.1186/s13059-020-1946-1942
- Ongaro, L., Scliar, M. O., Flores, R., Raveane, A., Marnetto, D., Sarno, S., et al. (2019). The genomic impact of European Colonization of the Americas. *Curr. Biol.* 29, 3974.e4–3986.e4. doi: 10.1016/j.cub.2019.09.076
- Pena, S. D., Di Pietro, G., Fuchshuber-Moraes, M., Genro, J. P., Hutz, M. H., Kehdy Fde, S., et al. (2011). The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One* 6:e17063. doi: 10.1371/journal.pone.0017063
- Pereira, R., Phillips, C., Pinto, N., Santos, C., dos Santos, S. E., Amorim, A., et al. (2012). Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS One* 7:e29684. doi: 10.1371/journal.pone.0029684
- Pereira, V., Freire-Aradas, A., Ballard, D., Borsting, C., Diez, V., Pruszkowska-Przybylska, P., et al. (2019). Development and validation of the EUROFORGEN NAME (North African and Middle Eastern) ancestry panel. *For. Sci. Int. Genet.* 42, 260–267. doi: 10.1016/j.fsigen.2019.06.010
- Pereira, V., Mogensen, H. S., Borsting, C., and Morling, N. (2017). Evaluation of the Precision ID Ancestry Panel for crime case work: a SNP typing assay developed for typing of 165 ancestral informative markers. *For. Sci. Int. Genet.* 28, 138–145. doi: 10.1016/j.fsigen.2017.02.013
- Pfaffelhuber, P., Grundner-Culemann, F., Lipphardt, V., and Baumdicker, F. (2020). How to choose sets of ancestry informative markers: a supervised feature selection approach. *For. Sci. Int. Genet.* 46:102259. doi: 10.1016/j.fsigen.2020.102259
- Phillips, C. (2015). Forensic genetic analysis of bio-geographical ancestry. *For. Sci. Int. Genet.* 18, 49–65. doi: 10.1016/j.fsigen.2015.05.012
- Phillips, C., McNevin, D., Kidd, K. K., Lagace, R., Wootton, S., de la Puente, M., et al. (2019). MAPlex - A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. *For. Sci. Int. Genet.* 42, 213–226. doi: 10.1016/j.fsigen.2019.06.022
- Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., et al. (2014). Building a forensic ancestry panel from the ground up: the euroforgen global AIM-SNP set. *For. Sci. Int. Genet.* 11, 13–25. doi: 10.1016/j.fsigen.2014.02.012
- Phillips, C., Salas, A., Sanchez, J. J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., et al. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *For. Sci. Int. Genet.* 1, 273–280. doi: 10.1016/j.fsigen.2007.06.008
- Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* 11, 459–463. doi: 10.1038/nrg2813
- Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.
- R Core Team, (2013). *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rajeevan, H., Soundararajan, U., Pakstis, A. J., and Kidd, K. K. (2012). Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb. *Investig. Genet.* 3:18. doi: 10.1186/2041-2223-3-18
- Rajeevan, H., Soundararajan, U., Pakstis, A. J., and Kidd, K. K. (2020). FrogAncestryCalc: a standalone batch likelihood computation tool for ancestry inference panels catalogued in FROG-kb. *For. Sci. Int. Genet.* 46:102237. doi: 10.1016/j.fsigen.2020.102237

- Rodrigues, E. M., Palha Tde, J., and dos Santos, S. E. (2007). Allele frequencies data and statistic parameters for 13 STR loci in a population of the Brazilian Amazon Region. *Forensic Sci. Int.* 168, 244–247. doi: 10.1016/j.forsciint.2006.03.003
- Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., et al. (2002). Genetic structure of human populations. *Science* 298, 2381–2385. doi: 10.1126/science.1078311
- Salzano, F. M., and Sans, M. (2014). Interethnic admixture and the evolution of Latin American populations. *Genet. Mol. Biol.* 37 (1 Suppl.), 151–170. doi: 10.1590/s1415-47572014000200003
- Santangelo, R., Gonzalez-Andrade, F., Borsting, C., Torroni, A., Pereira, V., and Morling, N. (2017). Analysis of ancestry informative markers in three main ethnic groups from Ecuador supports a trihybrid origin of Ecuadorians. *For. Sci. Int. Genet.* 31, 29–33. doi: 10.1016/j.fsigen.2017.08.012
- Santos, H. C., Horimoto, A. V., Tarazona-Santos, E., Rodrigues-Soares, F., Barreto, M. L., Horta, B. L., et al. (2016). A minimum set of ancestry informative markers for determining admixture proportions in a mixed American population: the Brazilian set. *Eur. J. Hum. Genet.* 24, 725–731. doi: 10.1038/ejhg.2015.187
- Simão, F., Ferreira, A. P., de Carvalho, E. F., Parson, W., and Gusmao, L. (2018). Defining mtDNA origins and population stratification in Rio de Janeiro. *For. Sci. Int. Genet.* 34, 97–104. doi: 10.1016/j.fsigen.2018.02.003
- Soundararajan, U., Yun, L., Shi, M., and Kidd, K. K. (2016). Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. *For. Sci. Int. Genet.* 23, 25–32. doi: 10.1016/j.fsigen.2016.01.013
- Tian, C., Gregersen, P. K., and Seldin, M. F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Hum. Mol. Genet.* 17, R143–R150. doi: 10.1093/hmg/ddn268
- Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., and Morling, N. (2017). GenoGeographer – A tool for genogeographic inference. *For. Sci. Int. Genet.* 6, e463–e465. doi: 10.1016/j.fsigen.2017.09.196
- Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., and Morling, N. (2018). Weight of the evidence of genetic investigations of ancestry informative markers. *Theor. Popul. Biol.* 120, 1–10. doi: 10.1016/j.tpb.2017.12.004
- Verdugo, R. A., Di Genova, A., Herrera, L., Moraga, M., Acuna, M., Berrios, S., et al. (2020). Development of a small panel of SNPs to infer ancestry in Chileans that distinguishes Aymara and Mapuche components. *Biol. Res.* 53:15. doi: 10.1186/s40659-020-00284-285
- Vullo, C., Gomes, V., Romanini, C., Oliveira, A. M., Rocabado, O., Aquino, J., et al. (2015). Association between Y haplogroups and autosomal AIMs reveals intra-population substructure in Bolivian populations. *Int. J. Leg. Med.* 129, 673–680. doi: 10.1007/s00414-014-1025-x
- Ye, Z., Wang, Z., and Hou, Y. (2020). Does Bonferroni correction "rescue" the deviation from Hardy-Weinberg equilibrium? *For. Sci. Int. Genet.* 46:102254. doi: 10.1016/j.fsigen.2020.102254
- Yuasa, I., Akane, A., Yamamoto, T., Matsusue, A., Endoh, M., Nakagawa, M., et al. (2018). Japaneseplex: a forensic SNP assay for identification of Japanese people using Japanese-specific alleles. *Leg. Med.* 33, 17–22. doi: 10.1016/j.legalmed.2018.04.008

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

Copyright © 2020 Pereira, Santangelo, Borsting, Tvedebrink, Almeida, Carvalho, Morling and Gusmao. 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.



Twenty Years Later: A Comprehensive Review of the X Chromosome Use in Forensic Genetics

Iva Gomes^{1,2}, Nádia Pinto^{1,2,3}, Sofia Antão-Sousa^{1,2,4,5}, Verónica Gomes^{1,2}, Leonor Gusmão⁵ and António Amorim^{1,2,4*}

¹ Institute for Research and Innovation in Health Sciences (i3S), University of Porto, Porto, Portugal, ² Institute of Molecular Pathology and Immunology, University of Porto (IPATIMUP), Porto, Portugal, ³ Center of Mathematics, Faculty of Sciences, University of Porto, Porto, Portugal, ⁴ Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal, ⁵ DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil

OPEN ACCESS

Edited by:

Kenneth K. Kidd,
Yale University, United States

Reviewed by:

Guanglin He,
Sichuan University, China
Carlo Robino,
University of Turin, Italy

*Correspondence:

António Amorim
aamorim@ipatimup.pt

Specialty section:

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

Received: 31 May 2020

Accepted: 24 July 2020

Published: 17 September 2020

Citation:

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

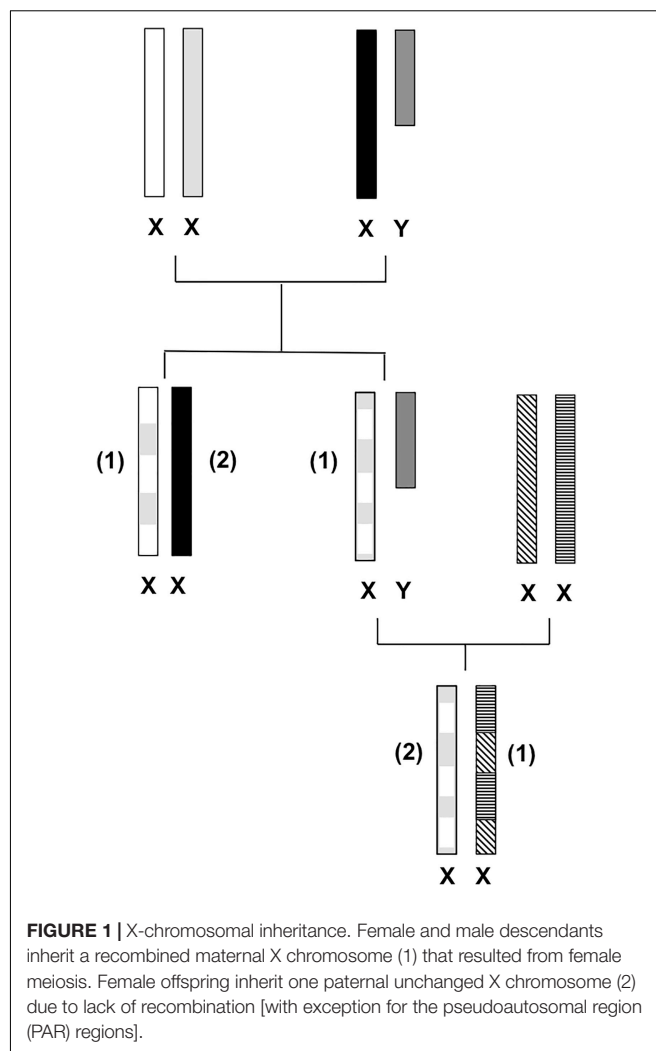
The unique structure of the X chromosome shaped by evolution has led to the present gender-specific genetic differences, which are not shared by its counterpart, the Y chromosome, and neither by the autosomes. In males, recombination between the X and Y chromosomes is limited to the pseudoautosomal regions, PAR1 and PAR2; therefore, in males, the X chromosome is (almost) entirely transmitted to female offspring. On the other hand, the X chromosome is present in females with two copies that recombine along the whole chromosome during female meiosis and that is transmitted to both female and male descendants. These transmission characteristics, besides the obvious clinical impact (sex chromosome aneuploidies are extremely frequent), make the X chromosome an irreplaceable genetic tool for population genetic-based studies as well as for kinship and forensic investigations. In the early 2000s, the number of publications using X-chromosomal polymorphisms in forensic and population genetic applications increased steadily. However, nearly 20 years later, we observe a conspicuous decrease in the rate of these publications. In light of this observation, the main aim of this article is to provide a comprehensive review of the advances and applications of X-chromosomal markers in population and forensic genetics over the last two decades. The foremost relevant topics are addressed as: (i) developments concerning the number and types of markers available, with special emphasis on short tandem repeat (STR) polymorphisms (STR nomenclatures and practical concerns); (ii) overview of worldwide population (frequency) data; (iii) the use of X-chromosomal markers in (complex) kinship testing and the forensic statistical evaluation of evidence; (iv) segregation and mutation studies; and (v) current weaknesses and future prospects.

Keywords: X chromosome short tandem repeats (X-STRs), X chromosome markers, forensic genetics, population genetics, kinship testing, X chromosome short tandem repeat (X-STR) mutation rates

INTRODUCTION

The X chromosome has many characteristics that are not shared by its counterpart, the Y chromosome, or by any of the autosomes of the mammalian genome. Its unique structural characteristics have been shaped by evolution, leading to the present known gender-specific genetic differences (Lahn and Page, 1999; Schaffner, 2004). In males, the single copy of the X chromosome does not allow for recombination to occur (except for the pseudoautosomal regions, PARs, which maintain homology by recombining during male meiosis). The non-recombining regions and the PAR 1 and PAR 2 regions of the X and Y chromosomes have taken different evolutionary paths becoming highly differentiated due to different functional roles, and consequently, only a few X-Y sequence similarities remain among them (Lahn and Page, 1999). Mutation events have gathered on the Y chromosome, and in addition to the lack of recombination, these events have contributed to the loss of most of the Y chromosome's sequence and genes emerging in a distinctive configuration of repeated sequences (Lahn and Page, 1999; Schaffner, 2004) becoming specialized in male sex determination. On the other hand, the X chromosome has preserved its autosomal character, becoming one of the most stable nuclear chromosomes, holding the largest and most conserved gene arrangement across eutherian ("placental") mammals (Lahn and Page, 1999; Kohn et al., 2004; Schaffner, 2004). It is the only chromosome to have one of its pair inactivated in one sex (females), and it is "corrupted" with repeat elements, making it especially tough to produce a detailed gene sequence (Gunter, 2005). In 2005, Ross et al. (2005) published the first draft that covered approximately 99.3% of the human X chromosome euchromatic sequence. The X chromosome holds a size length of approximately 155 million base pairs (Mb) (Ross et al., 2005), representing nearly 5% of the estimated human genome size (3,200 Mb) (Lander et al., 2001). Regarding some of the X chromosome's structural properties, it presents a low GC content (39%) when compared to 41% of the genome average (Ross et al., 2005). The low number of functional genes detected confers the chromosome one of the lowest gene densities among the chromosomes annotated to date (Ross et al., 2005). The X chromosome's sequence data revealed not only a low concentration of genes but also small gene length as only 1.7% of the chromosome sequence is represented by exons of the identified genes, responsible for transcribing 33% of the X chromosome (Ross et al., 2005). The particular genetic characteristics of the X chromosome, shaped by evolution, are responsible for the distinctive gender-specific features (**Figure 1**): in the male gender, the X chromosome is (almost entirely) transmitted to females as an unchanged block. While in females, the two copies recombine, like autosomes, reorganizing genetic variation in each generation, which contributes to the increase of haplotype diversity (Schaffner, 2004). The new reshuffled chromosome is then transmitted to female and male descendants (**Figure 1**).

These specific properties – two recombining copies in females and a single non-recombining copy in males (except for the PAR regions) creating haplotypes – provide X chromosome markers



a particular place in forensics and in population genetics, as well as in other research areas such as human evolutionary studies and medical genetics (e.g., X-linked recessive disorders such as hemophilia or Duchenne muscular dystrophy) (Szibor, 2007). Regarding forensic and population genetic applications, the X chromosome's mode of inheritance places this chromosome among the autosomes and the uniparental-inherited genomes [mitochondrial DNA (mtDNA) and Y chromosome] providing desirable and exclusive features that are not provided by any other of the latter.

In the early 2000s, the number of publications using X-chromosomal polymorphisms in these areas of research increased steadily. However, nearly 20 years later, a conspicuous decrease in the rate of these publications is observed. For example, X chromosome short tandem repeat (X-STR) forensic-based publications reached as many as 43 publications in a single year (2009), while in the past year of 2019, only 18 publications were found (complete results and detailed information on the criteria used for database search are presented and discussed under the section "Factors Underlying the Relative Stagnation

in X Chromosome Forensic Research”). In light of these observations, the main aim of the present work is to provide an up-to-date and objective review of the advances and applications of X-chromosomal markers in population and forensic genetics over the last two decades since the bloom observed in the early and mid-2000s.

CURRENT DEVELOPMENTS: NUMBERS AND TYPES OF X-CHROMOSOMAL MARKERS AVAILABLE (SHORT TANDEM REPEATS, SINGLE-NUCLEOTIDE POLYMORPHISMS, AND INSERTIONS/DELETIONS)

The use of X chromosome polymorphisms in human identification and in population genetics is mainly supported by the potential applications that outcome from its unique properties. Solely, or complementing the information provided by the autosomes or by markers located on the Y chromosome or mtDNA, X chromosome markers may provide essential information in many different lines of research. It must be highlighted that identity testing using X-STRs in particular contexts, namely in scenarios of (complex) kinship testing, may be the only tool to unravel certain cases. Examples of complex kinship testing scenarios where the prominent role of X chromosome polymorphisms is demonstrated are given in the section “The Use of X-Chromosomal Markers in (Complex) Kinship Testing.”

In the present section, we will try to draw the state of the art of the genetic markers that have been described, to date, in the X chromosome-specific region, i.e., leaving out PARs and amelogenin. Special attention is given to X-STRs as a result of their favorite usage in forensic genetics due to high standardization and existence of commercial typing kits. Although some of the first publications reporting X-STRs appeared in the late 90s (Edwards et al., 1991, 1992; Hearne and Todd, 1991; Sleddens et al., 1992), the beginning of the century marked the increase of X-STR publications that focused on the development of new multiplexes on the genetic characterization of many different population groups (databasing) and on kinship and forensic investigations.

An extensive literature review was undertaken, with special focus on forensic-population genetic publications. Results are analyzed and tabulated separately for each type of marker, including relevant references. **Supplementary Table 1** lists 85 STR loci in which usage in forensic-population genetic context was reported. In agreement with the study of Szibor et al. (2005), HumARA marker was not considered for ethical reasons. Although the number of X chromosome markers has grown since the 2007 seminal review of Szibor (2007), this growth may be illusory, since many markers were used quite rarely, sometimes only once.

Although a considerable number of X-STRs are available in the literature, a better view of their real, current usage may be given by the analysis of the multiplexes, which have been described

for their genotyping. **Table 1** shows the most used in-house and commercially developed X-STR multiplexes in which we update the revision of Diegoli (2015) and demonstrate clearly that the effective number of STRs routinely used is modest.

In any case, due to their high degree of discrimination, the number of standardized STRs is sufficient for most routine investigations, as will be discussed below in the section “The Use of X-Chromosomal Markers in (Complex) Kinship Testing.” Novel interesting STRs for forensic applications continue being described (Nishi et al., 2020). Despite the wide set of available X-STR markers as well as many population-based studies (see section “Overview of Worldwide (Published) X Chromosome Short Tandem Repeat Population Data”) that have emerged over these years, no effective X-STR database exists harboring this type of data. Some of the published population datasets are available in the FamLinkX web page¹ in a format that can be directly uploaded for kinship calculation using the software. Efforts were made by Szibor et al. (2006) to create an X-STR database² (ChrX-Str.org 2.0, 2020) that could anchor population data (namely, haplotype frequencies), calculation of forensically relevant parameters, information on markers such as multiplex kits, etc. Nevertheless, it seems that no updates have been made to this database, specifically in what regards population data submission, as only four populations are currently available (German, Ghanesen, Japanese, and Chinese Han) (“Haplotypes”; see text footnote 2). In addition, it is however noteworthy that no autosomal STR database such as NIST STRbase (National Institute of Standards, and Technology [Nist], 2020) or STRidER (2020)³ contains information on X-STRs either. This approach could be considered: autosomal types of database could potentially serve as harbor for X-STR data undergoing the same quality control (QC) submission criteria. In fact, several forensic-focused journals such as the *Forensic Science International: Genetics* and the *International Journal of Legal Medicine* have published minimum requirements for publication of forensic population data from different genomic markers (e.g., autosomal, Y-chromosomal, mtDNA) (Parson and Roewer, 2010; Gusmão et al., 2017). Submission of such data to these journals requires preliminary QC assessment and inclusion in public online databases. These requirements could certainly be applied to X-chromosomal type of markers, ensuring the same quality type of data submitted. STRs are undoubtedly the preferential markers in human identification applications. Some of the main features that make STRs desirable markers are (i) highly polymorphic, i.e., high discriminating capacity between individuals; (ii) technical easiness due to rapid analysis with PCR-based technology and capillary electrophoresis automated fluorescent detection; and (iii) ability for generating STR multiplexes with small amplicon lengths for degraded DNA. The same cannot be said about insertions/deletions (INDELs), although these share some of the features of STRs (technical ease of analyses by PCR and ability for multiplexing), standardization is much less advanced perhaps due to the

¹<http://famlink.se/Databases/>

²<http://www.chrx-str.org/>

³<https://strider.online/>

TABLE 1 | Most used multiplex PCR assays targeting X chromosome short tandem repeat (STR) markers. References do not necessarily refer to the original development papers.

Name	References	Nr. and STR loci
Goldeneye 17X kit	Gao et al. (2019)	16 (DXS6795, DXS9902, DXS8378, HPRTB, GATA165B12, DXS7132, DXS7424, DXS6807, DXS6803, GATA172D05, DXS6800, DXS10134, GATA31E08, DXS10159, DXS6789, and DXS6810)
Investigator® Argus X-12 QS (Qiagen) kit	Elakkary et al. (2014)	12 (DXS7132, DXS7423, DXS8378, DXS10074, DXS10079, DXS10101, DXS10103, DXS10134, DXS10135, DXS10146, DXS10148, and HPRTB)
Microreader™ 19X ID System kit	Lin et al. (2020)	19 (DXS6795, DXS6803, DXS6807, DXS9907, DXS7423, GATA172D05, DXS101, DXS9902, DXS7133, DXS6810, GATA31E08, DXS6800, DXS981, DXS10162, DXS6809, GATA165B12, DXS10079, DXS10135, and HPRTB)
AGCU X19 STR Kit	Li et al. (2017)	19 (DXS8378, DXS7423, DXS10148, DXS10159, DXS10134, DXS7424, DXS10164, DXS10162, DXS7132, DXS10079, DXS6789, DXS101, DXS10103, DXS10101, HPRTB, DXS6809, DXS10075, DXS10074, and DXS10135)
–	Deng et al. (2017)	19 (DXS8378, DXS9898, DXS7133, GATA31E08, GATA172D05, DXS7423, DXS6809, DXS7132, DXS9902, DXS6789, DXS8378, DXS7423, DXS7132, DXS10079, DXS6801, DXS6799, DXS6800, DXS10075, DXS6807, and DXS6803)
–	Prieto-Fernández et al. (2016)	17 (DXS9895, GATA144D04, DXS10077, DXS10078, DXS10161, DXS10160, DXS981, DXS6800, DXS6803, DXS9898, DXS6801, DXS6799, DXS6797, DXS7133, DXS6804, GATA172D05, DXS8377, DXS10146, and DXS10147)
GHEP-ISFG decaplex	Gusmão et al. (2009)	10 (DXS8378, DXS9898, DXS7133, GATA31E08, GATA172D05, DXS7423, DXS6809, DXS7132, DXS9902, and DXS6789)
–	Zhang et al. (2017b)	15 (DXS6807, DXS8378, DXS6795, DXS10164, DXS7132, DXS10074, DXS6803, DXS6801, DXS101, DXS7133, GATA165B12, DXS10103, HPRTB, GATA31E08, and DXS7423)
MiSeq FGx™ Forensic Genomics	Jäger et al. (2017)	7 (HPRTB, DXS7132, DXS7423, DXS8378, DXS10074, DXS10103, and DXS10135)

need of a much higher number of markers for a high degree of discrimination among individuals. Nevertheless, INDELs represent another potential tool for addressing human genetic identification issues. In **Table 2**, we list the X chromosome-specific INDEL polymorphisms genotyping systems described in forensic literature.

Unsurprisingly, not as many X chromosome INDEL marker systems have been described as compared to autosomal INDELs (e.g., Pereira et al., 2009; Freitas et al., 2010; Zaumsegel et al., 2013). In fact, no commercial kits being available, few systems have been subject to interlaboratorial comparisons, as in the case of autosomal INDELs, which stood international collaborative exercises (Pereira et al., 2018). One of the possible motifs for the lack of commercial kits is possibly due to the limited applications of X chromosome polymorphisms in forensic genetics when compared to autosomal markers. An interesting alternative typing approach, however, albeit of difficult analysis, is the one described in the studies of Fan et al. (2015, 2016) in which amplicons comprise various INDELs, i.e., biallelic loci that are tightly linked composing a new marker and that are amplified by a single pair of PCR primers.

TABLE 2 | X chromosome specific insertion/deletion (INDEL) polymorphisms genotyping systems. CE, capillary electrophoresis.

Number of loci	Genotyping system	References
32	Single multiplex (CE)	Pereira et al. (2012)
33	Single multiplex (CE)	Freitas et al. (2010)
16 (from a total of 45 mixed marker system)	Single multiplex (CE)	Tao et al. (2019)
17 (from a total of 60 mixed STR system)	Massive Parallel Sequencing	Zhang et al. (2017b)
21	Single multiplex (CE)	Edelmann et al. (2016)

With respect to X chromosome single nucleotide polymorphisms (X-SNPs), the analysis of the state of the art is even more complex due to the diversity of non-standardized genotyping systems and platforms, which have not been submitted to interlaboratorial comparisons. In **Table 3**, a summary of the actual forensic use of X chromosome-specific SNPs is shown. The number of table entries gives a false impression of abundance of X-SNPs; in fact, besides the

TABLE 3 | X chromosome-specific single-nucleotide polymorphism (SNP) genotyping systems.

Number of SNPs	Genotyping system	References
28 (from a total of 60 mixed marker systems)	Massive parallel sequencing	Zhang et al. (2017b)
39 (from a total of 273 mixed marker panels)	Massive parallel sequencing	Zhang et al. (2017a)
27 (from a total of 1,204 mixed marker panels)	Massive parallel sequencing	Hwa et al. (2018)
62	MALDI-TOF mass spectrometry	Stepanov et al. (2016)
17 (from a total of 220 mixed marker panels)	MALDI-TOF mass spectrometry	Hwa et al. (2019)
5 (from a total of 41 mixed marker panels)	MALDI-TOF mass spectrometry	Petkovski et al. (2005)
10	qPCR (TaqMan probes)	Zarrabeitia et al. (2007)
25	SNaPshot	Tomas et al. (2010)
16	SNaPshot	Oki et al. (2012)
14	qPCR (Taqman probes)	Li et al. (2010)

mentioned limitations, the number of SNPs overlap is very low. Although the binary nature of SNPs may favor degraded DNA as well as automation and high-throughput genotyping (e.g., in individual identification using complex kinship analyses in highly degraded scenarios such as natural or human-made disasters), the information content is considerably lower than for STR loci and consequently a larger number of SNPs are needed to match the discrimination power of the commonly used STRs (e.g., Chakraborty et al., 1999; Amorim and Pereira, 2005). Consequently, more loci mean more amplification products, which increases difficulty in data interpretation of DNA profile mixtures. In a multiple-donor sample interpretation, identification of each contributor may be very complex with biallelic systems. The limited number of alleles for each locus (normally two alleles) becomes hard to interpret because overlap will occur and multiple donors become hard to distinguish (Butler et al., 2007; Budowle and van Daal, 2008). Adding the mentioned data interpretation complexity in mixed profiles to the limited applications of X chromosome markers can potentially justify the lack of interest in X-SNPs observed.

OVERVIEW OF WORLDWIDE (PUBLISHED) X CHROMOSOME SHORT TANDEM REPEAT POPULATION DATA

For an overview of the worldwide population allele frequency datasets of X-STRs used in forensic genetics, we have consulted the articles available in PubMed database and in the congress proceedings of the International Society for Forensic Genetics⁴ (The International Society for Forensic Genetics [ISFG], 2020).

⁴www.isfg.org

This search resulted in a total of 269 articles. The first genetic studies with focus on genotyping X-STRs for forensic application start emerging in the year 1999. Since then, and until 2008, a remarkable increase of population data publications was observed (**Figure 2A**). Nevertheless, reported information on human X-STRs in different worldwide populations has been stagnating in the last years.

Information concerning the populations, number of male and female samples, and X-STRs analyzed was compiled using 236 publications out of the 269 consulted (see **Supplementary Table 2**). The remaining were excluded for different reasons, which include articles that were not in English, with overlapping data (in this case, the most updated dataset was considered), and with unclear information concerning population, markers, or total samples analyzed. Therefore and although some of these studies contain relevant information on X-STR variation (e.g., the study by Edelmann et al., 2006, which has data for DXS9908 and DXS7127 markers), these were not included in **Supplementary Table 2**. Furthermore, the study by Phillips et al. (2018) reports data on seven X-STRs for a large sample of 944 individuals from the HGDP-CEPH human genome diversity panel. However, since this dataset comprises samples from 51 populations with relatively low sample sizes, the results were compiled for seven continentally defined population groups, namely, African (sub-Saharan), European, Middle East (including North Africans), Central-South Asian, East Asian, Oceanian, and Native American.

In **Figure 2B**, it is possible to observe that the number of X-STRs analyzed is highly variable among publications with some studies genotyping a high number of X-STRs (e.g., Liu et al., 2013; Fukuta et al., 2019) and others genotyping a reduced number of loci (e.g., Watanabe et al., 2000; Koyama et al., 2002; Carvalho and Pinheiro, 2011). The number of markers included in each study varied from 1 to 27. In 47% of the cases, this number was between 10 and 12 X-STRs, in 31%, it was below 10, and 22% of the datasets included more than 12 makers (**Figure 2B**). The number of markers available per dataset is somehow related to the use of commercial kits in 37.4% of the population studies (**Supplementary Table 2**). The first commercial kit that was optimized for forensic applications was the Argus X-UL from Biotype (Dresden, Germany), containing four X-STRs (DXS8378, DXS7132, HPRTB, and DXS7423) located in distant positions along the chromosome to avoid linkage. This kit was soon expanded (Argus X-8) with four additional X-STRs (DXS10135, DXS10074, and DXS10134), creating four pairs of linked X-STRs. The Argus X-12 (Qiagen, Hilden, Germany) is the most recent version of the Argus kit and is the most widely used (an optimized version is now available, the Argus X-12 QS, but that contains the same markers). It comprises 12 X-STRs organized in four linkage groups: LG1, DXS10148/DXS10135/DXS8378; LG2, DXS7132/DXS10079/DXS10074; LG3, DXS10103/HPRTB/DXS10101; and LG4, DXS10146/DXS10134/DXS7423. The Goldeneye DNA ID System 17X (Goldeneye Technology Co., Ltd., Beijing, China) and the AGCU X19 STR kit (Wuxi Sino-German Meilian Biotechnology Co., Jiangsu, China) were also developed for forensic applications, although available data are virtually restricted to Chinese populations. Among in-house

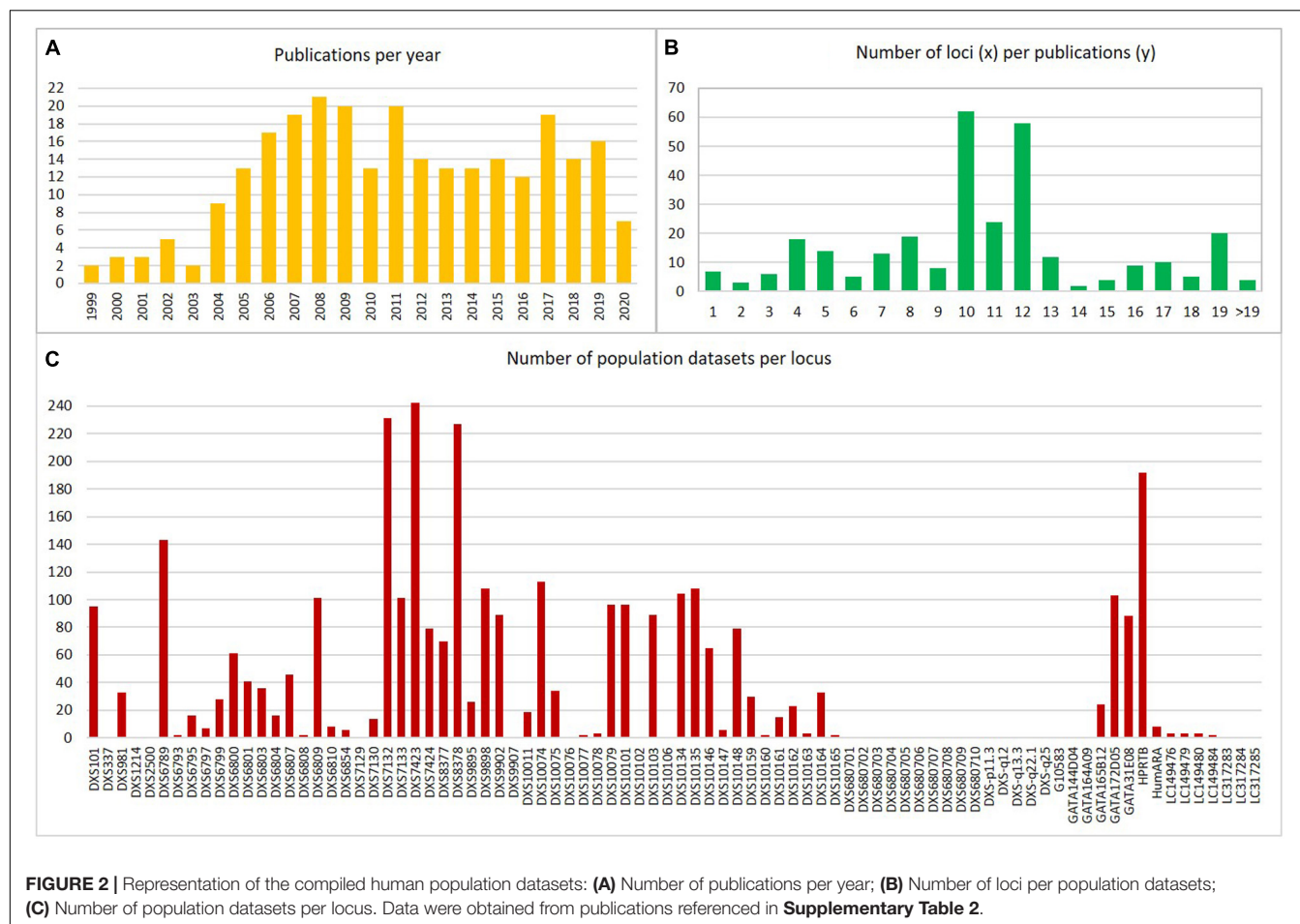


FIGURE 2 | Representation of the compiled human population datasets: **(A)** Number of publications per year; **(B)** Number of loci per population datasets; **(C)** Number of population datasets per locus. Data were obtained from publications referenced in **Supplementary Table 2**.

multiplexes, the Decaplex system developed by the GHEP-ISFG (Spanish and Portuguese Speaking working group of the ISFG) (Gusmão et al., 2008) has been the most widely used (14.6% of the population datasets were generated using this multiplex).

From the 84 markers that have been described as informative for forensic applications [including HumARA that is no longer used due to ethical issues (Szibor et al., 2005), as already mentioned], less than 50% were studied in more than 10 populations, and 29 were only reported in a single population (**Figure 2C**). The loci with more allele frequency data accumulated are those included in the commercial kits (namely, Investigator Argus X-12 kit, Qiagen) or in the in-house-developed Decaplex-GHEP-ISFG (Gusmão et al., 2008).

In **Supplementary Table 2**, the geographical distribution of the published human population data for X-STRs since 1999 is described. Notwithstanding the exhaustive nature of this review, it is possible that some studies are missing from this table. However, we believe that most forensic population studies on X-STRs have been identified, allowing a realistic picture of the state of the art. For a broader overview of the populations sampled, we have represented the number of datasets that have been published until now by country (**Figure 3**). The datasets were counted considering the number of subpopulations or ethnic groups in each publication. Populations defined at continental level (namely, the HGDP-CEPH and

Africa datasets) or belonging to ethnic affiliated populations from different countries (namely, the Jews) have been excluded. In **Figure 3**, it is possible to observe that apart from a lack of X-STR data information for many countries, there is high heterogeneity among and inside continents. Data are scarcer in some geographical areas, namely, for sub-Saharan African and American populations (except for Argentina, Brazil, and United States). On the other hand, a large quantity of X-STR data was obtained for other populations, such as the ones from China. China is by far the best represented country not only because of the higher number of publications but also due to the inclusion of various ethnic groups in a single study. Although for some countries a large number of datasets are available for the same X-STR loci, many of those studies characterize different regions or subpopulations, which is relevant to investigate population stratification inside the country, especially when a high diversity of ethnicities coexists.

Overall, the compiled information clearly shows an imbalance between the total number of publications and the asymmetric representation of the worldwide populations. In fact, for several populations from different geographic regions, data on X-STR remain largely scarce, being the available information representative of only a small fraction of the worldwide human populations. Moreover, apart from a large variation concerning the X-STRs included in each study, many only comprise a small

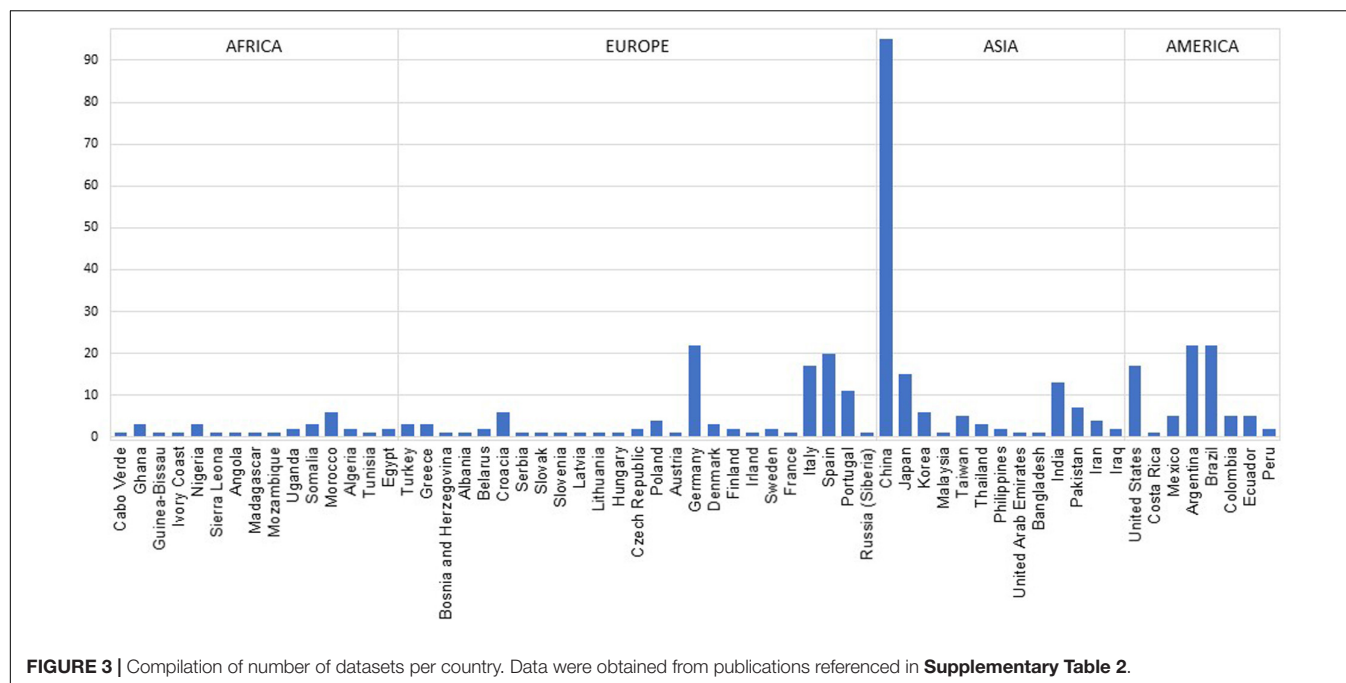


FIGURE 3 | Compilation of number of datasets per country. Data were obtained from publications referenced in **Supplementary Table 2**.

number of loci. Due to proximity on the chromosome, it is expected that some of the studied markers will be in linkage disequilibrium (LD) in many populations. However, data on haplotype frequencies are almost restricted to recent papers and not available for most publications consulted, invalidating the use of some of the available data in forensic applications.

Therefore, further studies on haplotype frequency distributions, as well as on mutation rates and LD, are mandatory to attain the final goal of establishing highly comprehensive and representative human reference X-STR databases.

SHORT TANDEM REPEAT NOMENCLATURES AND PRACTICAL CONCERNS

Accuracy and common nomenclature are of fundamental importance to secure error-free communication, data exchange, and data comparison among laboratories. STR nomenclature, independently of marker genome location, has been long addressed by several studies (e.g., Lazaruk et al., 2001; Gusmão et al., 2002; Gomes et al., 2008, 2009, 2016; Gettings et al., 2015) as well as by the ISFG and other DNA groups (e.g., Bär et al., 1997; Olaisen et al., 1998; Gill et al., 2001; Gusmão et al., 2006).

The observed increase of X-STR studies over the years justifies the need to evaluate X-STR nomenclature being used at least for the most common polymorphisms. Several studies have gathered considerable sequencing data for some of the commonly used X-STRs (Gomes et al., 2008, 2009, 2016, 2017; Szibor et al., 2009). In these latter studies, relevant findings were reported for several markers, which demonstrate that accurate allele nomenclature designation taking into consideration the ISFG recommendations (Gusmão et al., 2006) would have had a major

impact on allele assignment. One of the major gaps seen in several studies is the lack of sequencing data for, at least, the three major population groups (Asian, African, and Caucasian) when new markers are proposed as usually only one group is analyzed. This approach reduces possible interpopulational variation and avoids genotyping problems when different groups are genotyped. This was the case for the first version of the most used X-STR commercial kit, the Investigator Argus X-12 (Qiagen). The markers in this kit were characterized mostly in individuals of European ancestry and therefore some of the genetic variations detected in other population groups were missed out (Tillmar et al., 2017). Once other population groups of other ancestries were studied, several markers presented high frequencies of silent alleles that had gone previously undetected (e.g., Tomas et al., 2012; Gomes et al., 2016, 2017; Tillmar et al., 2017). For example, the silent alleles for some of the loci were mostly caused by a mismatch at one of the primer binding sites (Gomes et al., 2016, 2017). After several reports on this matter, a new version was developed, the Investigator Argus X-12 QS (Qiagen), containing the same markers but with new primer designs for some of the X-STRs to resolve the high frequency of allele dropouts. Another example of inaccurate nomenclature assignment was the case of HPRTB. In the study of Pereira et al. (2007), peculiar results during population comparison analyses of a Northern Portuguese population sample with other European groups were found. These findings led to a deeper investigation, leading to the discovery of issues behind the HPRTB nomenclature (Szibor et al., 2009). In this latter report, authors described that two different nomenclatures were being used among the forensic genetic community, leading to a shift in allele frequencies and consequently errors in data resulting from population comparisons-based analyses (e.g., Pereira et al., 2007).

Finally, as proposed by the ISFG recommendations on the use of X-chromosome markers (Tillmar et al., 2017), the previous recommendations on allele nomenclature already recognized for autosomal and Y-chromosomal-specific markers (Bär et al., 1997; Gill et al., 2001; Gusmão et al., 2006) can also be applied to X-STRs without the need for particular changes. It seems that very few studies take these recommendations into thoughtful consideration and no real significant advances have been made in this field. Accuracy in sequence variation and repeat structure and nomenclature of X-STRs are empirical and pending issues in forensic and population genetics research that are still often neglected.

THE USE OF X-CHROMOSOMAL MARKERS IN (COMPLEX) KINSHIP TESTING

The standard procedure to quantify the genetic evidence in kinship analyses relies upon independent autosomal markers and is grounded in Bayes' theorem. Typically, equal priors are considered, and a likelihood ratio (LR) comparing the probability of the observations assuming a pair of alternative, mutually exclusive, kinship hypotheses is computed (Gjertson et al., 2007). Indeed, autosomal information is the one generally considered, despite currently available X-chromosomal markers being able to provide great statistical power in some cases (Szibor et al., 2003; Krawczak, 2007; Szibor, 2007; Pinto et al., 2011a, 2013a; Gomes et al., 2012). From the set of the latter cases obviously excluded are those where there is a link "father-son" in both main and alternative hypotheses, as, for instance, in a "paternal grandfather-granddaughter" vs "unrelated" case analyzing a pair of individuals, as the first, when considering X-chromosomal transmission, equated to the second (Pinto et al., 2011a, 2012). In any case, the preference given to autosomal markers is easily justified and understood not only for allowing the same approach for each kinship problem, regardless of the sex of the involved individuals, but also because of independent transmission of the markers and, at least in most of the populations, absence of LD. Conversely, the analysis of X chromosome markers offers little room to consider only independently transmitted loci, and thus recombination rates and haplotype frequencies are in general required for statistical evaluation of the evidence.

Non-random association of alleles of different loci at a population-level LD (also known as gametic association) can result from population events like drift, selection, non-random mating, or admixture (Hedrick, 1987; Medina-Acosta, 2011). A close physical location of the markers, as well as population stratification, will influence the re-establishing of equilibrium. Consequently, LD results neither can be extrapolated from one population to another, nor are stable, even in a closed population, as recombination progressively breaks it. Moreover, haplotype frequencies

cannot be inferred from allelic ones, and direct counting needs to be carried out.

Closely located markers are said to be in linkage if they are more prone to be inherited together, as a unit, than independently. Linkage between markers depends on chromosomal recombination rate (or frequency). Two markers are unlinked if recombination between them is expected to occur in each meiosis so that half of the gametic products would be recombinant and thus recombination fraction takes the value of 0.50. Obviously, linked markers are more prone to be in LD. Segregation analyses in one or multi generation family studies were performed, aiming to estimate recombination rates between X-STRs of interest through proper bioinformatic pipelines that take into account the possibility of mutation (Nothnagel et al., 2012; Diegoli et al., 2016; Bini et al., 2019), but population-based studies, as HapMap project (The International HapMap Consortium, 2007), can also be considered (Phillips et al., 2012). Mapping functions as Haldane's (Haldane, 1919) or Kosambi's (Kosambi, 1944) are used to convert genetic distances between markers in recombination rates. It is however noteworthy that in some kinship problems, as the one involving a pair of females and the hypotheses maternity and unrelated, the linkage is not needed to be taken into account as it cancels in the LR numerator and denominator (Tillmar et al., 2017). A general framework to understand in which case linkage has to be considered is still lacking, despite being known that disregarding it may lead to a significant over- or under-quantification of the genetic evidence (Tillmar et al., 2011; Kling et al., 2015b).

Contrarily to what occurs for autosomes, where a plethora of markers from 22 chromosomes can be chosen, linkage and LD are unavoidable issues in the case of X-chromosomal analysis. Due to the length of the X chromosome, a maximum of four unlinked X-STRs are estimated to be liable of being simultaneously analyzed. On the other hand, higher LD values are expected for X-chromosomal markers than for autosomes since recombination only occurs in female meioses, which have also smaller mutation rates than males (Shimmin et al., 1993; Schaffner, 2004). Finally, it should be noted that estimates of haplotype frequencies are not as accurate as the allelic ones since much larger databases are required: just considering a simple illustrative example, a set of three loci with 10 alleles each can potentially entail the estimation of 1,000 haplotype frequencies.

Few software packages are available for kinship evaluations considering X-chromosomal transmission, FamLinkX being the most relevant, taking into account the possibility of mutation, linkage, and LD (Tillmar et al., 2011; Kling et al., 2015a). Also, software to weigh the *a priori* power of a marker to exclude a claimed relationship was already developed (Egeland et al., 2014), and the ISFG recently provided general guidelines for using X-chromosomal markers in kinship testing (Tillmar et al., 2017).

Kinship Testing and the Identity-by-Descent Framework

Considering a number of generations beyond which individuals are assumed to be unrelated, kinship measurements are

based on the concept of identity-by-descent. Two alleles are called identical-by-descent (IBD) if they are copies of a given ancestral allele. Barring mutation, two alleles which are identical by descent must be therefore identical-by-state (IBS). For autosomal transmission, nine IBD partitions can be established considering the four alleles of a pair of individuals and their relationship (Jacquard, 1974; Weir et al., 2006; Pinto et al., 2010). This number reduces to three if non-inbred individuals are considered, likewise occurring for X-chromosomal transmission between a pair of females (Pinto et al., 2011a, 2012). Regarding X-chromosomal transmission, there are four IBD partitions involving a female–male pair (two if assuming a non-inbred female) and two for a pair of males (Pinto et al., 2011a). Independently of the mode of genetic transmission considered, the probabilities of the genotypic observations, assuming a specific hypothesis of kinship, depend on the IBD probabilities of the pedigree and on the frequency of the alleles (Weir et al., 2006; Pinto et al., 2011a). Pedigrees with the same IBD coefficients are said to belong to the same kinship class, as they are, theoretically, undistinguishable through the use of unlinked markers (Pinto et al., 2010, 2012). In **Table 4**, IBD probabilities are presented for a pair of non-inbred individuals considering autosomal and X-chromosomal modes of genetic transmission and a set of commonly analyzed relationships. Algebraic formulae for the probabilities of the observations, given the identity by descent partitions, can be found in Weir et al. (2006) and Pinto et al. (2010, 2011a, 2012), respectively, for autosomes and X-chromosomal markers. Finally, it should be noted that, assuming X-chromosomal mode of transmission, relationships are not symmetrical as probabilities of IBD sharing may differ. For example, while a pair of paternal aunt–nephew does not share X-IBD alleles (being thus equated to unrelated from the X-chromosomal point of view), a pair of paternal uncle–niece shares one pair of IBD alleles with 50% of chance.

Regardless of the mode of genetic transmission considered, striking statistical results could be obtained when the sharing of IBD alleles is mandatory, unless mutation occurs, for one of the two kinship hypotheses considered. For example, in a standard paternity problem (“unrelated” as alternative hypothesis), the probability of sharing a pair of IBD autosomal alleles (and thus IBS, barring mutation) is one, under the main hypothesis, and null under the alternative. In cases with daughters, this is also true for X-chromosomal markers, providing a higher *a priori* paternity exclusion power than autosomal ones (Krawczak, 2007; Pinto et al., 2013a).

In some cases, as in disaster victim identification problems, specific kinship hypotheses cannot be established, and a broader measure of kinship can be established to weigh the degree of relatedness before specifying more detailed hypotheses. In these cases, the coancestry coefficient, i.e., the probability of selecting two IBD alleles when each one is randomly chosen from each individual, can be computed. In this case, the analysis of the X chromosome can be of major importance as, in all the cases where transmission

is not interrupted by a “father–son” link, the expected IBD sharing is at least the same as for autosomes – see **Table 5**, since no randomness is possible in the X-allele of a male. Coancestry coefficients can be estimated through the genotypes of the individuals (Pinto et al., 2011b, 2013b) and the combination of both types of genetic information can provide valuable insights on the genetic kinship linking the individuals.

Parenthood Testing

The X-chromosomal markers can be used to complement autosomal information when inconclusive or weak statistical results are achieved in standard parenthood testing where the alternative hypothesis is the individuals being unrelated. This can be due to the poor quality or low quantity of DNA in degraded samples, resulting in few analyzed markers or to other, more complex, situations where few Mendelian incompatibilities are found.

Compared with autosomes, X-chromosomal markers provide greater statistical power in trios, in paternity duos with daughters, and in maternity duos with sons. The X-chromosomal markers are not informative in paternity cases with sons, and for mother/daughter duos, the same statistical power is obtained for autosomal and X-chromosomal transmission.

When few Mendelian incompatibilities are found, this can be due to the alleged parent of the child being related to the true parent. A relatively common situation is the alleged father being either a full brother or the father of the true father of the child, in which case the probability of the alleged father and child sharing a pair of IBD alleles is 50%. In a paternity testing with a daughter, if the alleged father is a brother of the real one, the probability of uncle–niece sharing a pair of IBD X-alleles is also 50%. In all the other cases, this probability is null. Indeed, the analysis of X-chromosomal markers can be an efficient approach for excluding close relatives of the real father, unknowingly presented in a standard paternity case (Gomes et al., 2012).

Beyond Parenthood

In some cases, the alleged parent is not available for analysis, and sibship, or grandparenthood problems may emerge. In some of these cases, X-chromosomal markers can provide invaluable information, stronger than the one provided by autosomes. The most striking examples are those where the sharing of a pair of IBD X-alleles is mandatory. This occurs when the paternity of a daughter is questioned, being the alleged father unavailable for analysis, contrarily to his (unquestioned) mother or daughter. In both cases, the sharing of IBS alleles between analyzed females is mandatory for all the markers, unless mutation occurs. In these cases, the reached statistical power is the same for a paternity testing with autosomes when the alleged father is directly analyzed whether the mother of the child is available for analysis or not.

Another illustrating example is the kinship problem where the hypotheses are “full sisters” versus “unrelated.”

TABLE 4 | Probability of two individuals sharing two, one, or no pairs of identical-by-descent (IBD) alleles, assuming a specific kinship for both autosomal (Aut) and X-chromosomal (X chr) modes of genetic transmission.

Pair of shared IBD alleles			Two		One		None	
Mode of transmission			Aut	X chr	Aut	X chr	Aut	X chr
Relationship								
Female–Female								
Identical twins/Identity			1		0		0	
Mother–Daughter			0		1		0	
Full-Sisters			1/4	1/2	1/2	1/2	1/4	0
Grandmother–Granddaughter	Maternal	0	0	1/2	1/2	1/2	1/2	
	Paternal				1		0	
Aunt–Niece	Maternal				3/4		1/4	
	Paternal				1/2		1/2	
Half-sisters	Maternal				1/2		1/2	
	Paternal				1		0	
Unrelated			0		0		1	
Female–Male								
Father–Daughter/Mother–Son			0	–	1		0	
Full brother–sister			1/4		1/2	1/2	1/4	1/2
Grandfather–Granddaughter/Grandmother–Grandson	Maternal	0		1/2	1/2	1/2	1/2	
	Paternal				0		1	
Uncle–Niece	Maternal				1/4		3/4	
	Paternal				1/2		1/2	
Aunt–Nephew	Maternal				3/4		1/4	
	Paternal				0		1	
Half-brother–sister	Maternal				1/2		1/2	
	Paternal				0		1	
Unrelated					0		1	
Male–Male								
Identical twins/Identity			1	–	0	1	0	
Father–Son			0		1	0	0	1
Full-brothers			1/4		1/2	1/2	1/4	1/2
Grandfather–Grandson	Maternal	0		1/2	1/2	1/2	1/2	
	Paternal				0		1	
Uncle–Nephew	Maternal				1/4		3/4	
	Paternal				0		1	
Half-brothers	Maternal				1/2		1/2	
	Paternal				0		1	
Unrelated					0		1	

Considering X-chromosomal transmission and the main hypothesis, females share either two or one pair of IBD X-alleles with the same probability: 50%. Assuming autosomal transmission, they may not share IBD alleles (with 25% of chance), such as occurs assuming they are unrelated (with 100% of chance). It is then expected that X-chromosomal markers provide stronger results than autosomes. This occurs in all the kinships where the transmission of the X chromosome is not interrupted due to its obligatory transmission between father and daughter, which allows the skipping of one meiosis.



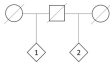
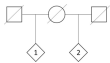
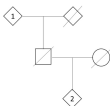
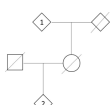
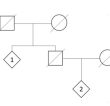
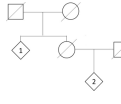
Incest Cases

In some cases, the high number of homozygosities shown by a child (e.g., in a paternity testing with alleged father excluded) may raise the suspicion of an incestuous situation. This may, under some circumstances, configure a crime (mother

under age or with intellectual disability, for example). In the case of a daughter, X-chromosomal analyses may provide important insights even without analyzing the alleged father. If the father of the daughter is also the father of the mother and, in the absence of mutation, either the child is homozygous (for one allele present in the mother) or is heterozygous for the same alleles of the mother. In the case of autosomal transmission, three alleles can be seen in mother/daughter pair, as for the case of the parents being unrelated.

The hypotheses of the father of the child being either the father or the full brother of the mother are theoretically indistinguishable when considering unlinked autosomal markers. Contrastingly, in the case of daughters, X-chromosomal markers can provide insights allowing the different weighing of the two hypotheses (Pinto et al., 2011a).

TABLE 5 | Probability of choosing a pair of identical-by-descent (IBD) alleles when one allele is randomly chosen from each individual. Numbers in superscript in header refer to the sex of the individuals represented in genealogies.

Kinship	Coancestry	Female ¹ –Female ²	Male ¹ –Male ²	Female ¹ –Male ²	Male ¹ –Female ²
General	Aut-chr	$1/2k_2 + 1/4k_1$			
	X-chr	$1/2x_2 + 1/4x_1$	x_1	$1/2x_1$	$1/2x_1$
Parenthood	Aut-chr	$1/4$			
	X-chr	$1/4$	0	$1/2$	$1/2$
Full-sibship	Aut-chr	$1/4$			
	X-chr	$3/8$	$1/2$	$1/4$	$1/4$
Paternal half-sibship	Aut-chr	$1/8$			
	X-chr	$1/4$	0	0	0
Maternal half-sibship	Aut-chr	$1/8$			
	X-chr	$1/8$	$1/2$	$1/4$	$1/4$
Paternal grandparenthood	Aut-chr	$1/8$			
	X-chr	$1/4$	0	0	0
Maternal grandparenthood	Aut-chr	$1/8$			
	X-chr	$1/8$	$1/2$	$1/4$	$1/4$
Paternal avuncular	Aut-chr	$1/8$			
	X-chr	$1/8$	0	0	$1/4$
Maternal avuncular	Aut-chr	$1/8$			
	X-chr	$3/16$	$1/4$	$3/8$	$1/8$

* k_i , probability of sharing i pairs of IBD autosomal alleles; x_i , probability of sharing i pairs of IBD X-chromosomal alleles.

Distinguishing Pedigrees Belonging to the Same Autosomal Kinship Class

Pedigrees are theoretically indistinguishable, considering unlinked markers, whenever they have the same IBD partitions (Pinto et al., 2010). This is the case of the second-degree relatives: avuncular, half-siblings and grandparent–grandchild, as the probability of individuals sharing two pairs of IBD alleles is null, while the probability of sharing one pair of IBD autosomal alleles is equal to the probability of sharing none (50%) – see Table 4. Nevertheless, the analysis of X-chromosomal markers can provide differential weighing favoring one of the alternative hypotheses (Pinto et al., 2011a). For example, when a pair of females is analyzed, maternal and paternal

aunt/niece can be distinguished from, respectively, maternal and paternal half-sisters and grandmother–granddaughter, which are not distinguishable among them even when considering X-chromosomal markers. In all the cases, females cannot share two pairs of IBD alleles, but a pair of maternal aunt/niece shares one pair of IBD alleles with a probability equal to 75%, while for both maternal half-sisters and grandmother–granddaughter pairs, this probability reduces to 50%. On the other, if both pairs of paternal half-sisters and grandmother–granddaughter have to share one pair of IBD alleles, this probability drops from 100 to 50% in the case of paternal aunt/niece. Different IBD probabilities will result in different weighing of the evidence, depending on the genotypic observations.

SEGREGATION STUDIES: CURRENT DATA AND MISSING DATA

The high power of discrimination that characterizes STRs and makes them desirable genetic markers compared to SNPs or INDELs, particularly in human identification analysis (such as kinship testing), is due to their higher mutation rate. An STR is, by definition, a tandemly arrayed repetition of a DNA fragment of one to six base pairs. There is general consensus that these are created by random mutations (Levinson and Gutman, 1987; Schlötterer, 2000). Generally, STRs with four base pairs motifs are plentiful and more stable than two or three nucleotide repeats; hence, they have been favored when designing the commercially available forensic kits (Pereira and Gusmão, 2016). Motifs with two or three base pairs are less stable and have a higher propensity for stutter during PCR, and STRs with more base pairs are less frequent. When a somatic mutation occurs, it affects only cell lines of the individual where it occurred. However, when a mutation occurs in the germ line, it has the potential of being passed on to the offspring and resulting in different parental and filial alleles. Mutation rates vary between types of polymorphisms and also on inherent individual characteristics such as sex and age (Brinkmann et al., 1998; Nachman and Crowell, 2000).

Polymerase template slippage is thought to be the primary mutational mechanism leading to changes in STR length (Schlötterer and Tautz, 1992; Strand et al., 1993), and mutations involving the loss or gain of one repeat are assumed to be preponderant over mutations involving the loss or gain of multiple repeats. Slippage occurs during DNA replication when the two DNA strands come apart. When misalignment occurs out of register the repeat number of the STR product will be different. The currently accepted mutational model, also known as the stepwise mutation model (SMM) (Ohta and Kimura, 1973) occurring as a result of DNA replication slippage, includes mutational forces working in opposite directions: polymerase template slippage and point mutations; the latter reduce the length of STRs due to the breakage of the original segment creating two new shorter segments. Studies have shown that the longer the allele length, the higher is the frequency of these events. It has also been reported that longer alleles tend to mutate to shorter alleles and vice versa, while intermediate-sized alleles have approximately the same tendency to shorten or lengthen (Primmer et al., 1996; Brinkmann et al., 1998; Xu et al., 2000; Antão-Sousa et al., 2019).

In forensic casework context, the estimation of mutation rates is crucial for the analysis, interpretation, and quantification of experimental data and for the proper quantification of LR. In such scenarios, the detection of mutation(s) has practical consequences in the interpretation of the genetic profiles. Some studies have addressed this by analyzing different familial configurations, familial duos, mother–son, mother–daughter, and father–daughter, and familial trios, father–mother–daughter (e.g., Jin et al., 2016; Burgos et al., 2019; García et al., 2019). **Supplementary Table 3** presents the most updated information on mutation rates per marker and per familial configuration for the most commonly used X-STRs. To date, not much research on

the mutation rates of the most commonly used X-STRs has been given, and therefore, data collection and analyses are still lacking. Perhaps one of the limitations in the estimation of mutation rates of STRs, in general, is the use of the (most frequently used) method for mutation estimates based on direct pedigree analysis. This means that mutated alleles are identified straightforward by the observation of allele transmissions in parent–child requiring a large amount of data to reliably estimate allele mutation rates. Having access to a high number of specific constellations of families may be a drawback to the (accurate) estimation of mutation rates of X-STRs.

DISCUSSION

Factors Underlying the Relative Stagnation in X Chromosome Forensic Research

After an initial boom, forensic research interest on X chromosome markers has witnessed a decline as judged by the number of relevant publications: 2000 (6), 2001 (7), 2002 (11), 2003 (18), 2004 (27), 2005 (25), 2006 (40), 2007 (35), 2008 (42), 2009 (43), 2010 (19), 2011 (41), 2012 (26), 2013 (22), 2014 (18), 2015 (15), 2016 (22), 2017 (31); 2018 (16), and 2019 (18) [search results obtained using Scopus database⁵ and the following criteria [ALL (dxs*) AND ALL (forensic)] AND PUBYEAR > 1999 AND PUBYEAR < 2020 on 30/04/2020]. In the beginning of the early 2000s, only a scarce number of X chromosome STRs and a very limited number of human population groups were characterized for forensic genetic applications. Data focusing on the assessment of X-linked polymorphisms for forensic and kinship genetic studies were an impending demand which created a gap in these fields producing sufficient ground for the interest in X chromosome markers and, in particular, X-STRs. Consequently, an increase of studies in 2003 until 2011 (with exception of the year 2010) can be noted. After this year, fluctuations are mostly toward a reduction of X-STR studies (except for 2017).

This implies that the practical forensic use of X chromosome is well below its potential and – what is most concerning – is that its use may be unsupported by research data and based on inadequately validated technical means and theoretically reduced or even incorrect analytical approaches. Enabling corrective actions demands therefore the identification of the causes of this slowing down of the forensically inclined research on X genetic markers. This fact has no parallel on the other sexual chromosome counterpart, the Y, to which a lot of attention is devoted, for example, by the STRbase (National Institute of Standards, and Technology [Nist], 2020) and has as well a very active dedicated site⁶, YHRD (2020), in contrast to the ChrX-STR.org 2, as mentioned previously (see text footnote 2).

In this section, we will analyze the putative change counteracting the loss of interest and analyzing the presumed reasons or factors justifying this situation,

⁵<http://www.scopus.com>

⁶<https://yhrd.org/>

which, from our point of view, can be classified into four broad categories: (a) theoretical and/or analytical, (b) technical, (c) statistical, and (d) medical/ethical, to be detailed below.

Theoretical and Analytical Difficulties

The main obstacle to the correct use of X chromosome in forensics lies in the hybrid nature of its formal genetic model of inheritance, common to most mammals, with very few exceptions (Cortez et al., 2014; Matveevsky et al., 2017). Indeed, as presented in the section “Introduction,” this chromosome harbors two distinct modes of transmission: the diploid, autosomal style (corresponding to the so-called pseudoautosomal regions), two in humans, PAR 1 and PAR 2 (Flaquer et al., 2009) and the sex-linked haplodiploid (for the rest of the chromosome, known as X-specific), which, due to the single copy in males, does not recombine.

When addressing X-chromosome markers, we are referring to the X-specific located ones. Therefore, only these will be analyzed (although some confusions do sometimes arise and quite often the status of X specificity may be doubtful – see below the technical section).

Even so, the formal genetic model of transmission and the consequences at the level of population genetics seem to be poorly understood by the forensic community, as judged by a recent analysis of the literature (Ferragut et al., 2019). It was shown that in 60% of 52 analyzed publications, forensic parameters were computed as for autosomal markers, and the analysis of associations between alleles from distinct loci (LD) was generally deficient or erroneous. In fact, linkage and LD concepts, particularly important for the X chromosome since all markers are located on the same chromosome, are often a source of confusion and generally lead to misinterpretation or even non-consideration of LD results in many genetic studies. Most studies using X-STRs correctly test for the presence of significant association among pairs of loci (LD) but fail to estimate haplotype frequencies and probability calculations, accordingly, when significant association is found among markers as loci must be analyzed together and not as individual markers in such cases. In 2017, recommendations were provided by the DNA commission of the ISFG addressing exactly the issues behind the concepts of linkage and LD in cases of kinship testing using X-STRs and emphasizing that “Haplotype frequencies should be used for likelihood calculations when LD exists” (Tillmar et al., 2017).

Similar issues have also arisen with the assessment of conformity with Hardy–Weinberg equilibrium expectations. Quite symptomatically, the ChrX-STR.org 2 website (see text footnote 2, accessed on 02/05/2020) has posted: “Based on the review of December 2018, it has been decided in cooperation with the X working group to remove the PI calculation from this website.”

From an applicable point of view, one can add that one of the additional problems to justify the decrease of interest in X chromosome markers could be due to the low number of identification cases that request X-STR markers. Perhaps

the troubles behind the implementation of a new system (financial cost and human resource training) which has a much more complex type of analysis when compared to the Y chromosome, for example, may not justify the need for the use of this system.

Technical Problems

Besides the genotyping problems, which may be transversal to all markers, irrespectively of the mode of transmission, sex chromosomes pose special difficulties due to their complex evolutionary history. In fact, apart from the PAR regions, X and Y chromosomes still keep substantial extensions of homologous regions, which obstruct the safe establishment of specificity for a marker, as well as its primers in case of PCR-based techniques. Particularly for recently X/Y transposed regions, this may constitute an (nearly) insurmountable obstacle (Lopes et al., 2004) as well as the dynamic state of the pseudoautosomal moving boundaries (Otto et al., 2011).

Statistical Issues

Most of the statistical problems (both at the descriptive level – parameter estimation level or hypothesis testing design or evidence quantitative evaluation) stem out of the theoretical flaws discussed above. Nonetheless, some are specifically empirical and are related to the haplodiploid specificity of the X chromosome: different sampling and estimation methods are required for each sex. Indeed, while haplotype frequencies can be estimated by simple counting in males, in females, they have to be inferred. Needless to say, simple haplotype frequency estimation requires prohibitively large sample sizes, growing exponentially with the number of loci involved (Amorim and Pinto, 2018).

Medical/Ethical Questions

To begin with, it must be highlighted that the very genotyping of sex chromosome markers for forensic purposes may represent a violation of some of the established recommendations and rules on the exclusion of any markers that can reveal physical traits [e.g., European Council Resolution of 25 June 2001 on the exchange of DNA analysis results (2001/C 187/01)]. Furthermore, gender and sex are always sensitive, and sometimes conflicting, categories in or for some individuals.

The evolutionary dynamics of sex chromosomes introduces also undesirable clinical and ethical problems. In fact, sex chromosomes are the Achilles’ heel of male meiosis (Kauppi et al., 2012). A non-negligible proportion (1/448 live births) of the human population carries some sort of chromosomal aberration and, for example, one of the aneuploidies, Klinefelter syndrome, has an incidence of ~1/500 male live births (Nielsen and Wohler, 1990). The consequences for forensic practice are ethically troublesome: discordance of external sex from X chromosome typing and unwilling disclosure of a clinical condition. In addition to X-chromosomal changes, several X-STR markers, that were or are still in use, have been linked to medical conditions. The HumARA is linked to spinal and bulbar muscular dystrophy (SBMA) as well as to other health risks

(Szibor et al., 2005). Another example is the possible LD between the STR alleles at HPRTB locus to the X-linked recessive disorder Lesch–Nyhan syndrome (caused by molecular defects within the HPRT gene) (Mansfield et al., 1993). Some data have shown that inheritance of two polymorphic tandem repeats, one being the HPRTB locus (mapped within intron 3 of the HPRT gene), could be used to establish linkage to the disease (Mansfield et al., 1993).

The X chromosome has had an interesting journey in the last two decades in the research fields of forensic and population genetics by providing new (population) data and aiding in the clarification of several issues, namely, in kinship testing. Its particular properties of inheritance (recombination on the female side and haploid state on the male side) have allowed this chromosome a role that cannot be accomplished by the autosomes neither by its counterpart, the Y chromosome. After an initial bloom of publications, several multiplex developments, workshops at international meetings, creation of an X-STR database, the interest in X-chromosomal markers is gradually fading. Analytical and statistical issues may be the major underlined motivations to the lack of interest in addition to a lower demand of X-STR-based identification cases.

Considerable effort has already been put in X-STRs, namely, (i) the generation of allelic and haplotypic frequency databases that include a fair enough number of geographically different located populations; (ii) several in-house multiplexes containing a large number of highly polymorphic markers as well as a sound established commercial kit; and (iii) relevant number of studies addressing and recommending solutions for the main issues surrounding X-STR kinship-based testing. Therefore, this effort should not be lost and move toward the revival of the standing position of X chromosome markers in forensic genetics.

REFERENCES

- Amorim, A., and Pereira, L. (2005). Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Sci. Int.* 150, 17–21. doi: 10.1016/j.forsciint.2004.06.018
- Amorim, A., and Pinto, N. (2018). Big data in forensic genetics. *Forensic Sci. Int. Genet.* 37, 102–105. doi: 10.1016/j.fsigen.2018.08.001
- Antão-Sousa, S., Amorim, A., Gusmão, L., and Pinto, N. (2019). Mutation in Y STRs: Repeat motif gains vs. losses. *Forensic Science International: Genetics Supplement Series* 7, 240–242. doi: 10.1016/j.fsigss.2019.09.092
- Bär, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., Lincoln, P., et al. (1997). DNA recommendations. Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics. *Int J Legal Med* 110, 175–176. doi: 10.1007/s004140050061
- Bini, C., Di Nunzio, C., Aneli, S., Sarno, S., Alù, M., Carnevali, E., et al. (2019). Analysis of recombination and mutation events for 12 X-Chr STR loci: A collaborative family study of the Italian Speaking Working Group Ge.F.I. *Forensic Science International: Genetics Supplement Series* 7, 398–400. doi: 10.1016/j.fsigss.2019.10.027
- Brinkmann, B., Klitsch, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics* 62, 1408–1415. doi: 10.1086/301869
- Budowle, B., and van Daal, A. (2008). Forensically relevant SNP classes. *Biotechniques* 44, 603–608, 610. doi: 10.2144/000112806
- Burgos, G., Posada, Y., Florez-Misas, A., Ávila, C., and Ibarra, A. (2019). An update of STR mutation rates from paternity tests analyzed in a 14 year period (2005–2018) at IdentiGEN lab, Universidad de Antioquia, Colombia. *Forensic Science International: Genetics Supplement Series* 7, 530–531. doi: 10.1016/j.fsigss.2019.10.078
- Butler, J. M., Coble, M. C., and Vallone, P. M. (2007). STRs vs. SNPs: thoughts on the future of forensic DNA testing. *Forensic sci med pathol* 3, 200–205. doi: 10.1007/s12024-007-0018-1
- Carvalho, R., and Pinheiro, M. F. (2011). Study of DXS9895 and DXS7130: Population data from North of Portugal. *J Forensic Leg Med.* 18, 21–22. doi: 10.1016/j.jflm.2010.11.010
- Chakraborty, R., Stivers, D. N., Su, B., Zhong, Y., and Budowle, B. (1999). The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis* 20, 1682–1696. doi: 10.1002/(SICI)1522-2683(19990101)20:8<1682::AID-ELPS1682<3.0.CO;2-Z
- ChrX-Str.org 2.0. (2020). *ChrX-STR.org 2.0*. <http://www.chrx-str.org/> (accessed May 15, 2020)
- Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., et al. (2014). Origins and functional evolution of Y chromosomes across mammals. *Nature* 508, 488–493. doi: 10.1038/nature13151
- Deng, C., Song, F., Li, J., Hou, Y., and Luo, H. (2017). Multiplex PCR for 19 X-chromosomal STRs in Chinese population. *Forensic Science International: Genetics Supplement Series* 6, e24–e26. doi: 10.1016/j.fsigss.2017.09.016
- Diegoli, T. M. (2015). Forensic typing of short tandem repeat markers on the X and Y chromosomes. *Forensic Sci. Int. Genet.* 18, 140–151. doi: 10.1016/j.fsigen.2015.03.013

AUTHOR CONTRIBUTIONS

IG drafted the manuscript scheme and, in addition, all authors have made a substantial, direct, and intellectual contribution to the work, and approved the final version for publication.

FUNDING

This work was partially financed by FEDER-Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020-Operacional Programme for Competitiveness and Internationalisation (POCI), Portugal 2020, and by Portuguese funds through FCT-Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the projects “Institute for Research and Innovation in Health Sciences” (POCI-01-0145-FEDER-007274). IG is funded by the FCT Scientific Stimulus program CEECIND/02609/2017. NP and VG are supported by FCT under the program contract provided in Decree-Law no.57/2016 of August 29. SA-S is funded by the FCT doctoral grant SFRH/BD/136284/2018. LG is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq (ref. 306342/2019-7) and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro-FAPERJ (CNE-2018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00926/full#supplementary-material>

- Diegoli, T. M., Rohde, H., Borowski, S., Krawczak, M., Coble, M. D., and Nothnagel, M. (2016). Genetic mapping of 15 human X-chromosomal forensic short tandem repeat (STR) loci by means of multi-core parallelization. *Forensic Sci. Int. Genet.* 25, 39–44. doi: 10.1016/j.fsigen.2016.07.004
- Edelmann, J., Kohl, M., Dressler, J., and Hoffmann, A. (2016). X-chromosomal 21-indel marker panel in German and Baltic populations. *Int. J. Legal. Med.* 130, 357–360. doi: 10.1007/s00414-015-1221-3
- Edelmann, J., Lessig, R., Willenberg, A., Wildgrube, R., Hering, S., and Szibor, R. (2006). Forensic validation of the X-chromosomal STR-markers GATA165B12, GATA164A09, DXS9908 and DXS7127 in German population. *International Congress Series* 1288, 298–300. doi: 10.1016/j.jics.2005.09.022
- Edwards, A., Civitello, A., Hammond, H. A., and Caskey, C. T. (1991). DNA Typing and Genetic Mapping With Trimeric and Tetrameric Tandem Repeats. *Am. J. Hum. Genet.* 49, 746–756.
- 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
- Egeland, T., Pinto, N., and Vigeland, M. D. (2014). A general approach to power calculation for relationship testing. *Forensic Sci. Int. Genet.* 9, 186–190. doi: 10.1016/j.fsigen.2013.05.001
- Elakkary, S., Hoffmeister-Ullrich, S., Schulze, C., Seif, E., Sheta, A., Hering, S., et al. (2014). Genetic polymorphisms of twelve X-STRs of the investigator Argus X-12 kit and additional six X-STR centromere region loci in an Egyptian population sample. *Forensic Sci Int Genet.* 11, 26–30. doi: 10.1016/j.fsigen.2014.02.007
- Fan, G., Ye, Y., Luo, H., and Hou, Y. (2015). Screening of Multi-InDel markers on X-chromosome for forensic purpose. *Forensic Science International: Genetics Supplement Series* 5, e42–e44. doi: 10.1016/j.fsigss.2015.09.017
- Fan, G. Y., Ye, Y., and Hou, Y. P. (2016). Detecting a hierarchical genetic population structure via Multi-InDel markers on the X chromosome. *Scientific Reports* 6, 32178.
- Ferragut, J. F., Pinto, N., Amorim, A., and Picornell, A. (2019). Improving publication quality and the importance of Post Publication Peer Review: The illustrating example of X chromosome analysis and calculation of forensic parameters. *Forensic Sci Int Genet.* 38, e5–e7. doi: 10.1016/j.fsigen.2018.11.006
- Flaquer, A., Fischer, C., and Wienker, T. F. (2009). A new sex-specific genetic map of the human pseudoautosomal regions (PAR1 and PAR2). *Hum Hered.* 68, 192–200. doi: 10.1159/000224639
- Freitas, N. S., Resque, R. L., Ribeiro-Rodrigues, E. M., Guerreiro, J. F., Santos, N. P., Ribeiro-Dos-Santos, A., et al. (2010). X-linked insertion/deletion polymorphisms: forensic applications of a 33-markers panel. *Int J Legal Med.* 124, 589–593. doi: 10.1007/s00414-010-0441-9
- Fukuta, M., Gaballah, M., Takada, K., Miyazaki, H., Kato, H., Aoki, Y., et al. (2019). Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population. *Legal Medicine* 37, 64–66. doi: 10.1016/j.legalmed.2019.01.009
- Gao, H., Wang, C., Zhang, R., Wu, H., Sun, S., Xiao, D., et al. (2019). Application of CPI cutoff value based on parentage testing of duos and trios typed by four autosomal kits. *PLoS ONE* 14:e0225174. doi: 10.1371/journal.pone.0225174
- García, M. G., Catanesi, C. I., Penacino, G. A., Gusmão, L., and Pinto, N. (2019). X-chromosome data for 12 STRs: Towards an Argentinian database of forensic haplotype frequencies. *Forensic Science International: Genetics* 41, e8–e13. doi: 10.1016/j.fsigen.2019.04.005
- Gettings, K. B., Aponte, R. A., Vallone, P. M., and Butler, J. M. (2015). Current knowledge and future issues. *Forensic Sci Int Genet* 18, 118–130. doi: 10.1016/j.fsigen.2015.06.005
- Gill, P., Brenner, C., Brinkmann, B., Budowle, B., Carracedo, A., Jobling, M. A., et al. (2001). DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. *Forensic Sci. Int.* 124, 5–10. doi: 10.1016/s0379-0738(01)00498-4
- Gjertson, D. W., Brenner, C. H., Baur, M. P., Carracedo, A., Guidet, F., Luque, J. A., et al. (2007). ISFG: recommendations on biostatistics in paternity testing. *Forensic Sci. Int. Genet* 1, 223–231. doi: 10.1016/j.fsigen.2007.06.006
- Gomes, C., Magalhães, M., Alves, C., Amorim, A., Pinto, N., and Gusmão, L. (2012). Comparative evaluation of alternative batteries of genetic markers to complement autosomal STRs in kinship investigations: autosomal indels vs. X-chromosome STRs. *Int. J. Legal Med.* 126, 917–921. doi: 10.1007/s00414-012-0768-5
- Gomes, I., Brehm, A., Gusmão, L., and Schneider, P. M. (2016). New Sequence Variants Detected at DXS10148, DXS10074 and DXS10134 Loci. *Forensic Sci Int Genet* 20, 112–116. doi: 10.1016/j.fsigen.2015.10.005
- Gomes, I., Pereira, P. J. P., Harms, S., Oliveira, A. M., Schneider, P. M., and Brehm, A. (2017). Genetic characterization of Guinea-Bissau using a 12 X-chromosomal STR system: Inferences from a multiethnic population. *Forensic Sci Int Genet.* 31, 89–94. doi: 10.1016/j.fsigen.2017.08.016
- Gomes, I., Pereira, R., Mayr, W. R., Amorim, A., Carracedo, A., and Gusmão, L. (2009). Evaluation of DXS9902, DXS7132, DXS6809, DXS7133, and DXS7423 in humans and chimpanzees: sequence variation, repeat structure, and nomenclature. *Int. J. Legal Med.* 123, 403–412. doi: 10.1007/s00414-009-0357-4
- Gomes, I., Prinz, M., Pereira, R., Bieschke, E., Amorim, A., Carracedo, A., et al. (2008). Sequence variation at three X chromosomal short tandem repeats in Caucasian and African populations. *Forensic Science International: Genetics Supplement Series* 1, 147–149. doi: 10.1016/j.fsigss.2007.10.097
- Gunter, C. (2005). Genome Biology: She Moves in Mysterious Ways. *Nature* 434, 279–280. doi: 10.1038/434279a
- Gusmão, L., Butler, J. M., Carracedo, A., Gill, P., Kayser, M., Mayr, W. R., et al. (2006). International Society of Forensic Genetics. DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Int J Legal Med* 120, 191–200. doi: 10.1007/s00414-005-0026-1
- Gusmão, L., Butler, J. M., Linacre, A., Parson, W., Roewer, L., Schneider, P. M., et al. (2017). Revised guidelines for the publication of genetic population data. *Forensic Sci Int Genet.* 30, 160–163. doi: 10.1016/j.fsigen.2017.06.007
- Gusmão, L., González-Neira, A., Alves, C., Lareu, M., Costa, S., Amorim, A., et al. (2002). Chimpanzee homologous of human Y specific STRs: A comparative study and a proposal for nomenclature. *Forensic Sci. Int.* 126, 129–136. doi: 10.1016/s0379-0738(02)00046-4
- Gusmão, L., Sánchez-Diz, P., Alves, C., Gomes, I., Zarrabeitia, M. T., Abovich, M., et al. (2008). A GEP-ISFG collaborative study on the optimization of an X-STR decaplex: data on 15 Iberian and Latin American populations. *Int. J. Legal Med.* 123, 227–234.
- Gusmão, L., Sánchez-Diz, P., Alves, C., Gomes, I., Zarrabeitia, M. T., Abovich, M., et al. (2009). A GEP-ISFG collaborative study on the optimization of an X-STR decaplex: data on 15 Iberian and Latin American populations. *Int. J. Legal Medicine* 123, 227–234. doi: 10.1007/s00414-008-0309-4
- Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* 8, 299–309.
- Hearne, C. M., and Todd, J. A. (1991). Tetranucleotide repeat polymorphism at the HPRT locus. *Nucleic Acids Res.* 19, 5450. doi: 10.1093/nar/19.19.5450
- Hedrick, P. W. (1987). Gametic disequilibrium measures: proceed with caution. *Genetics* 117, 331–341.
- Hwa, H. L., Chung, W. C., Chen, P. L., Lind, C. L., Lie, H. Y., Yin, H. I., et al. (2018). A 1204-single nucleotide polymorphism and insertion-deletion polymorphism panel for massively parallel sequencing analysis of DNA mixtures. *Forensic Sci Int Genet.* 32, 94–101. doi: 10.1016/j.fsigen.2017.11.002
- Hwa, H. L., Wu, M. Y., Lin, C. P., Hsieh, W. H., Yin, H. I., Lee, T. T., et al. (2019). A single nucleotide polymorphism panel for individual identification and ancestry assignment in Caucasians and four East and Southeast Asian populations using a machine learning classifier. *Forensic Sci Med Pathol.* 15, 67–74. doi: 10.1007/s12024-018-0071-y
- Jacquard, A. (1974). *The genetic structure of populations; in Biomathematics*, Vol. 5. Berlin: Springer-Verlag, 102–107.
- Jäger, A. C., Alvarez, M. L., Davis, C. P., Guzmán, E., Han, Y., Way, L., et al. (2017). Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. *Forensic Sci Int Genet.* 28, 52–70. doi: 10.1016/j.fsigen.2017.01.011
- Jin, B., Su, Q., Luo, H., Li, Y., Wu, J., Yan, J., et al. (2016). Mutational analysis of 33 autosomal short tandem repeat (STR) loci in southwest Chinese Han population based on trio parentage testing. *Forensic Science International: Genetics* 23, 86–90. doi: 10.1016/j.fsigen.2016.03.009

- Kauppi, L., Jasin, M., and Keeney, S. (2012). The tricky path to recombining X and Y chromosomes in meiosis. *Ann. N. Y. Acad. Sci.* 1267, 18–23. doi: 10.1111/j.1749-6632.2012.06593.x
- Kling, D., Dell'Amico, B., and Tillmar, A. O. (2015a). FamLinkX - implementation of a general model for likelihood computations for X-chromosomal marker data. *Forensic Sci. Int. Genet.* 17, 1–7. doi: 10.1016/j.fsigen.2015.02.007
- Kling, D., Tillmar, A., Egeland, T., and Mostad, P. (2015b). A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations. *Int. J. Legal Med.* 129, 943–954. doi: 10.1007/s00414-014-1117-7
- Kohn, M., Kehrner-Sawatzki, H., Vogel, W., Graves, J. A., and Hameister, H. (2004). Wide genome comparisons reveal the origins of the human X chromosome. *Trends Genet.* 20, 598–603. doi: 10.1016/j.tig.2004.09.008
- Kosambi, D. D. (1944). The estimation of map distances from recombination values. *Ann. Eugen.* 12, 172–175. doi: 10.1111/j.1469-1809.1943.tb02321.x
- Koyama, H., Iwasa, M., Tsuchimochi, T., Maeno, Y., Isobe, I., Seko-Nakamura, Y., et al. (2002). Y-STR haplotype data and allele frequency of the DXS10011 locus in a Japanese population sample. *Forensic Sci. Int.* 125, 273–276. doi: 10.1016/s0379-0738(01)00649-1
- Krawczak, M. (2007). Kinship testing with X-chromosomal markers: mathematical and statistical issues. *Forensic Sci. Int. Genet.* 1, 111–114. doi: 10.1016/j.fsigen.2007.01.014
- Lahn, B. T., and Page, D. C. (1999). Four evolutionary strata on the human X chromosome. *Science* (1999). 286, 964–967. doi: 10.1126/science.286.5441.964
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062
- Lazaruk, K., Wallin, J., Holt, C., Nguyen, T., and Walsh, P. S. (2001). Sequence variation in humans and other primates at six short tandem repeat loci used in forensic identity testing. *Forensic Sci. Int.* 119, 1–10. doi: 10.1016/s0379-0738(00)00388-1
- Levinson, G., and Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4, 203–221. doi: 10.1093/oxfordjournals.molbev.a040442
- Li, L., Li, C., Zhang, S., Zhao, S., Liu, Y., and Lin, Y. (2010). Analysis of 14 highly informative SNP markers on X chromosome by TaqMan SNP genotyping assay. *Forensic Sci. Int. Genet.* 4, e145–e148. doi: 10.1016/j.fsigen.2010.04.004
- Li, J., Deng, C., Luo, H., Song, F., and Hou, Y. (2017). *Analyzing an off Forensic Science International: Genetics Supplement Series* 6, e92–e93.
- Lin, L., Li, J., Hu, Y., Wang, H., Marah, F. A., Moseray, M., et al. (2020). Genetic characterization of 19 X-STRs in Sierra Leone population from Freetown. *Int. J. Legal Med.* 34, 1659–1661. doi: 10.1007/s00414-019-02243-6
- Liu, Q.-L., Wang, J.-Z., Quan, L., Zhao, H., Wu, Y.-D., Huang, X.-L., et al. (2013). Allele and Haplotype Diversity of 26 X-STR Loci in Four Nationality Populations from China. *PLoS ONE* 8:e65570. doi: 10.1371/journal.pone.0065570
- Lopes, A. M., Calafell, F., and Amorim, A. (2004). Microsatellite variation and evolutionary history of PCDHX/Y gene pair within the Xq21.3/Yp11.2 hominid-specific homology block. *Mol. Biol. Evol.* 21, 2092–2101. doi: 10.1093/molbev/msh218
- Mansfield, E. S., Blasband, A., Kronick, M. N., Wrabetz, L., Kaplan, P., Rappaport, E., et al. (1993). Fluorescent approaches to diagnosis of Lesch-Nyhan syndrome and quantitative analysis of carrier status. *Mol. Cell. Probes* 7, 311–324. doi: 10.1006/mcpr.1993.1045
- Matveevsky, S., Kolomiets, O., Bogdanov, A., Hakhverdyan, M., and Bakloushinskaya, I. (2017). Chromosomal Evolution in Mole Voles *Ellobius* (Cricetidae, Rodentia): Bizarre Sex Chromosomes, Variable Autosomes and Meiosis. *Genes (Basel)* 8, 306. doi: 10.3390/genes8110306
- Medina-Acosta, E. (2011). Evidence of partial and weak gametic disequilibrium across clusters of pericentromeric short tandem repeats loci on human X chromosome: proceed with caution in forensic genetics. *Forensic Sci. Int. Genet.* 5, 545–547. doi: 10.1016/j.fsigen.2009.12.002
- Nachman, M. W., and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297–304.
- National Institute of Standards and Technology [Nist]. (2020). *Standard Reference Database*. Available at: <https://strbase.nist.gov/> (accessed May 10, 2020).
- Nielsen, J., and Wohrlert, M. (1990). Sex chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Aarhus. *Denmark. Birth Defects Original Article Series* 26, 209–223. doi: 10.1007/BF01213097
- Nishi, T., Fukui, K., and Iwadata, K. (2020). Genetic Polymorphism Analyses of Three Novel X Chromosomal Short Tandem Repeat Loci in the Xp22.3 Region. *Legal Medicine* 45, 101709. doi: 10.1016/j.legalmed.2020.101709
- Nothnagel, M., Szibor, R., Vollrath, O., Augustin, C., Edelmann, J., Geppert, M., et al. (2012). Collaborative genetic mapping of 12 forensic short tandem repeat (STR) loci on the human X chromosome. *Forensic Sci. Int. Genet.* 6, 778–784. doi: 10.1016/j.fsigen.2012.02.015
- Ohta, T., and Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* 22, 201–204. doi: 10.1073/pnas.75.6.2868
- Oki, T., Hayashi, T., Ota, M., and Asamura, H. (2012). Development of multiplex assay with 16 SNPs on X chromosome for degraded samples. *Leg Med (Tokyo)* 14, 11–16. doi: 10.1016/j.legalmed.2011.10.001
- Olaisen, B., Bär, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., et al. (1998). DNA recommendations 1997 of the International Society for Forensic Genetics. *Vox Sang* 74, 61–63.
- Otto, S., Pannell, J. R., Peichel, C. L., Ashman, T. L., Charlesworth, D., Chippindale, A. K., et al. (2011). About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet.* 27, 358–367. doi: 10.1016/j.tig.2011.05.001
- Parson, W., and Roewer, L. (2010). Publication of Population Data of Linearly Inherited DNA Markers in the International Journal of Legal Medicine. *Int J Legal Med.* 124, 505–509. doi: 10.1007/s00414-010-0492-y
- Pereira, R., Alves, C., Aler, M., Amorim, A., Arévalo, C., Betancor, E., et al. (2018). A GHEP-ISFG collaborative study on the genetic variation of 38 autosomal indels for human identification in different continental populations. *Forensic Sci. Int. Genet.* 32, 18–25. doi: 10.1016/j.fsigen.2017.09.012
- Pereira, R., Gomes, I., Amorim, A., and Gusmão, L. (2007). Genetic diversity of 10 X chromosome STRs in northern Portugal. *Int. J. Legal Med.* 121, 192–197. doi: 10.1007/s00414-006-0144-4
- Pereira, R., Pereira, V., Gomes, I., Tomas, C., Morling, N., Amorim, A., et al. (2012). A method for the analysis of 32 X chromosome insertion deletion polymorphisms in a single PCR. *Int J Legal Med.* 126, 97–105. doi: 10.1007/s00414-011-0593-2
- Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, A., and Gusmão, L. (2009). A new multiplex for human identification using insertion/deletion polymorphisms. *Electrophoresis* 30, 3682–3690. doi: 10.1002/elps.200900274
- Pereira, V., and Gusmão, L. (2016). “Types of Genomes, Sequences and Genetic Markers (Repeats, SNPs, Indels, Haplotypes),” in *Handbook Of Forensic Genetics: Biodiversity And Heredity In Civil And Criminal Investigation*, Vol. 2, (Singapore: World Scientific), 163. doi: 10.1142/9781786340788_0009
- Petkovski, E., Keyser-Tracqui, C., Hienne, R., and Ludes, B. (2005). SNPs and MALDI-TOF MS: tools for DNA typing in forensic paternity testing and anthropology. *J. Forensic Sci.* 50, 535–541.
- Phillips, C., Ballard, D., Gill, P., Syndercombe Court, D., Carracedo, A., and Lareu, M. V. (2012). The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data. *Forensic Science International: Genetics* 6, 354–365. doi: 10.1016/j.fsigen.2011.07.012
- Phillips, C., Devesse, L., Ballard, D., van Weert, L., de la Puente, M., and Melis, S. (2018). Global patterns of STR sequence variation: Sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit. *Electrophoresis* 39, 2708–2724. doi: 10.1002/elps.201800117
- Pinto, N., Gusmão, L., and Amorim, A. (2011a). X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial. *Forensic Sci. Int. Genet.* 5, 27–32. doi: 10.1016/j.fsigen.2010.01.011
- Pinto, N., Gusmão, L., Silva, P. V., and Amorim, A. (2011b). Estimating coancestry from genotypes using a linear regression method. *Forensic Science International: Genetics Supplement Series* 3, e373–e374. doi: 10.1016/j.fsigs.2011.09.048

- Pinto, N., Gusmão, L., Egeland, L. T., and Amorim, A. (2013a). Paternity exclusion power: comparative behaviour of autosomal and X-chromosomal markers in standard and deficient cases with inbreeding. *Forensic Sci. Int. Genet.* 7, 290–295. doi: 10.1016/j.fsigen.2012.12.002
- Pinto, N., Gusmão, L., Egeland, T., and Amorim, A. (2013b). Estimating relatedness with no prior specification of any genealogy: The role of the X-chromosome. *Forensic Science International: Genetics Supplement Series* 4, e252–e253. doi: 10.1016/j.fsigs.2013.10.129
- Pinto, N., Silva, P. V., and Amorim, A. (2010). General Derivation of the Sets of Pedigrees with the Same Kinship Coefficients. *Hum. Hered.* 70, 194–204. doi: 10.1159/000316390
- Pinto, N., Silva, P. V., and Amorim, A. (2012). A general method to assess the utility of the X-chromosomal markers in kinship testing. *Forensic Science International: Genetics* 6, 198–207. doi: 10.1016/j.fsigen.2011.04.014
- Prieto-Fernández, E., Baeta, M., Núñez, C., Zarrabeitia, M. T., and Herrera, R. J. (2016). Development of a new highly efficient 17 X-STR multiplex for forensic purposes. *Electrophoresis* 37, 1651–1658. doi: 10.1002/elps.201500546
- Primmer, C. R., Saino, N., Möller, A. P., and Ellegren, H. (1996). Directional evolution in germline microsatellite mutations. *Nat. Genet.* 13, 391. doi: 10.1038/ng0896-391
- Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., et al. (2005). The DNA sequence of the human X chromosome. *Nature* 434, 325–337. doi: 10.1038/nature03440
- Schaffner, S. F. (2004). The X chromosome in population genetics. *Nat. Rev. Genet.* 5, 43–51. doi: 10.1038/nrg1247
- Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma* 109, 365–371. doi: 10.1007/s004120000089
- Schlötterer, C., and Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* 20, 211–215. doi: 10.1093/nar/20.2.211
- Shimmin, L. C., Chang, B. H., and Li, W. H. (1993). Male-driven evolution of DNA sequences. *Nature* 362, 745–747. doi: 10.1038/362745a0
- Sleddens, H. F., Oostra, B. A., Brinkmann, A. O., and Trapman, J. (1992). Trinucleotide repeat polymorphism in the androgen receptor gene (AR). *Nucleic Acids Res.* 20, 1427. doi: 10.1093/nar/20.6.1427-a
- Stepanov, V., Vagaitseva, K., Kharkov, V., Cherednichenko, A., Bocharova, A., Berezina, G., et al. (2016). Forensic and population genetic characteristics of 62 X chromosome SNPs revealed by multiplex PCR and MALDI-TOF mass spectrometry genotyping in 4 North Eurasian populations. *Leg Med (Tokyo)* 18, 66–71. doi: 10.1016/j.legalmed.2015.12.008
- Strand, M., Prolla, T. A., Liskay, R. M., and Petes, T. D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* 365, 274. doi: 10.1038/365274a0
- STRidER. (2020). (STRs for Identity ENFSI Reference Database). Available at: <https://strider.online/> (accessed May 10, 2020).
- 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
- Szibor, R., Edelmann, J., Hering, S., Gomes, I., and Gusmão, L. (2009). Nomenclature Discrepancies in the HPRTB Short Tandem Repeat. *Int. J. Legal Med.* 123, 185–186. doi: 10.1007/s00414-008-0314-7
- Szibor, R., Hering, S., and Edelmann, J. (2005). The HumARA genotype is linked to spinal and bulbar muscular dystrophy and some further disease risks and should no longer be used as a DNA marker for forensic purposes. *Int. J. Legal Med.* 119, 179–180. doi: 10.1007/s00414-005-0525-0
- Szibor, R., Hering, S., and Edelmann, J. (2006). A new Web site compiling forensic chromosome X research is now online. *Int J Legal Med.* 120, 252–254. doi: 10.1007/s00414-005-0029-y
- Szibor, R., Krawczak, M., Hering, S., Edelmann, J., Kuhlich, E., and Krause, D. (2003). Use of X-linked markers for forensic purposes. *Int. J. Legal Med.* 117, 67–74. doi: 10.1007/s00414-002-0352-5
- Tao, R., Zhang, J., Sheng, X., Zhang, J., Yang, Z., Chen, C., et al. (2019). Development and validation of a multiplex insertion/deletion marker panel, SifaInDel 45plex system. *Forensic Science International: Genetics* 41, 128–136. doi: 10.1016/j.fsigen.2019.04.008
- The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–861. doi: 10.1038/nature06258
- The International Society for Forensic Genetics [ISFG]. (2020). Available at: www.isfg.org (accessed May 10, 2020).
- Tillmar, A. O., Egeland, T., Lindblom, B., Holmlund, G., and Mostad, P. (2011). Using X-chromosomal markers in relationship testing: calculation of likelihood ratios taking both linkage and linkage disequilibrium into account. *Forensic Sci. Int. Genet.* 5, 506–511. doi: 10.1016/j.fsigen.2010.11.004
- Tillmar, A. O., Kling, D., Butler, J. M., Parson, W., Prinz, M., Schneider, P. M., et al. (2017). DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis. *Forensic Sci Int Genet.* 29, 269–275. doi: 10.1016/j.fsigen.2017.05.005
- Tomas, C., Pereira, V., and Morling, N. (2012). Analysis of 12 X-STRs in Greenlanders, Danes and Somalis using Argus X-12. *Int J Legal Med.* 126, 121–128. doi: 10.1007/s00414-011-0609-y
- Tomas, C., Sanchez, J. J., Castro, J. A., Børsting, C., and Morling, N. (2010). Forensic usefulness of a 25 X-chromosome single-nucleotide polymorphism marker set. *Transfusion.* 50, 2258–2265. doi: 10.1111/j.1537-2995.2010.02696.x
- Watanabe, G., Umetsu, K., Yuasa, I., and Suzuki, T. (2000). DXS10011: a hypervariable tetranucleotide STR polymorphism on the X chromosome. *Int J Legal Med.* 113, 249–250. doi: 10.1007/s004149900096
- Weir, B., Anderson, A., and Hepler, A. (2006). Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* 7, 771–780. doi: 10.1038/nrg1960
- Xu, X., Peng, M., and Fang, Z. (2000). The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* 24, 396. doi: 10.1038/74238
- YHRD. (2020). Y-STR Haplotype Reference Database (YHRD). Available at: <https://yhrd.org/YHRD> (accessed May 10, 2020).
- Zarrabeitia, M. T., Mijares, V., and Riancho, J. A. (2007). Forensic efficiency of microsatellites and single nucleotide polymorphisms on the X chromosome. *Int J Legal Med.* 121, 433–437. doi: 10.1007/s00414-007-0169-3
- Zaumsegel, D., Rothschild, M. A., and Schneider, P. M. (2013). A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry. *Forensic Sci Int Genet.* 7, 305–312. doi: 10.1016/j.fsigen.2012.12.007
- Zhang, S., Bian, Y., Chen, A., Zheng, H., Gao, Y., Hou, Y., et al. (2017a). Developmental validation of a custom panel including 273 SNPs for forensic application using Ion Torrent PGM. *Forensic Sci Int Genet.* 27, 50–57. doi: 10.1016/j.fsigen.2016.12.003
- Zhang, S., Lin, Y., Bian, Y., and Li, C. (2017b). Parallel sequencing of 60 X-chromosome genetic markers including STRs, SNPs and InDels. *Forensic Science International: Genetics Supplement Series* 6, e317–e319. doi: 10.1016/j.fsigs.2017.09.127

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

Copyright © 2020 Gomes, Pinto, Antão-Sousa, Gomes, Gusmão and Amorim. 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 Validation of a Novel Five-Dye Short Tandem Repeat Panel for Forensic Identification of 11 Species

Wei Cui^{1,2,3}, Xiaoye Jin^{1,2,3}, Yuxin Guo^{1,2,3}, Chong Chen^{1,2,3}, Wenqing Zhang^{1,2}, Yijie Wang^{1,2}, Jiangwei Lan⁴ and Bofeng Zhu^{1,2,4*}

¹ Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, China, ² Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases, College of Stomatology, Xi'an Jiaotong University, Xi'an, China, ³ College of Medicine and Forensics, Xi'an Jiaotong University Health Science Center, Xi'an, China, ⁴ Multi-Omics Innovative Research Center of Forensic Identification, Department of Forensic Genetics, School of Forensic Medicine, Southern Medical University, Guangzhou, China

OPEN ACCESS

Edited by:

Cemal Gurkan,
Turkish Cypriot DNA Laboratory
(TCDL), Cyprus

Reviewed by:

Christopher Phillips,
University of Santiago
de Compostela, Spain
Guanglin He,
Sichuan 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: 24 March 2020

Accepted: 06 August 2020

Published: 24 September 2020

Citation:

Cui W, Jin X, Guo Y, Chen C,
Zhang W, Wang Y, Lan J and Zhu B
(2020) Development and Validation
of a Novel Five-Dye Short Tandem
Repeat Panel for Forensic
Identification of 11 Species.
Front. Genet. 11:1005.
doi: 10.3389/fgene.2020.01005

Species identification of unknown biological samples is of fundamental importance for forensic applications, especially in crime detection, poaching, and illegal trade of endangered animals as well as meat fraud. In this study, a novel panel was developed to simultaneously identify 10 different animal species (*Gallus domesticus*, *Anas platyrhynchos domesticus*, *Ovis aries*, *Sus scrofa domesticus*, *Bos taurus*, *Equus caballus*, *Columba livia domestica*, *Rattus norvegicus*, *Mus musculus*, and *Canis lupus familiaris*) and human beings by amplifying 22 short tandem repeat (STR) loci in a multiplex PCR using a set of five fluorescently labeled dyes. This novel 22-STR panel was validated by optimization of PCR conditions as well as species specificity, sensitivity, reproducibility, precision, DNA mixture, and tissue/organ consistency. The results of developmental validation showed that the 22-STR loci achieved high species specificity among 10 animal species and human beings, and the sensitivity of this panel was 0.09 ng. This 22-STR panel identified different meats in mixed samples, and the minimum detected mixture ratio in the current test was 10% (0.1 ng/1 ng). This sensitive, accurate, and specific 22-STR panel can be used for forensic species identification and the detection of meat fraud and adulteration.

Keywords: species identification, meat fraud, developmental validation, forensic science, short tandem repeat

INTRODUCTION

Biological samples left behind at crime scenes always contain a great deal of valuable information that can provide helpful clues for the criminal investigations. In addition to human biological specimens, non-human biological samples acquired from a crime scene can also suggest certain directions for tracing the suspects. With the continuous progress of biotechnological achievements made in the field of forensic genetics, there has been much interest in the forensic identification of non-human species. For example, domestic pet hairs left at a crime scene could be evidence indicating that the suspect (pet owner) might have been present at the scene of the crime (Budowle et al., 2005).

Rhinoceros horn and tiger bone were two substances that were formerly used in traditional Chinese medicine. Although the production and import of these protective animal-derived traditional Chinese medicine are strictly prohibited in China, occasional illegal trade occurs. Species identification is then of great importance in the criminal investigation of poaching and illegal trade of endangered animals (Staats et al., 2016). Moreover, species identification of various animal utilizing genetic markers can be applied in the detection of species mislabeling, and the process also contributes to food safety by its utilization in meat adulteration cases (Iyengar, 2014).

With the growing size of the human population and increasing of social affluence in recent years, meat consumption has been increasing annually (Godfray et al., 2018). Along with the increase in meat consumption of different animals, meat fraud and adulteration events occasionally occur around the world. An example of this was the spread of a horsemeat scandal across Europe in 2013 (O'Mahony, 2013), which not only seriously undermined the market order but also increased the risk of religious and ethnic conflicts. Because there are now widespread meat fraud and adulteration, some effective measures should be taken to ensure the authenticity of meat products. Developing accurate meat identification techniques will play an important role in solving these problems mentioned above.

Based on the morphological and structural differences of cells and tissues in different species, the morphological observation was one of the most traditional techniques used for species identification a few decades ago (Lou et al., 2016). The serologic-based technique has been used for species identification since the 1980s. With relatively high accuracy and sensitivity compared with traditional morphology, serological methods, such as the colloidal gold test strip, have been used in species confirmatory tests (Matsuzawa et al., 1993). However, serology analysis is vulnerable to low antibody specificity or trace sample, as well as the ability to only distinguish between human and non-human specimens.

To date, the DNA-based species identification technique has been widely adopted as an effective molecular detection tool for the identification of non-human species due to the rapid development of polymerase chain reaction (PCR) and the high level of sensitivity achieved in recent years (Skouridou et al., 2019). DNA barcodes refer to a short DNA sequence from a standard locus which not only encompasses sufficient phylogenetic information to identify different species but is also easy to amplify and analyze (Nicolas et al., 2012). Currently, the most popular DNA barcodes are cytochrome *b* and cytochrome oxidase subunit 1, which are located on mitochondrial genome. Although there have been more than five million DNA barcodes published in the Barcode of Life Data System (BOLD)¹, scientists found a severe lack of adequate taxonomic coverage of some animal species within BOLD, which might be the results of anomalous or invalid identification (Wilson-Wilde et al., 2010; Iyengar, 2014). However, an underestimation or overestimation of species DNA content might be a concluded when mixed samples were analyzed using mitochondrial markers based on

real-time PCR (RT-PCR) or digital PCR (dPCR) due to the mitochondrial heterogeneity in different tissue or organs (Floren et al., 2015). Hence, DNA genetic markers located in the nuclear genome have increasingly become the promising molecular markers for animal species identification.

Appearing as a repeat unit of a 2–6 bp core sequence, the short tandem repeat (STR) loci, distributing widely in the human genome with high polymorphisms, have been widely used in the field of forensic genetics (Fordyce et al., 2015; Fang et al., 2018; Wang et al., 2018). In recent decades, the PCR-STR capillary electrophoresis (CE)-based technique has already matured and has been used as the gold standard method for individual identifications and kinship tests (Zhang et al., 2018).

In this study, we selected 22-STR loci of 10 different animals as well as human beings, with two STR loci for each species, which enabled high species specificity among pig, cattle, goat, chicken, duck, rat, mouse, horse, pigeon, canine, and human samples. And then, we constructed a novel five-dye typing panel based on the CE platform. To evaluate the forensic efficiency of this 22-STR panel, we conducted a series of validation tests such as sensitivity, species specificity, DNA mixture, reproducibility studies and so on.

MATERIALS AND METHODS

Sample Collections and DNA Extraction

Samples of chicken (*Gallus domesticus*), duck (*Anas platyrhynchos domesticus*), sheep (*Ovis aries*), pig (*Sus scrofa domesticus*), cattle (*Bos taurus*), horse (*Equus caballus*), and pigeon (*Columba livia domestica*) were purchased in the local market. Samples of Sprague-Dawley (SD) rat (*Rattus norvegicus*), Kunming mouse (*Mus musculus*), and dog (*Canis lupus familiaris*) were acquired from the Medical Experimental Animal Center of Xi'an Jiaotong University. The environment of this center meets the standard for feeding practices conducted for experimental animals (GB 14925-2010). Human (*Homo sapiens*) blood stains were previously collected by our laboratory. SD rats and Kunming mice were anesthetized and euthanized by cervical dislocation. Euthanasia of the experimental animals, sample collections, and the following experimental processes were approved by the ethics committee of Xi'an Jiaotong University, Health Science Center.

DNA was extracted using a TIANamp Genomic DNA Kit (TIANGEN Biotech, Beijing, China). DNA was quantified with a NanoDrop 2000c spectrophotometer (Thermo Fisher Scientific, South San Francisco, CA, United States). If the concentration of extracted DNA was not greater than 1 ng/μl, the DNA was re-extracted.

Selection of Species-Specific STR Loci and Primer Design

Twenty-two STR loci showing species specificity among 11 species were selected from published studies following the criteria: (1) primer sequences designed for each STR locus of one species did not share homologous sequences with other species; (2) priority was given to STR loci whose core sequences

¹<http://www.barcodinglife.org/>

were tetranucleotide; and (3) priority was given to STR loci that had fewer alleles. Primer 5.0 software was used to design the STR primers. Oligo 7 software (Rychlik, 2007) was used to ensure that each primer was free from self-dimer and non-specific hybridization in other species genomic regions using Basic Local Alignment Search Tool (BLAST) provided by National Center for Biotechnology Information. Primers of 22 STR loci were commercially synthesized (Microread Genetics, Beijing, China). Four dyes were used to individually label these primers, and QD550 (orange, Microread Genetics, Beijing, China) was used to mark the internal size standard. Detailed information for each STR locus is shown in **Table 1**.

Allelic Ladder Construction

For the 22-STR loci, 20 unrelated individuals of each species were collected to determine the variabilities of the alleles observed in each species. Moreover, variabilities of the alleles for some STR markers were screened from previously published studies. Allelic ladder was generated according to previous reports (Chen et al., 2019).

Multiplex Amplification and Genotyping

Unless stated otherwise, standard PCR amplification and genotyping procedures were as follows. We used a 10- μ l reaction volume containing 1 μ l of DNA template (1 ng/ μ l), 2 μ l of Primer set (Microread Genetics, Beijing, China), 4 μ l of Master Mix I (Microread Genetics, Beijing, China), and 3 μ l of deionized water. The PCR was conducted using a GeneAmp PCR System 9700 Thermal Cycler (Thermo Fisher Scientific,

South San Francisco, CA, United States) under the following conditions: 5 min of initial denaturation at 95°C, followed by 29 cycles of 94°C for 30 s, 59°C for 60 s, and 72°C for 60 s, with the final elongation at 60°C for 60 min. Electrophoresis was performed by an ABI 3500xL Genetic Analyzer (Thermo Fisher Scientific, South San Francisco, CA) using 36-cm capillary arrays with POP-4® Polymer (Thermo Fisher Scientific, South San Francisco, CA). Loading samples for CE contained 1 μ l of PCR product, 0.3 μ l of QD550 internal size standard, and 8.7 μ l of Hi-Di Formamide. The alleles were genotyped using GeneMapper ID-X software v1.5 (Thermo Fisher Scientific, South San Francisco, CA, United States). Next, equal amounts of DNA from each species were mixed, and then the mixture was diluted to 1 ng/ μ l (DNA mix). The DNA mix was used as positive control DNA, and deionized water was used as the negative control.

Construction of the Multiple Amplification STR Panel

Amplification of Each STR Locus

To evaluate the specificity and amplification efficiency of a pair of primers, we amplified each STR locus according to the standard PCR components and reaction conditions.

Optimization Study of PCR Conditions

PCR cycling parameter studies were conducted on the 10- μ l volume system. Total cycle numbers of 27, 28, 29, 30, and 31 cycles, and annealing temperatures at 57, 58, 59, 60, and 61°C

TABLE 1 | Detail information of 22 STR loci in this novel panel for species identification.

Species	Locus	Chromosome	Accession	Repeat unit	Dye
Pig	EF046	13	NC010455.5	GT	HEX
	SW742	16	AF235351.1	GT	FAM
Cattle	BT165*	26	FJ232025	TATG	FAM
	BT150*	22	FJ232024	ATAC	TAMRA
Sheep	MAF33	OAR9/CHI14	M77200	CA	TAMRA
	MCM164	OAR2/CHI8	L39134	GT	HEX
Chicken	LEI0094	1	X83246.1	AC	HEX
	GCT025	2	AJ233970.1	(GAAA) _m (GAAG) _n (AAAG) _o	FAM
Duck	APH14	Unknown	AJ272583.1	(CA) _m A(CA) _n	HEX
	CAUD056	Unknown	AY493301.1	TTTCCCTCTTTC	FAM
SD rat	D0UIA21	Unknown	AF053391	GATA	TAMRA
	D8UIA2	8	AF054019	GATA	TAMRA
Kunming mouse	NC000084	18	NC000084	TAGA	HEX
	NC000070	4	NC000070	GATA	HEX
Horse	HMS3	9	X74632.1	(TG) ₂ (CA) ₂ TC(CA) _n Or (TG) ₂ (CA) ₂ TC(CA) _n GA(CA) ₅	TAMRA
	HMS6	4	X74635.1	GT	HEX
Pigeon	PG5	Unknown	Ref [#]	TTTG	FAM
	PG6	Unknown	Ref [#]	AAAC	FAM
Canine	FH2100	3	NC_006585.3	GAAT	ROX
	FH2361	29	FJ031001.1	GAAA	ROX
Human	D3S3045	3	NC_000003.12	GATA	TAMRA
	TPOX	2	M68651	AATG	ROX

*BT150 has changed to BT22, and BT165 has changed to BT26 in GenBank. [#]These two STR loci were selected from Chun-Lee et al. (2007).

were separately tested in order to choose the most optimal PCR parameters.

Studies of the Total Volume of the PCR System and Uniformity of the PCR Amplification

Two different total reaction volumes, 10 and 25 μ l, were adjusted to evaluate the performance of this 22-STR panel. The value of each reagent in the 25- μ l PCR system was 2.5 times larger than that in the 10- μ l system, and the PCR conditions were in accordance with those mentioned above. To evaluate the PCR performance in various PCR thermocyclers, we conducted the PCR in the Applied Biosystems Veriti Thermal Cycler, Applied Biosystems ProFlex Thermal Cycler, Applied Biosystems 9700 Thermal Cycler, and Applied Biosystems 2720 Thermal Cycler (Thermo Fisher Scientific, South San Francisco, CA). The DNA mix was regarded as the DNA template, and the reaction conditions used were as mentioned above.

Developmental Validation Studies of This 22-STR Panel

Sensitivity, Reproducibility, and Precision Study

We performed a serial of concentrations of DNA mix to obtain 5 ng/ μ l, 2 ng/ μ l, 1 ng/ μ l, 0.5 ng/ μ l, 0.25 ng/ μ l, 125 pg/ μ l, and 62.5 pg/ μ l that were used to evaluate the sensitivity of this 22-STR panel. To evaluate the reproducibility of this panel, the DNA mix was genotyped three different times, comparing with the ladder profile. The mean sizes in base pairs and the standard deviations were calculated for each allele. We selected some samples of different species and sequenced the alleles of each STR locus using the Sanger sequencing method in order to verify the STR profile results of the CE platform. Sanger sequencing was conducted by Sangon Biotech® Company (Sangon Biotech® Co., Ltd, Shanghai, China). Genomic DNA was extracted from the liver, heart, spleen, lung, kidney, and muscle of the same SD rat and Kunming mouse, respectively, which was used to evaluate the concordance of genotyping results of this panel in detecting different organs or tissue of the same individual.

Specificity and Mixture Study

We genotyped DNA of each studied species based on this multiple STR panel to evaluate if the panel was capable of avoiding the genotype of other non-targeted species. Samples from practical cases were usually composed of more than one animal species, therefore, it was important to evaluate the reliability of this panel for the detection of mixture samples of different species. Mixture studies were also performed to evaluate the lowest detection limit (minimum amount of DNA/total amount of DNA) of this 22-STR panel. For this purpose, we set two types of DNA mixture patterns with different mix ratios, and these two mixture patterns are shown in **Supplementary Table S1**.

Casework Sample Verification

Human blood stains preserved on an FTA™ card at room temperature for up to 7 years were used to represent common samples in the practical cases. We selected four unknown cooked meats to evaluate the efficiency of the detection of cooked meats.

These four meats were named SE (braised), SY (roasted), SG (poached), and SN (stir-fried). The surface of all samples was rinsed with ultrapure water before DNA extraction to remove all inhibitory substances. All samples were amplified in a 10- μ l volume, and the PCR conditions were the same as those mentioned above.

RESULTS

Construction of This 22-STR Panel

In this research, 22-STR loci with high-species specificity among 10 animal species and human were selected: D8UIA2, D0UIA21; PG5, PG6; FH2361, FH2100; GCT025, LEI0049; HMS6, HMS3; BT165, BT150; D3S3045, TPOX; NC000070, NC000084; APH14, CAUD056; MAF33, MCM164; SW742 and EF046 loci.

Before the construction of this novel panel, we amplified each STR locus to evaluate the specificity and amplification efficiency of a pair of primers, and the results showed that specific peaks of each pair of primers were only detected at the corresponding locus for each species, and no peak was found in other species.

A series of temperatures at 57, 58, 59, 60, and 61°C were used to determine the optimal primer annealing temperature. All loci could be detected at these five different annealing temperatures, and the amplification efficiencies at 58, 59, and 60°C were higher than those at 57 and 61°C. The highest value of average peak height was observed when the annealing temperature was 58°C, but more optimal peak height balance was found at 59°C. We finally chose 59°C as the most optimal annealing temperature. Related genotype profiles are shown in **Supplementary Figure S1**.

This 22-STR panel was tested over a range of 27, 28, 29, 30, and 31 total amplification cycles. With the increase in the number of cycles, the peak height increased obviously. All alleles could be detected, and more optimal allelic peak height balance was obtained when the number of amplification cycles was 29. We finally chose 29 as the optimal number of amplification cycles. Related genotype profiles are shown in **Figure 1**.

We also evaluated the PCR efficiencies in the different reaction volume systems of 10 and 25 μ l. As shown in **Supplementary Figure S2**, there was no allele drop in two systems, and higher peak heights were obtained in the 10- μ l volume system as compared to the 25- μ l volume system. We conducted the PCR using four different types of PCR machines, and the results showed that there was no obvious deviation of peak height balance among these four machines. The relative genotype profiles are shown in **Supplementary Figure S3**.

Sensitivity, Reproducibility, Precision, and Concordance Study

We used a serial of input DNA concentrations of 5 ng/ μ l, 2 ng/ μ l, 1 ng/ μ l, 0.5 ng/ μ l, 0.25 ng/ μ l, 125 pg/ μ l, and 62.5 pg/ μ l of DNA mix to evaluate the sensitivity of this 22-STR panel. As shown in **Figure 2**, small numbers of allele peaks dropped out when the template amount was 0.5 ng/ μ l, and many allele peaks dropped out when the template amounts were 125 or 62.5 pg/ μ l. Because the DNA mix contained equal amounts of DNA, the

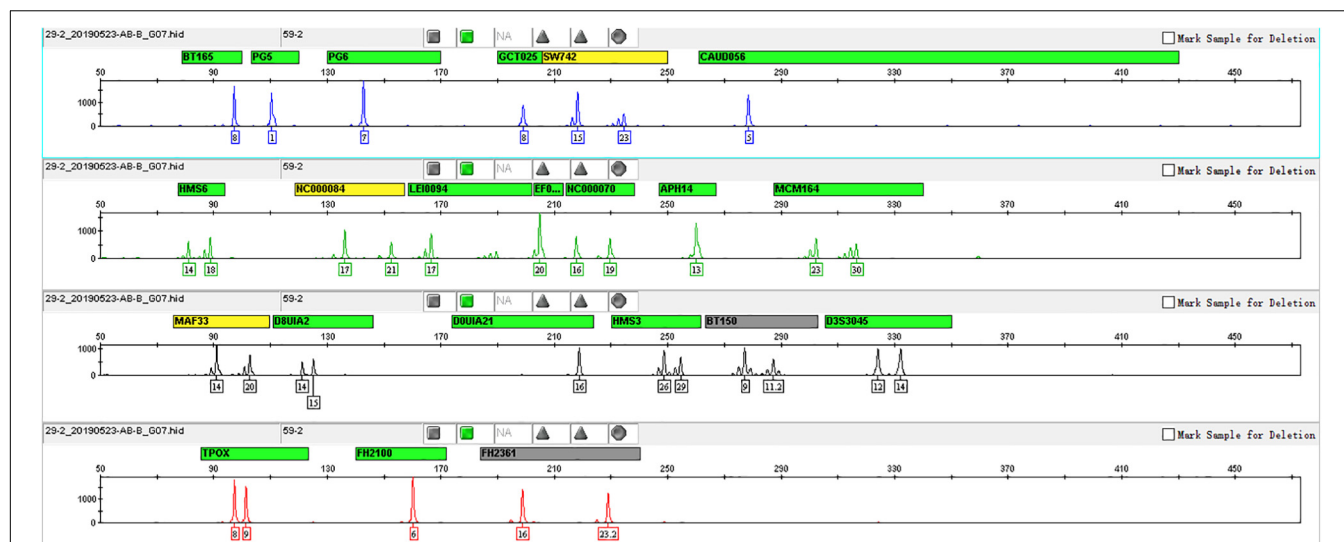


FIGURE 1 | Genotyping profile of DNA mix amplified cycle numbers at 29 cycles. A more optimal peak height balance was observed with 29 cycles, and thus, we finally chose 29 as the optimal number of cycles.

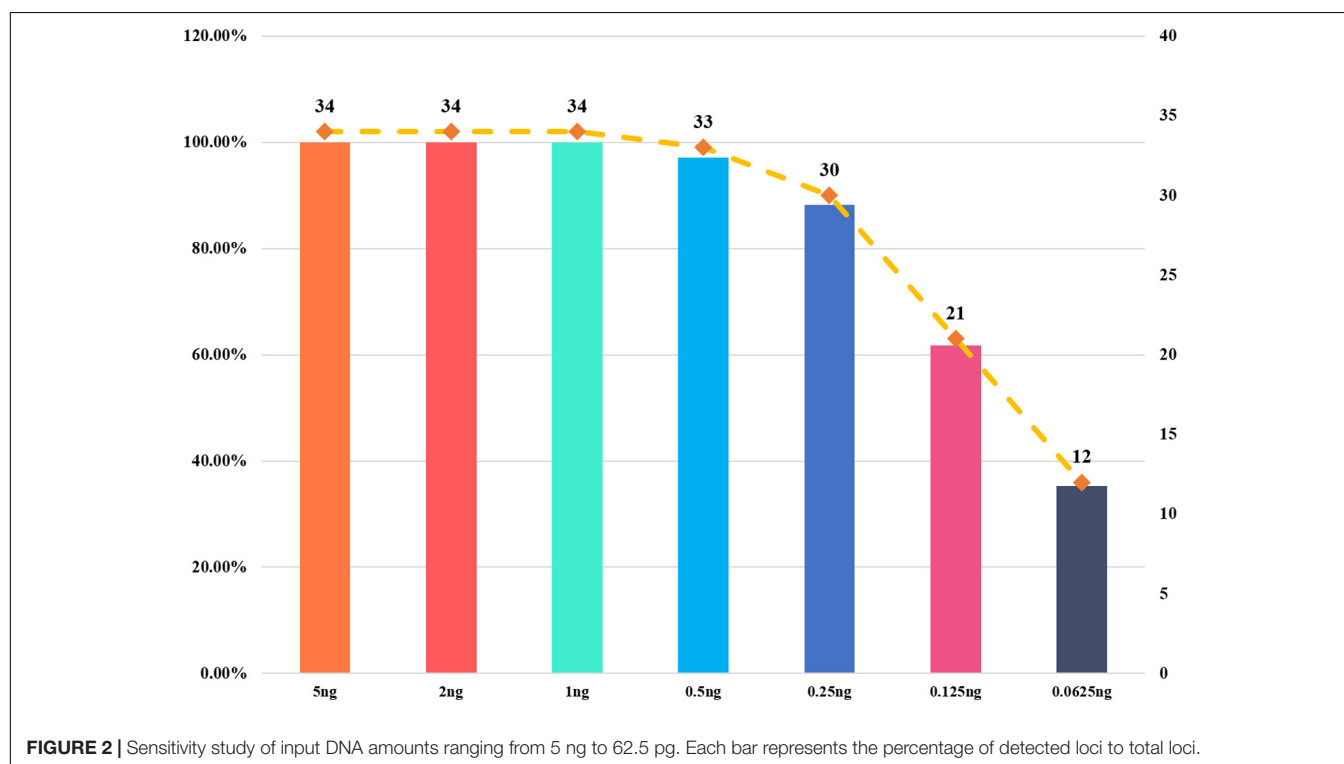
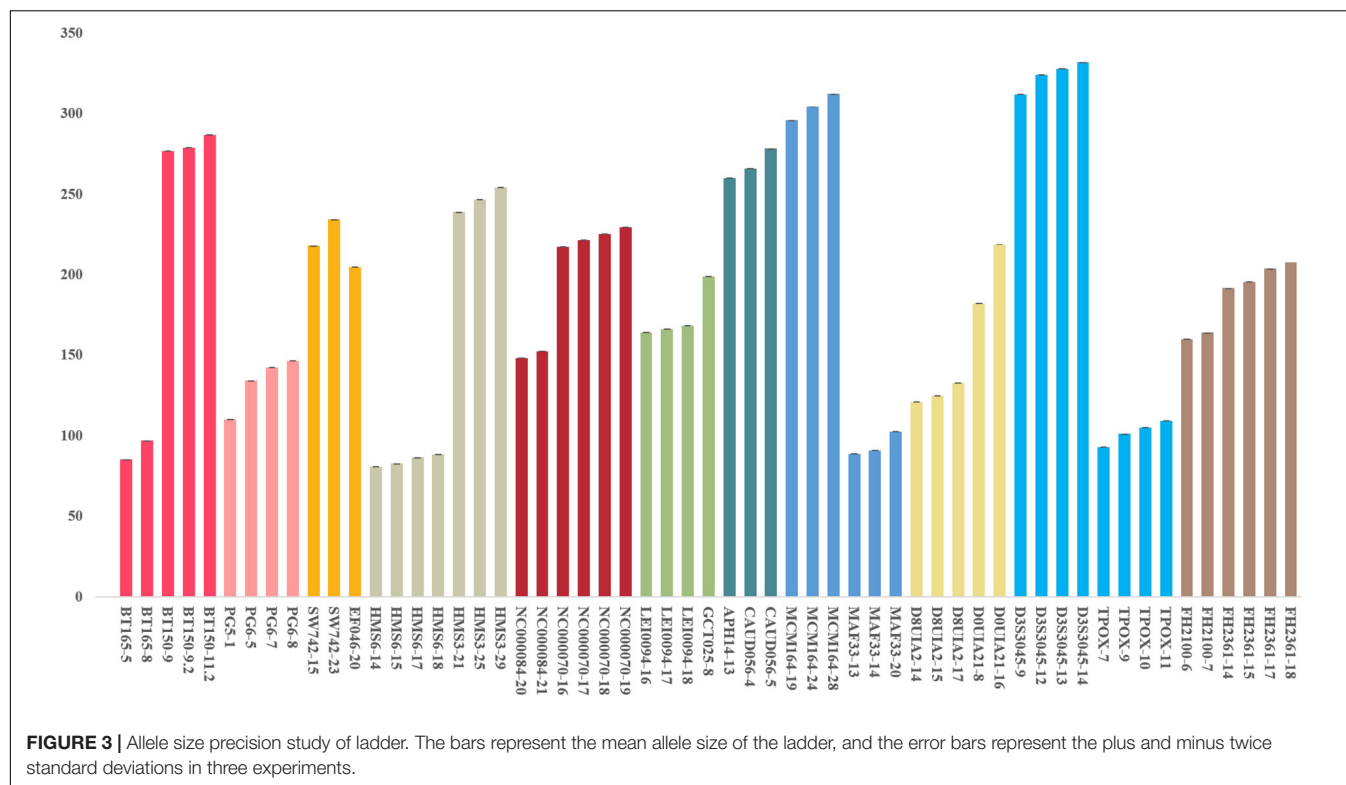


FIGURE 2 | Sensitivity study of input DNA amounts ranging from 5 ng to 62.5 pg. Each bar represents the percentage of detected loci to total loci.

sensitivity of this panel was 0.09 ng (1 ng/11). The allelic ladder and DNA mix were genotyped three separate tests to evaluate the size precision. We measured the deviation of each allele size in these three experiments, and the results are shown in the bar chart in **Figure 3**. The bars in **Figure 3** represent the mean allele size of the ladder, and the error bars represent the plus and minus twice standard deviations in three experiments. **Figure 3** indicates that the standard deviation of each allele size is less than 0.1 bp. The

results showed that in each sample, the allele sizes were consistent with their known amplicon sizes. We sequenced part of the alleles of each STR locus to evaluate the precision of the CE platform. The STR genotyping profiles acquired from the CE platform were consistent with the corresponding results of Sanger sequencing.

Genomic DNA extracted from the liver, heart, spleen, lung, kidney, and muscle of the same SD rat and Kunming mouse, respectively, were used to evaluate whether the STR genotyping



profiles of different tissues or organs from the same individual showed exactly the same STR genotyping result. As shown in **Supplementary Figure S4**, allelic genotyping peaks could only be observed at the D8U1A2 (alleles: 14, 15) and D0U1A21 (alleles: 16, 16) loci when we co-amplified the genomic DNA extracted from different organs or tissue of a SD rat. For the Kunming mouse's various organs or tissue (shown in **Supplementary Figure S5**), STR genotyping peaks could be detected at the NC000084 (allele: 17, 18) and NC000070 (allele: 17, 19) loci belonging to the Kunming mouse.

Specificity, Mixture Study, and Casework Sample Verification

Genomic DNA templates of the studied species were amplified separately based on this multiplex STR panel so that we could evaluate the species specificity of this 22-STR panel, and the corresponding STR genotyping results are shown in **Supplementary Figure S6**. The profiles revealed that no peak was detected for the negative control. Specific allelic peaks were only detected at the corresponding loci for each species, and no allelic peak was found in other species-specific loci.

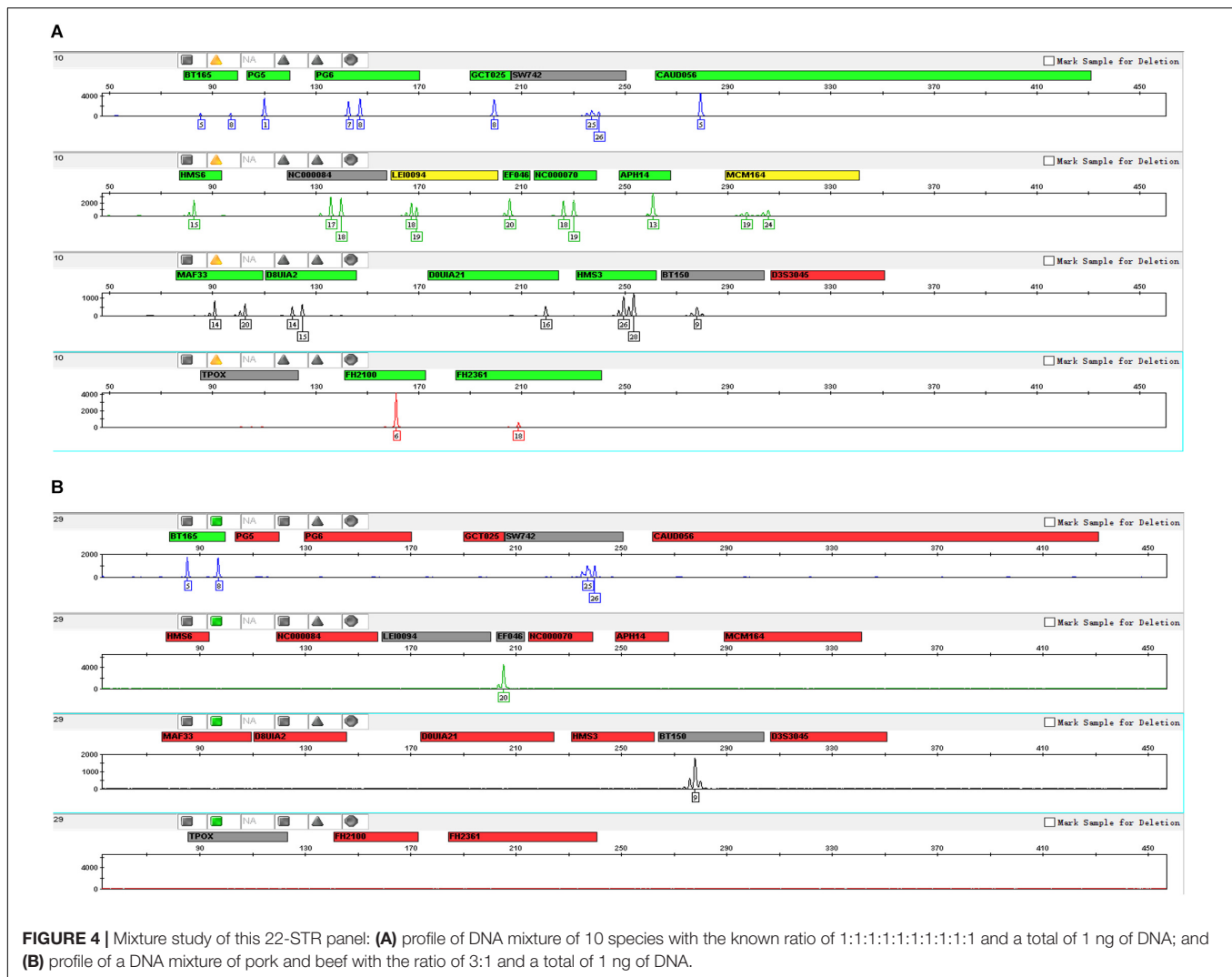
We made two types of DNA mix models with different mixture ratios to evaluate the performance of this panel in the detection of each DNA mixture. In **Figure 4**, we only displayed the genotyping profiles of DNA mixtures of 10 species (without human samples) with the known ratio of 1:1:1:1:1:1:1:1:1:1, and the DNA mixture of pork and beef with a ratio of 3:1. In these two types of mix models, all species were detected, and the detected ratio in the current test was 10% (0.1 ng/1 ng).

We genotyped human blood stains preserved for up to 7 years on FTATM cards at room temperature to evaluate the performance of this novel panel on aged samples. The results showed that all samples could be successfully genotyped at D3S3045 and TPOX loci.

We also genotyped four types of cooked meats of unknown species origin to evaluate the efficiency of the detection of cooked meats, and the profiles are shown in **Supplementary Figure S7**. Two pig-specific STR loci, SW742 (alleles: 15.1, 15.1) and EF046 (alleles: 20, 20), were observed in SG meat, indicating that SG was pork; two sheep-specific STR loci, MCM164 (alleles: 20, 28) and MAF33 (alleles: 13, 20), were observed in SY meat, indicating that it was mutton; two pig-specific STR loci, SW742 (alleles: 15.1, 24) and EF046 (alleles: 20, 20), were observed in SN meat, indicating that it was pork. LEI0094 was a chicken-specific STR locus that was observed in SE meat, but another locus, GCT025, was not detected, giving partial indication it was chicken.

DISCUSSION

With the increasing occurrence of illegal incidents such as meat fraud and adulteration or illegal trade of protected animals (Cawthorn et al., 2013; Tibola et al., 2018; LaFleur et al., 2019), it is fundamentally important in forensic genetics to be able to identify the animal species in an unknown sample. Compared to morphological observation and protein-based method, DNA-based method is regarded as one of the most suitable techniques for species identification due to their tolerance for heat or other environmental influences. To date, many panels for animal



species identification have been developed on the basis of different DNA markers such as autosomal STRs (Dawnay et al., 2008; Liu et al., 2019), species-specific insertions-deletions (InDels) (Alves et al., 2017), and DNA barcodings (Arulandhu et al., 2017).

Over the past two decades, STR loci have been used extensively in population genetics, individual identification, and paternity tests for protected wild animals or domestic animals (Eiken et al., 2009; Gupta et al., 2011; Ogden et al., 2012; Wang et al., 2019). Despite their widespread use in the genetic research of non-human species, there were only a few STR-based panels used for the animal species identification.

Compared to human forensic genetics, research progress of non-human genetics has been more gradual, largely because no rich unified databases of wild animal or domestic animal were available. Besides, genetic markers used in the field of animal genetics have not been systematically validated by forensic medicine, which made them difficult to be used in the forensic genetics (Iyengar, 2014).

Most kits have been developed based on RT-PCR, dPCR, or liquid chromatography-tandem mass spectrometry

(Floren et al., 2015; Kim et al., 2017; Xu et al., 2018). Although higher sensitivity and accuracy were acquired, the need for expensive instruments and their time-consuming operation made it difficult to apply these methods in the primary laboratories of China. At present, a PCR-STR-CE-based method is widely applied in most laboratories in China due to its relatively lower cost, higher efficiency, and mature technical system.

The purpose of this research is to develop a panel that could distinguish 10 animal species as well as human beings, which could then be used in the application of forensic species identification and the detection of meat fraud and adulteration. The choice of the studied species is fully considered based on actual adulteration cases. Pork, beef, mutton, and chicken are the most common meats found in China. Beef or mutton has been found to be adulterated with inexpensive meat such as duck, horse, and even mouse meat, and therefore, chicken, duck, sheep, pig, horse, cattle, rat, and mouse are selected for this study. Additionally, because canine and pigeon meat are also popular in some cities of China, these two animal species are chosen for this study.

We selected 22-STR loci with high species specificity among 11 species and then constructed a novel five-dye multiplex amplification panel that could be analyzed using the CE platform. Before the validations, we evaluated the performance of different thermal cycling parameters. As anticipated, an increasing cycle number led to an apparent increase in overall allelic peak height. All loci could be detected in reasonable ranges of thermal cycling parameters. At 29 cycles, we observed a more balanced peak height.

The annealing temperature affected the specificity of the PCR. In tests of different annealing temperatures, the amplification efficiencies at 58, 59, and 60°C were higher than those at 57 and 61°C. After we considered that low annealing temperature led to non-specific amplification (Rychlik et al., 1990), we finally chose 59°C as the optimal annealing temperature. After evaluating the PCR efficiencies in two reaction volumes, the results revealed that more optimal amplification occurred in the 10- μ l volume (rather than 25 μ l) containing 1 μ l of DNA template, 2 μ l of Primer set, 4 μ l of Master Mix I, and 3 μ l of deionized water.

It is essential to evaluate the efficiency of a novel panel before it is used for casework. Here, we performed a series of developmental validations studies including sensitivity, reproducibility, precision, specificity, mixture, and tissue/organ consistency and so on. In forensic practice, we could not always acquire sufficient DNA amounts, and therefore, any potential panel should be capable of genotyping trace amounts of DNA template. In the current study, we evaluated the sensitivity of this 22-STR panel with serial input DNA amounts. According to the results, one dropped peak (HMS3, allele 29) was observed when the input DNA was 0.5 ng, indicating that the minimum input amount of DNA template should be more than 0.5 ng. Reproducibility and precision studies were performed to validate the reliability and accuracy of this 22-STR panel. The results of reproducibility studies showed that the STR profiles of three trials were consistent, and allele calling was consistent with their known amplicon sizes, which demonstrated that this panel could ensure proper allele detection.

It was critical to ensure that this 22-STR panel exhibited no cross-reactivity between different species. The primer specificity of the STR loci was the key to the specificity of this panel. To ensure that no cross-reactivity occurred among the 11 species, we designed the primers according to the highly conserved region of each species' genome and used BLAST to evaluate the specificity of each primer sequence. The present species specificity study showed that the specific peaks of the STRs were detected only at the corresponding loci for each species, and no allelic peaks were found in the STR loci of other species, indicating that all the primers in this panel exhibited no cross-reactivity between different species.

In the ongoing investigations of meat fraud and adulteration, it is usually found that various inexpensive meats such as chicken or duck are often added to beef or mutton, which not only decreases food safety, but also disrupts market order. Illegal addition of animal-derived ingredients in feedstuff might spread infectious diseases such as bovine spongiform encephalopathy or scrapie (Gao et al., 2017). Therefore, it is

of fundamental importance to develop a panel with a high efficiency for the detection of the individual components in meat mixtures. According to the results of the mixture studies, all species in each mixture pattern could be detected, indicating that this panel would adequately function in the detection of mixed samples.

In addition to meat fraud and adulteration, processed meat or animal tissues are also commonly investigated in forensic casework. Poached and roasted lamb, beef, and pork are popular in the Chinese diet. High temperatures and various condiments used during cooking could damage DNA. Therefore, the efficiency of the detection of cooked meats or animal tissues is also essential. In this research, we used four different cooked meats to evaluate the detection efficiency of the mixture of cooked meat. Full profiles were acquired for SG, SY, and SN meats, and one locus was detected in SE meat. The results indicated that this 22-STR panel could be used for the detection of individual species in cooked meat.

CONCLUSION

In this research, we developed a novel 5-dye panel that could simultaneously identify 10 animal species and human being, and co-amplify 22-STR loci using one PCR system. This panel was validated by a series of tests including optimization of PCR conditions, sensitivity, reproducibility, precision, species specificity, DNA mixture, and tissue/organ consistency. In present results, this 22-STR panel achieved high species specificity among 11 species and a detection capacity for a mixture of meat samples. The results of the developmental validations demonstrated that this panel can be used for forensic species identification and the detection of meat fraud and adulteration.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics committee of the Xi'an Jiaotong University, Health Science Center. The patients/participants provided their written informed consent to participate in this study. The animal study was reviewed and approved by the Ethics committee of the Xi'an Jiaotong University, Health Science Center.

AUTHOR CONTRIBUTIONS

BZ designed and was responsible for this research. WC and XJ built up this 22-STR panel and prepared the preliminary

data. WC, YG, and CC analyzed the data. WC wrote the draft manuscript. WZ, JL, YW, and BZ reviewed and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Shaanxi Science and Technology Coordination Innovation Project (2015KTCL03-03), the National Natural Science Foundation of China (NSFC, 81525015), and the Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (GDUPS, 2017).

ACKNOWLEDGMENTS

We thank the technical assistance of Beijing Microread Genetics Technology Co., Ltd.

REFERENCES

- Alves, C., Pereira, R., Prieto, L., Aler, M., Amaral, C. R. L., Arevalo, C., et al. (2017). Species identification in forensic samples using the SPInDel approach: a GHEP-ISFG inter-laboratory collaborative exercise. *Forensic Sci. Int. Gen.* 28, 219–224. doi: 10.1016/j.fsigen.2017.03.003
- Arulandhu, A. J., Staats, M., Hagelaar, R., Voorhuijzen, M. M., Prins, T. W., Scholtens, I., et al. (2017). Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *Gigascience* 6, 1–18. doi: 10.1093/gigascience/gix080
- Budowle, B., Garofano, P., Hellman, A., Ketchum, M., Kanthaswamy, S., Parson, W., et al. (2005). Recommendations for animal DNA forensic and identity testing. *Int. J. Leg. Med.* 119, 295–302. doi: 10.1007/s00414-005-0545-9
- Cawthorn, D. M., Steinman, H. A., and Hoffman, L. C. (2013). A high incidence of species substitution and mislabelling detected in meat products sold in South Africa. *Food Control* 32, 440–449. doi: 10.1016/j.foodcont.2013.01.008
- Chen, L., Du, W. A., Wu, W. B., Yu, A. L., Pan, X. Y., Feng, P. P., et al. (2019). Developmental validation of a novel six-dye typing system with 47 A-InDels and 2 Y-InDels. *Forensic Sci. Int. Gen.* 40, 64–73. doi: 10.1016/j.fsigen.2019.02.009
- Chun-Lee, J., Tsai, L. C., Kuan, Y. Y., Chien, W. H., Chang, K. T., Wu, C. H., et al. (2007). Racing pigeon identification using STR and chromo-helicase DNA binding gene markers. *Electrophoresis* 28, 4274–4281. doi: 10.1002/elps.200700063
- Dawnay, N., Ogden, R., Thorpe, R. S., Pope, L. C., Dawson, D. A., and McEwing, R. (2008). A forensic STR profiling system for the Eurasian badger: a framework for developing profiling systems for wildlife species. *Forensic Sci. Int. Gen.* 2, 47–53. doi: 10.1016/j.fsigen.2007.08.006
- Eiken, H. G., Andreassen, R. J., Kopatz, A., Bjervamo, S. G., Wartianen, I., Tobiasen, C., et al. (2009). Population data for 12 STR loci in Northern European brown bear (*Ursus arctos*) and application of DNA profiles for forensic casework. *Forensic Sci. Int. Gen. Suppl. Ser.* 2, 273–274. doi: 10.1016/j.fsigs.2009.07.007
- Fang, Y., Guo, Y., Xie, T., Jin, X., Lan, Q., Zhou, Y., et al. (2018). Forensic molecular genetic diversity analysis of Chinese Hui ethnic group based on a novel STR panel. *Int. J. Leg. Med.* 132, 1297–1299. doi: 10.1007/s00414-018-1829-1
- Floren, C., Wiedemann, I., Brenig, B., Schutz, E., and Beck, J. (2015). Species identification and quantification in meat and meat products using droplet digital PCR (ddPCR). *Food Chem.* 173, 1054–1058. doi: 10.1016/j.foodchem.2014.10.138
- Fordyce, S. L., Mogensen, H. S., Børsting, C., Lagacé, R. E., Chang, C.-W., Rajagopalan, N., et al. (2015). Second-generation sequencing of forensic STRs using the Ion Torrent™ HID STR 10-plex and the Ion PGM™. *Forensic Sci. Int. Gen.* 14, 132–140. doi: 10.1016/j.fsigen.2014.09.020
- Gao, F., Zhou, S. M., Yang, Z. L., Han, L. J., and Liu, X. (2017). Study on the characteristic spectral properties for species identification of animal-derived feedstuff using Fourier transform infrared spectroscopy. *Appl. Spectrosc.* 71, 2446–2456. doi: 10.1177/0003702817732323
- Godfray, H. C. J., Aveyard, P., Garnett, T., Hall, J. W., Key, T. J., Lorimer, J., et al. (2018). Meat consumption, health, and the environment. *Science* 361:eaam5324. doi: 10.1126/science.aam5324
- Gupta, S. K., Bhagavatula, J., Thangaraj, K., and Singh, L. (2011). Establishing the identity of the massacred tigress in a case of wildlife crime. *Forensic Sci. Int. Gen.* 5, 74–75. doi: 10.1016/j.fsigen.2010.05.004
- Iyengar, A. (2014). Forensic DNA analysis for animal protection and biodiversity conservation: a review. *J. Nat. Conserv.* 22, 195–205. doi: 10.1016/j.jnc.2013.12.001
- Kim, G. D., Seo, J. K., Yum, H. W., Jeong, J. Y., and Yang, H. S. (2017). Protein markers for discrimination of meat species in raw beef, pork and poultry and their mixtures. *Food Chem.* 217, 163–170. doi: 10.1016/j.foodchem.2016.08.100
- LaFleur, M., Clarke, T. A., Reuter, K. E., Schaefer, M. S., and terHorst, C. (2019). Illegal trade of wild-captured lemur catta within madagascar. *Folia Primatol.* 90, 199–214. doi: 10.1159/000496970
- Liu, Y. L., Xu, J., Chen, M. X., Wang, C. F., and Li, S. C. (2019). A unified STR profiling system across multiple species with whole genome sequencing data. *BMC Bioinformatics* 20:671. doi: 10.1186/s12859-019-3246-y
- Lou, X. P., Zhang, W., Zheng, J., Xu, H., and Zhao, F. (2016). Comparative study on morphology of human, swine, sheep and cattle muscle tissues and its forensic significance. *Fa Yi Xue Za Zhi* 32, 250–253. doi: 10.3969/j.issn.1004-5619.2016.04.003
- Matsuzawa, S., Kimura, H., Itoh, Y., Wang, H., and Nakagawa, T. (1993). A rapid dot-blot method for species identification of bloodstains. *J. Forensic Sci.* 38, 448–454.
- Nicolas, V., Schaeffer, B., Missouf, A. D., Kennis, J., Colyn, M., Denys, C., et al. (2012). Assessment of three mitochondrial genes (16S, Cytb, CO1) for identifying species in the Praomiyini tribe (*Rodentia: Muridae*). *PLoS One* 7:e36586. doi: 10.1371/journal.pone.0036586
- Ogden, R., Mellanby, R. J., Clements, D., Gow, A. G., Powell, R., and McEwing, R. (2012). Genetic data from 15 STR loci for forensic individual identification and parentage analyses in UK domestic dogs (*Canis lupus familiaris*). *Forensic Sci. Int. Genet.* 6, e63–e65. doi: 10.1016/j.fsigen.2011.04.015
- O'Mahony, P. J. (2013). Finding horse meat in beef products—a global problem. *QJM* 106, 595–597. doi: 10.1093/qjmed/hct087

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.01005/full#supplementary-material>

FIGURE S1 | Genotyping profiles of annealing temperatures at 57, 58, 59, 60, and 61°C.

FIGURE S2 | PCR efficiency studies of the two different reaction volume systems: 10 and 25 µl.

FIGURE S3 | Genotyping profiles of PCR efficiency studies using four different types of PCR machines.

FIGURE S4 | Genotyping profiles of DNA samples extracted from the liver, heart, spleen, lung, kidney, and muscle of the same SD rat.

FIGURE S5 | Genotyping profiles of DNA samples extracted from the liver, heart, spleen, lung, kidney, and muscle of the same Kunming mouse.

FIGURE S6 | Genotyping profiles of species specificity studies on this 22-STR panel.

FIGURE S7 | Genotyping profiles of four cooked meat samples amplified by this STR panel.

- Rychlik, W. (2007). OLIGO 7 primer analysis software. *Methods Mol. Biol.* 402, 35–60. doi: 10.1007/978-1-59745-528-2_2
- Rychlik, W., Spencer, W. J., and Rhoads, R. E. (1990). Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res.* 18, 6409–6412. doi: 10.1093/nar/18.21.6409
- Skouridou, V., Tomaso, H., Rau, J., Bashammakh, A. S., El-Shahawi, M. S., Alyoubi, A. O., et al. (2019). Duplex PCR-ELONA for the detection of pork adulteration in meat products. *Food Chem.* 287, 354–362. doi: 10.1016/j.foodchem.2019.02.095
- Staats, M., Arulandhu, A. J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., et al. (2016). Advances in DNA metabarcoding for food and wildlife forensic species identification. *Anal. Bioanal. Chem.* 408, 4615–4630. doi: 10.1007/s00216-016-9595-8
- Tibola, C. S., da Silva, S. A., Dossa, A. A., and Patricio, D. I. (2018). Economically motivated food fraud and adulteration in Brazil: incidents and alternatives to minimize occurrence. *J. Food Sci.* 83, 2028–2038. doi: 10.1111/1750-3841.14279
- Wang, L., Chen, M., Wu, B., Liu, Y. C., Zhang, G. F., Jiang, L., et al. (2018). Massively parallel sequencing of forensic STRs using the Ion Chef™ and the Ion S5™ XL systems. *J. Forensic Sci.* 63, 1692–1703. doi: 10.1111/1556-4029.13767
- Wang, M. L., Jin, X. Y., Xiong, X., Yang, J. L., Li, J. P., Wang, Q., et al. (2019). Polymorphism analyses of 19 STRs in Labrador Retriever population from China and its heterozygosity comparisons with other retriever breeds. *Mol. Biol. Rep.* 46, 1577–1584. doi: 10.1007/s11033-019-04601-4
- Wilson-Wilde, L., Norman, J., Robertson, J., Sarre, S., and Georges, A. (2010). Current issues in species identification for forensic science and the validity of using the cytochrome oxidase I (COI) gene. *Forensic Sci. Med. Pathol.* 6, 233–241. doi: 10.1007/s12024-010-9172-y
- Xu, R., Wei, S., Zhou, G., Ren, J., Liu, Z., Tang, S., et al. (2018). Multiplex TaqMan locked nucleic acid real-time PCR for the differential identification of various meat and meat products. *Meat Science* 137, 41–46. doi: 10.1016/j.meatsci.2017.11.003
- Zhang, P., Zhu, Y., Li, Y., Zhu, S., Ma, R., Zhao, M., et al. (2018). Forensic evaluation of STR typing reliability in lung cancer. *Leg. Med.* 30, 38–41. doi: 10.1016/j.legalmed.2017.11.004

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

Copyright © 2020 Cui, Jin, Guo, Chen, Zhang, Wang, Lan 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.



Broadening the Applicability of a Custom Multi-Platform Panel of Microhaplotypes: Bio-Geographical Ancestry Inference and Expanded Reference Data

María de la Puente^{1,2*}, Jorge Ruiz-Ramírez¹, Adrián Ambroa-Conde¹, Catarina Xavier², Jorge Amigo³, María Ángeles Casares de Cal⁴, Antonio Gómez-Tato⁴, Ángel Carracedo^{1,3}, Walther Parson^{2,5}, Christopher Phillips^{1*} and María Victoria Lareu¹

OPEN ACCESS

Edited by:

Cemal Gurkan,
Turkish Cypriot DNA Laboratory
(TCDL), Cyprus

Reviewed by:

Guanglin He,
Sichuan University, China
Tábita Hünemeier,
University of São Paulo, Brazil
Peng Chen,
Nanjing Medical University, China

*Correspondence:

María de la Puente
m.delapuate.vila@gmail.com
Christopher Phillips
c.phillips@mac.com

Specialty section:

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

Received: 07 July 2020

Accepted: 25 September 2020

Published: 20 October 2020

Citation:

de la Puente M, Ruiz-Ramírez J, Ambroa-Conde A, Xavier C, Amigo J, Casares de Cal MÁ, Gómez-Tato A, Carracedo Á, Parson W, Phillips C and Lareu MV (2020) Broadening the Applicability of a Custom Multi-Platform Panel of Microhaplotypes: Bio-Geographical Ancestry Inference and Expanded Reference Data. *Front. Genet.* 11:581041. doi: 10.3389/fgene.2020.581041

¹ Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain, ² Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria, ³ Fundación Pública Galega de Medicina Xenómica (FPGMX), Santiago de Compostela, Spain, ⁴ Faculty of Mathematics, University of Santiago de Compostela, Santiago de Compostela, Spain, ⁵ Forensic Science Program, The Pennsylvania State University, University Park, PA, United States

The development of microhaplotype (MH) panels for massively parallel sequencing (MPS) platforms is gaining increasing relevance for forensic analysis. Here, we expand the applicability of a 102 autosomal and 11 X-chromosome panel of MHs, previously validated with both MiSeq and Ion S5 MPS platforms and designed for identification purposes. We have broadened reference population data for identification purposes, including data from 240 HGDP-CEPH individuals of native populations from North Africa, the Middle East, Oceania and America. Using the enhanced population data, the panel was evaluated as a marker set for bio-geographical ancestry (BGA) inference, providing a clear differentiation of the five main continental groups of Africa, Europe, East Asia, Native America, and Oceania. An informative degree of differentiation was also achieved for the population variation encompassing North Africa, Middle East, Europe, South Asia, and East Asia. In addition, we explored the potential for individual BGA inference from simple mixed DNA, by simulation of mixed profiles followed by deconvolution of mixture components.

Keywords: microhaplotypes, massively parallel sequencing, bio-geographical ancestry, mixed DNA, human identification

INTRODUCTION

Microhaplotypes (MHs), defined as sets of SNPs in sequence segments of less than 200 base-pairs (bp), which define multi-allelic haplotypes, have been proposed as forensic markers in concert with the forensic adoption of massively parallel sequencing (MPS) technologies (Kidd et al., 2014; Oldoni et al., 2018). MPS platforms allow the detection of the phase of the SNP alleles in MH loci from the generated monoclonal (single strand) sequences, in contrast to other SNP genotyping methods used in forensics (Sobrinho et al., 2005) or Sanger sequencing. The favorable characteristics of MH loci has prompted the search and characterization of new MH markers for forensic use and their genotyping using MPS-based panels (Kidd and Speed, 2015; Kidd et al., 2017; Chen et al., 2018, 2019a,b;

van der Gaag et al., 2018; Voskoboinik et al., 2018; Bennett et al., 2019; De La Puente et al., 2019; Phillips et al., 2019; Turchi et al., 2019; Gandotra et al., 2020; Sun et al., 2020).

Three notable advantages of MHs are: a higher degree of polymorphism compared to single-site SNPs; the absence of stutter artifacts; and short amplicon lengths compared to STRs. Therefore, possible applications of MHs include a wide range of forensic scenarios: individual identification from degraded DNA (van der Gaag et al., 2018), kinship testing (Sun et al., 2020), mixture analysis (Voskoboinik et al., 2018; Bennett et al., 2019; Chen et al., 2019a) and bio-geographical ancestry (BGA) prediction (Chen et al., 2019b; Phillips et al., 2019). Moreover, the same markers have been proposed for multiple forensic applications examined simultaneously, constituting a multi-purpose set of panels (Oldoni et al., 2017; Turchi et al., 2019; Gandotra et al., 2020).

Here, we have made new evaluations of a previously published multi-platform (MiSeq and Ion S5) panel of 102 autosomal and 11 X-chromosome MHs validated for forensic identification (De La Puente et al., 2019) (herein MHs-panel), in order to: (i) expand the available reference dataset with native populations from major groups not covered by the 1,000 Genomes Project; (ii) provide a comprehensive description of the BGA prediction capabilities of the panel; and (iii) test the possibility of obtaining individual BGA predictions from the deconvoluted contributors detected in simple mixed profiles.

MATERIALS AND METHODS

DNA Samples, Library Construction and Sequencing

A total of 246 DNAs were analyzed from the HGDP-CEPH Human Genome Diversity Panel (Cann et al., 2002) (herein CEPH), comprising: (i) 28 Oceanians—17 Papuan from New Guinea and 11 Melanesian from Bougainville; (ii) 62 Native Americans—14 Karitiana, 8 Surui from Brazil; 20 Maya, 13 Pima from Mexico; and 7 Piapoco from Colombia; (iii) 127 Middle East—40 Druze from Israel (Carmel), 42 Palestinian from Israel (Central), 45 Bedouin from Israel (Negev); and (iv) North-Africans—29 Mozabite from Algeria (Mzab).

Library preparation was performed with AmpliSeq Precision ID Library Kit [Thermo Fisher Scientific (TFS)] and Ion Xpress Barcode Adapters (TFS) optimizing the manufacturer's recommendations to half-volumes. A total of 1 ng of input DNA was used, quantified with Qubit 3.0 Fluorometer (TFS) and Qubit dsDNA HS Assay Kit (TFS) following the manufacturer's recommendations. The primer pool was described in De La Puente et al. (2019). Briefly, a total of 107 (10 Mb-spaced) autosomal and 11 (5 Mb-spaced) X-chromosome short highly polymorphic MHs were identified from 1,000 Genomes public data as optimal forensic MH markers and incorporated in a single-pool Hotspot AmpliSeq design targeting Formalin-Fixed Paraffin-Embedded (FFPE) DNA (i.e., with amplicons of 125–175 nucleotide lengths highly suitable for degraded DNA). Individual libraries were quantified with the Ion Library TaqMan Quantitation Kit (TFS), following manufacturer's protocols.

Equimolar pools of 39 to 46 libraries at 20–30 pM were prepared for sequencing. Template preparation was performed using the Ion 510, Ion 520, Ion 530 Kit-Chef (TFS), Ion 530 chips (TFS) and the Ion Chef Instrument. Sequencing was performed on the Ion S5 instrument with a read length of 200 (500 flows).

Data Curation and Concordance With Databases

Sequencing quality parameters including sequence coverage, strand bias, allele balance and misincorporation rates were evaluated using single SNP data produced with the HID Genotyper plugin v.5.2.2 (TFS) of Torrent Suite v. 5.6.0 (TFS) using default parameters of minimum coverage of six reads and minimum allele read frequency of 0.1.

Microhaplotype calling was performed using the pipeline described in De La Puente et al. (2019), that allows inferring the phase of the SNPs on the same amplicon from the sequence reads obtained. Briefly, FASTQ reads were aligned using Burrows-Wheeler aligner (BWA) (Li and Durbin, 2009) to a customized reference genome comprising each MH amplicon joined. Alignments were processed with SAMtools (Li et al., 2009) to create the input files for the microhaplot R package (Thomas, 2019), which outputs a raw table of allele strings and depth per MH. Minor allele read frequency and minimum coverage filtering parameters were set to the default values of 0.1 and 15, respectively. A total of five MHs: 3pC, 5qD, 10qC, 12qA, and 19qB, were included in the primer set but previously identified as unreliable and therefore excluded from analysis; and genotypes were manually corrected, when necessary, according to the guidelines in De La Puente et al. (2019).

Genotyping and phase concordance with publicly available data from Simons Genome Diversity Project (SGDP) (Mallick et al., 2016) and recent whole genome sequencing of the HGDP panel (Almarri et al., 2020; Bergstrom et al., 2020) (herein HGDP WGS) was evaluated. SGDP dataset is phased using the probabilistic software IMPUTE2 (Howie et al., 2009) with 1,000 Genomes data as reference. SGDP lists whole-genome variant data for 280 worldwide samples, but 21 are overlaps with 1,000 genomes sample sets, and 133 are samples from the CEPH panel. In total, 35 CEPH samples overlapped between SGDP and those we genotyped from Middle East, Oceanian and American populations. The HGDP WGS dataset infers the phase of heterozygous SNPs with GATK HaplotypeCaller (McKenna et al., 2010; Poplin et al., 2018) for a total of 929 HGDP-CEPH panel samples of which 234 overlap with those we genotyped for the MH loci. GATK HaplotypeCaller reassembles active regions with significant variation in order to identify all the possible haplotypes, then for each haplotype a likelihood is calculated given the sequence read data by aligning each read against each haplotype and based on those likelihoods the genotypes are assigned.

Population Metrics and Bio-Geographical Ancestry Analysis

Population data for haplotype frequency estimation and BGA analysis was obtained from 1,000 Genomes project

phase III public releases (The Genomes Project Consortium, 2015) (herein 1 KG) and the genotyping of HGDP-CEPH populations. Additionally, data for 679 HGDP-CEPH individuals from 42 Sub-Saharan African, European, Central and South Asian and East Asian populations was collected from HGDP WGS. These populations comprise a limited number of individuals and descriptive analyses such as frequencies or F_{ST} were not conducted.

Population haplotype frequencies, expected Heterozygosity values (as 1 minus the sum of the squares of the haplotype frequencies) and cumulative match probabilities (as the product of the sum of the squares of the genotype probabilities of each locus) were calculated and plotted using R v. 3.6.1 (R Core Team, 2019) or Excel spreadsheets. F_{ST} and average number of pairwise differences within and between population were calculated using Arlequin version 3.5.1.2 (Excoffier and Lischer, 2010).

Bio-geographical ancestry analyses were conducted considering the autosomal MHs as independent markers and their haplotypes as alleles. Analyses with STRUCTURE v. 2.3.4 (Pritchard et al., 2000) were performed following guidelines in Porras-Hurtado et al. (2013), including the following parameters: five iterations for each K, one million burnin steps and one million MCMC steps, correlated allele frequencies under the Admixture model. When combining both reference and non-reference populations, the option “Update allele frequencies using only individuals with POPFLAG = 1” was selected and reference populations were set to 1. The optimum K was estimated considering the output graphs generated with Structure Harvester (Earl and Von Holdt, 2012). Ancestry membership was plotted using the CLUMPAK portal (Kopelman et al., 2015). Multidimensional scaling (MDS) analyses and Neighbor-joining (NJ) trees were constructed with R v. 3.6.1 (R Core Team, 2019) over an allele-distance matrix computed using the R package *pegas* (Paradis, 2010).

Population-specific Divergence (PSD) and simple pairwise Divergence values were calculated using infocalc v. 1.1 for obtaining Rosenberg’s informativeness-for-assignment metric (*In*) (Rosenberg et al., 2003). For PSD, individual profiles were marked as AFR and non-AFR, etc.; and for pairwise comparisons, each pair of populations was grouped. In values for each autosomal MH were summed to obtain cumulative values. As explained in Cheung et al. (2019), *In* is the most convenient metric for assessment of BGA informativeness in different types of genomic markers.

Mixture Simulation, Profile Deconvolution and BGA Inference From Components

Three mixed profiles including 102 autosomal MHs were simulated from single source profiles of known ancestry, comprising: (i) a 1:3 mixture of HG02922 unadmixed ESN (AFR) and NA18939 unadmixed JPT (EAS)—herein, mixture 1; (ii) a 1:5 mixture of HG00097 unadmixed GBR (EUR) and HG00096 unadmixed GBR (EUR)—herein, mixture 2; and (iii) a 1:7 mixture of HG01565 admixed PEL (AMR) and HG00096 unadmixed GBR (EUR)—herein, mixture 3.

Two analysts conducted a blind deconvolution of each of the mixed profiles, instructed to separate two components (minor and major) assigning only the haplotypes that were unequivocally from one of the components when taking into account stochastic phenomena (allele drop-out, heterozygous imbalance). Results from both analysts were merged maintaining the most conservative profile when interpretations differed, and BGA inference analysis comprising STRUCTURE and MDS were performed as described in section “Population Differentiation and BGA Inference Performance.”

RESULTS

Assay Performance and Genotyping Data Curation

Details of the overall performance of the seven sequencing runs are collected in **Supplementary Table S1**. All chips reached a satisfactory loading performance, with percentages ranging from 72 to 90%. In order to reduce the high proportion of polyclonal reads observed initially (38%), the molar concentration of the library pool was progressively lowered to 20 pM. Even when the number of chips is not statistically sufficient to test this effect, a tendency toward lower polyclonality was generally observed, except for chip 4.

Supplementary Table S2 and **Supplementary Figure S1** outline the target coverage per sample. Samples reached comparable levels of overall mean coverage value across MHs of $3,572.33 \pm 2,601.39$ reads. Uniformity was maintained both within and among sequencing runs, with few samples giving values beyond the overall mean coverage. Most samples from chip 4 showed lower median coverage values, probably due to the fact that sample HGDP00693 had mean coverage values nearly eight times higher than the overall mean ($28,313.96 \pm 9,743$). This excessive sequence coverage was most likely caused by erroneous quantification of the sample library (i.e., the library concentration was underestimated and pooled at a much higher concentration than 20 pM) and explains the high polyclonality of chip 4.

Supplementary Figure S2 shows normalized coverage values per marker, calculated as MH coverage per sample/total sample coverage. As expected, from previous analyses using the same primer pools, results closely match those found from the initial panel validation (De La Puente et al., 2019), with 6pB, 17qC, XpB, and 16pB having the lowest normalized coverage values. Coverage values per marker in each sample are shown in **Supplementary Figure S3**. For the problem MHs mentioned above, a high proportion of samples did not reach a minimum of 15 reads, affecting the calling process and genotype completeness of the typed samples. This was anticipated before sequencing but the loss of data from these discounted MH loci did not unduly affect the panel’s informativeness, taking into account the fact that most BGA panels can accommodate some degree of missing values.

Regarding strand bias, represented in **Supplementary Figure S4**, most MHs ranged between the 40–60% of forward coverage/total coverage. When compared to the initial

evaluation, MHs XqA and 12pA presented a slight degree of reverse strand bias, which had not been previously observed. In contrast, 11qC and 19qA presented some forward strand bias uniquely in this study.

Allele read frequency balance is described in **Supplementary Figure S5** as the percentage of reference allele sequence reads. For single source DNA samples, these frequencies would ideally cluster closely around 50% for heterozygous genotypes and 0 or 100% in homozygotes for the alternative or reference allele, respectively. Most MHs values are close to the expected values, with few outliers. MHs 6pB, 17qC, XpB, and 16pB display highly scattered plots that can be explained by stochastic PCR effects due to low coverage, as is often observed. In contrast with the initial evaluation, MHs 1qC, 7pC, 14qA, and 19qA showed adequate balance in this study, possibly due to the effect of a higher sample size.

Supplementary Figure S6 outlines the mean percentage misincorporation (as non-allelic bases detected at the SNP site/total coverage). Overall misincorporation rates reached levels of $0.29 \pm 0.71\%$, a value closely matching that previously observed for these loci ($0.25 \pm 0.73\%$). Outlier misincorporation rates between the 5 and 1% thresholds were observed in MHs 15qB (4.69%), 1pC (3.21%), 4qB (2.32%), 6qD (1.60%), 13qD (1.47%), 9qA (1.52%), XqA (1.25%), 21qA (1.22%), 13qB (1.11%), and 7qC (1.03%). Some of these MHs were previously reported as sited within repetitive regions. However, these values did not come close to the 10% minimum allele read frequency used for MH-allele calling, and therefore, genotyping accuracy was not unduly affected.

After MH component SNP genotype calling, six samples: HGDP00588, HGDP00627, HGDP00634, HGDP00637, HGDP00640, and HGDP00642 showed highly imbalanced profiles with more than two haplotypes for several markers, and were excluded from further analysis, as this was most probably due to reference DNA contamination.

Concordance With Online Variant Databases

Concordance with SGDP phased data comprised a total of 3,220 comparisons for 92 markers—note that all X-chromosome loci plus 10 autosomal MHs are not listed by SGDP. Comparisons were made in 35 samples (17 OCE, 10 AMR, 6 ME and 2 NAF). In addition, 82 genotypes could not be compared due to the lack of results from genotyping, most of these in MHs that showed the lowest coverage values: 6pB, 17qC, and 16pB. Concordance rates reached levels of 99.01%, with 31 discordances in 3,138 genotypes confined to 10 MHs, as listed in **Supplementary Table S3**.

All the discordances were explored further in IGV in order to clarify possible causes. Most discordances (21/31) we presume to be caused by the use of probabilistic software to phase the SGDP SNP genotype data (i.e., with IMPUTE2 software) in the following two ways: (i) erroneous phasing of heterozygous alleles in MHs 13qD, 20pA, and 22qA; or (ii) the software does not account for multi-allelic SNPs (i.e., more than two common alleles at the SNP site)—affecting MHs 1qC, 5qB, and 11qA. This supports the idea that more accurate phasing is obtained through

applying MPS to short MHs sequenced as single strands, rather than inferring phase from individually genotyped SNPs.

For MH 16qB, previously identified as underperforming, a total of seven discordancies were found, due to allele drop-out (i.e., one of the alleles did not reach the minimum coverage threshold of 15 reads) during genotype calling. These genotypes were corrected for further analysis. Also, single discordancies were found for MHs 1qD, 7pC, and 11pA. In 11pA the discordancy was due to high allele imbalance of the sequence reads and was corrected; while the cause of the others remained unclear.

For concordance with HGDP WGS, a total of 234 out of the 240 analyzed samples—i.e., all excluding HGDP01003, HGDP01006, HGDP01042, HGDP01051, HGDP01273, and HGDP01278—were compared in 113 loci, adding up to a total of 26,442 comparisons. A total of 1,321 comparisons were inconclusive due to: (i) lack of genotypes in either dataset; (ii) HGDP WGS does not list the first SNP of MH XqD, located in position 93531382 (GRCh37/hg19) or 94276383 (GRCh38/hg38) and (iii) HGDP WGS does not provide phase information for the loci located on the X, thus haplotype reconstruction was not possible for MH loci comprising two or more heterozygous SNPs.

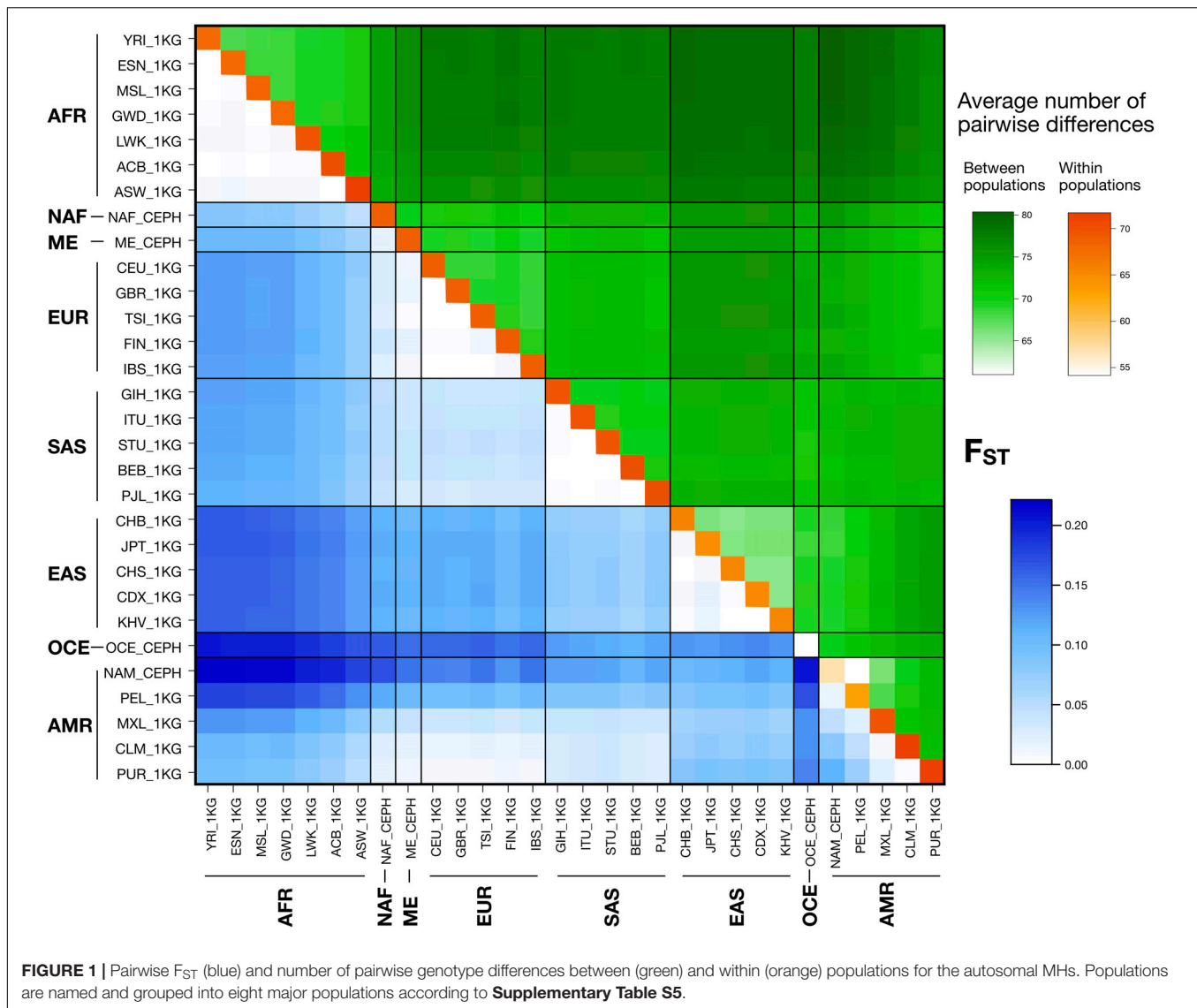
The concordance rate reached 99.75%, with only 62 discrepancies observed (details can be found in **Supplementary Table S4**). Similar to the comparisons with SGDP, the majority of these discordances (43/62) were observed in MHs 16pB, XpB, 17qC, and 6pB (with 18, 13, 4, and 4 discordances each, respectively), previously defined as misperforming markers in terms of coverage which caused high allele imbalance and allelic drop-out.

Low coverage and high allelic imbalance were also the causes for three discordances in MH 7pC; three in MH 12qC and one in MH 11pA. In addition, eight discrepancies were related to the phasing of MHs 1qD, 3qC, 13qD, and 17qA; these were analyzed thoroughly in IGV in order to confirm the phase obtained through the MHs panel. Moreover, raw data from the HGDP WGS sequencing project was inspected in IGV, resulting in confirmation of the phase obtained for the MHs panel. Therefore, the phasing algorithm performed erroneously in a very limited number of cases, which could be due to the fact that WGS reads do not necessarily reach all the SNPs in the amplicon.

Population Metrics

Details of the thirty populations included in this study are listed in **Supplementary Table S5**. Eight major populations were considered: AFR, sub-Saharan Africa; EUR, Europe; NAF, North Africa; ME, Middle East; SAS, South Asia; EAS, East Asia; OCE, Oceania; and AMR, America. For each major population, all individuals from different CEPH populations were gathered into a single population group, in order to achieve high sampling scales, although this was still relatively small for Oceanians.

Allele frequency estimates for 30 populations are given in **Supplementary File S1** and genotypes/haplotypes listed in **Supplementary Table S6**. The latter contains information on the total number of chromosomes typed and data completeness per MH; and total number of counts per SNP allele. This information is intended to emphasize the need for caution with the



frequency estimates derived from populations with few sampled individuals, especially NAF and OCE; as well as highlighting the underperforming MHs such as 6pB, 17qC, XpB, and 16pB.

Figure 1 represents pairwise F_{ST} values and average numbers of pairwise differences within and between populations, considering data from the 102 autosomal MHs. Pairwise F_{ST} values ranged from $7.00E-5$ to $2.21E-1$. As expected, low values were found when comparing populations within the same major population group and for comparisons including those between admixed AMR populations with higher proportions of European contributions (CLM, PUR) and the EUR populations. Higher values were found in comparisons between the AFR populations and EAS, OCE and AMR populations with a low degree of admixture, following the known demographic histories of continental populations. Likewise, the average number of pairwise differences between populations ranged from 60.94 to 80.35 and showed similar patterns to F_{ST} —with the low values corresponding to comparisons inside the same major population

group and high values in the comparisons involving an AFR population. The lowest value was recorded for the comparison of Native Americans (NAM) with the least admixed 1 KG AMR population of Peruvians from Lima (PEL). Average number of pairwise differences within populations ranged from 54.13 to 71.70 with the lowest values in NAM and OCE populations.

Heterozygosity values for the autosomal MHs are listed **Supplementary Table S7** and represented graphically in **Supplementary Figure S7**. Heterozygosity showed variance both among markers (**Supplementary Figure S7A**) and populations (**Supplementary Figure S7B**). Overall mean Heterozygosity values were 0.67 ± 0.09 for autosomal MHs, close to the 0.667 level of a perfectly balanced tri-allelic marker. A single MH, 20pC, gave values lower than 0.5 and the rest had values ranging from 0.49 to 0.81, approaching the 0.5 and 0.75 theoretical limits of bi- and tetra-allelic single-site SNPs. Consistent with their inheritance patterns, X-chromosome MHs showed a lower overall mean Heterozygosity of 0.564 ± 0.118 . In terms

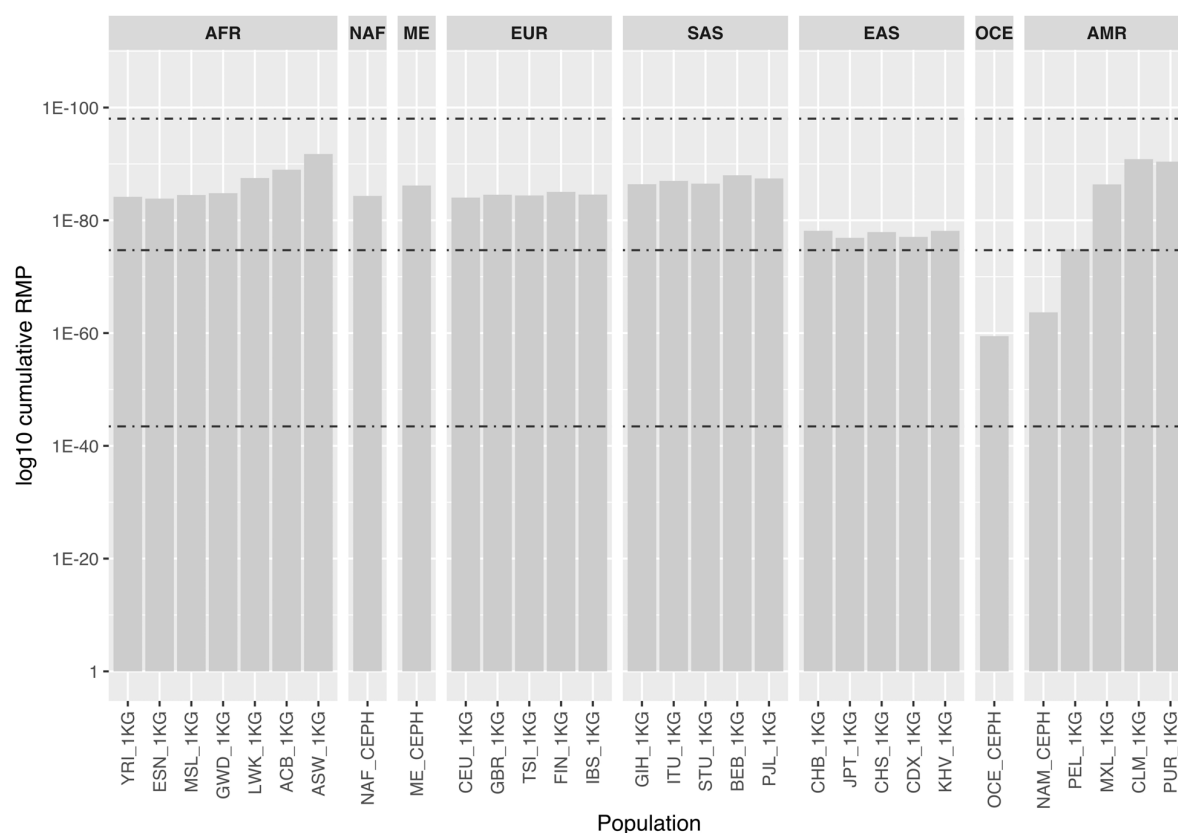


FIGURE 2 | Bar chart represents log10 cumulative random match probability values (i.e., the probability that two individuals share the same profile) for the 30 populations considered, based on the autosomal MH data only. Populations are named and grouped into eight major populations according to **Supplementary Table S5**. Dashed lines represent, from bottom to top, the theoretical values for a panel composed of 102 perfectly balanced bi, tri and tetra-allelic SNPs for comparison: 3.56E-44, 1.98E-75, and 9.32E-99, respectively.

of populations, all showed comparable levels, but NAM and OCE populations had the lowest values, matching patterns of increasing homozygosity with distance from East Africa.

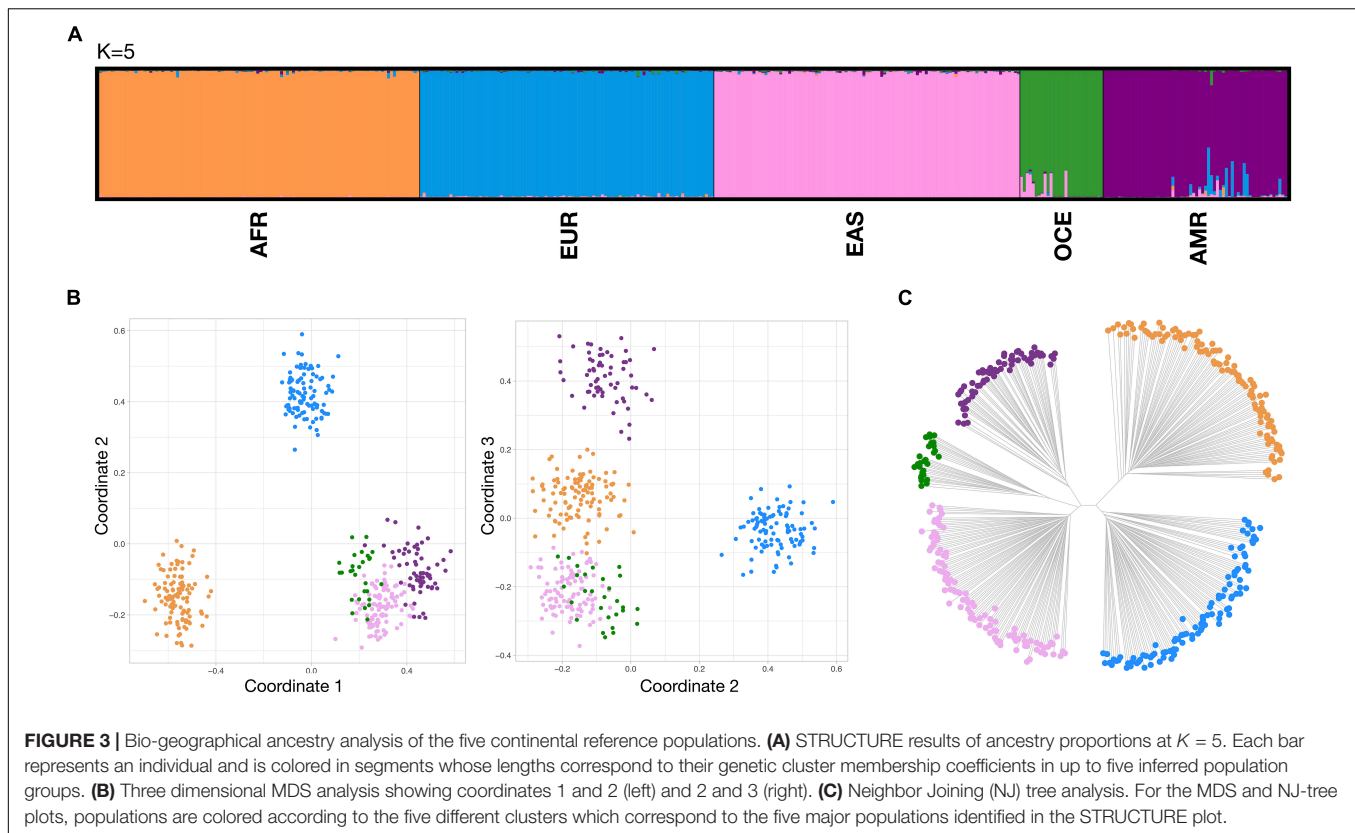
Figure 2 shows cumulative random match probability (RMP) for the 30 populations considering the autosomal MHs. Values for most populations ranged between 1.98E-75 and 9.32E-99, the maximum theoretical values for a panel of 102 tri- and tetra-allelic markers. As a consequence of their lower level of variability, NAM and OCE showed the lowest values. This decrease in discrimination power in such populations should be taken into account when assessing the use of the panel for analyzing distant pedigrees.

Population Differentiation and BGA Inference Performance

Bio-geographical ancestry inference analyses were performed considering genetic information from the 102 autosomal MHs in the panel. In order to minimize possible sample size effects (Onogi et al., 2011), a reference set was constructed by selecting from each major population a single unadmixed population from the total of 30 previously described, as recorded in **Supplementary Table S5**. Additionally, classification was

performed at two levels: (i) five major populations—AFR, EUR, EAS, OCE and AMR—for a first approach at a continental level (herein continental), followed by a second approach when appropriate (ii) with the five main Eurasian populations of NAF, ME, EUR, SAS, EAS to achieve a more detailed analysis of the variability continuously distributed North of the Sahara Desert, forming a natural barrier, and extending across Eurasia from NW to SE of this region (herein NAF-Eurasia). These hierarchical levels are devised so the substructure within NAF-Eurasia can be efficiently detected after a major continental comparison, as suggested in Rosenberg et al. (2002) and Evanno et al. (2005).

Figure 3 compiles results from STRUCTURE, three dimensional MDS and neighbor-joining tree (NJ tree) for the reference populations at the continental level. In STRUCTURE, exploratory runs from $K = 1$ to $K = 8$ (detailed in **Supplementary Figure S8**—left) showed the most consistent cluster patterns at $K = 5$, supported both by the plateau at the mean of estimated Ln probability of data and the peak at Delta K. This five-group differentiation was also observed in the NJ tree, splitting into a 3–2 branch pattern, while some overlap between the OCE and EAS clusters persists in the MDS analysis. Both PSD and pairwise Divergence cumulative values, presented in **Supplementary Figure S9**—top, provided a relatively good balance between



major population groups. **Supplementary Figure S10** includes non-reference populations for the continental level. Unadmixed populations were predominantly assigned to their reference populations in all analysis systems, while admixed populations exhibited the expected patterns, showing mixed co-ancestry membership proportions in STRUCTURE and showing a spread distribution of points between the component clusters in the MDS and NJ tree plots.

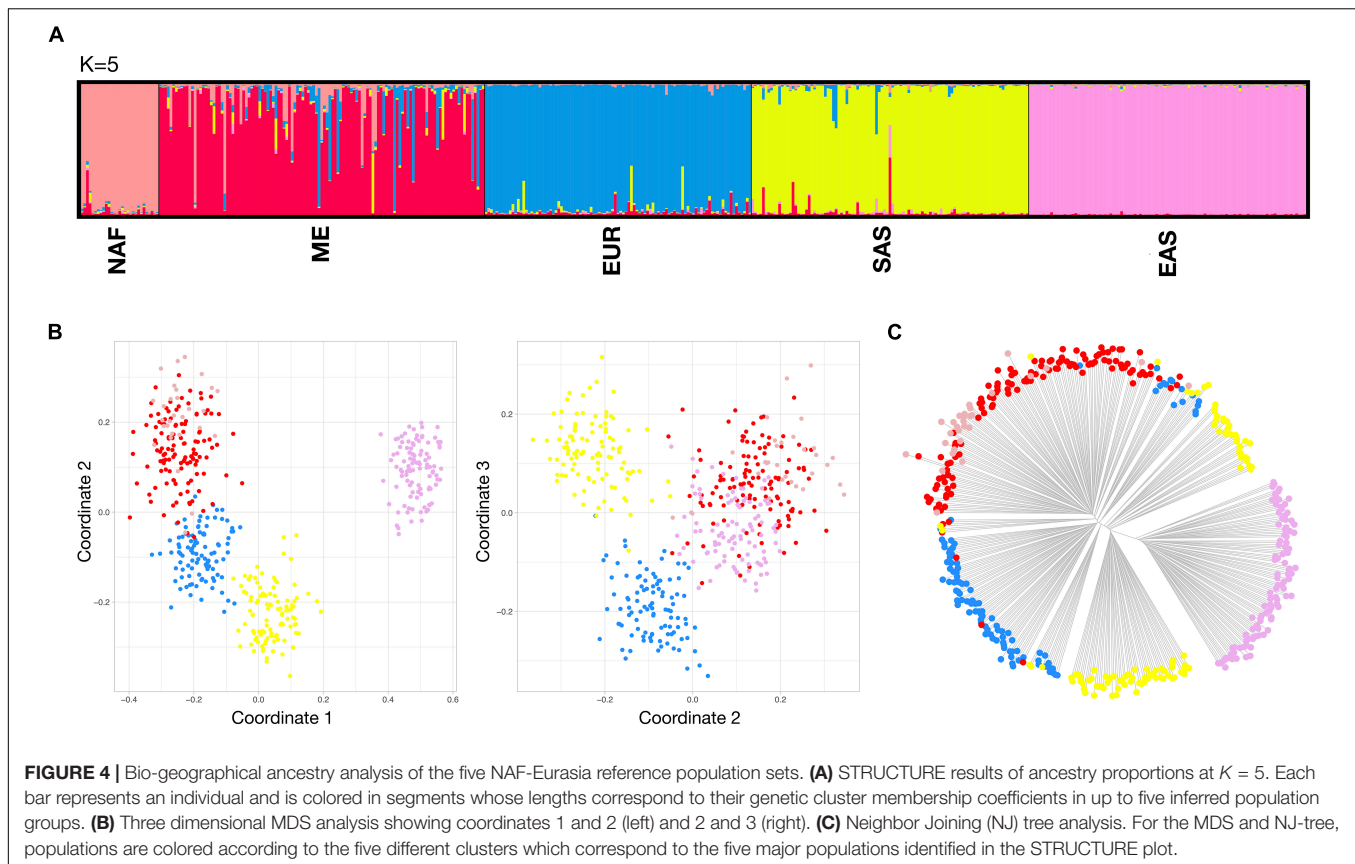
For differentiations at the NAF-Eurasia level, results are compiled in **Figure 4**. Exploratory STRUCTURE runs (**Supplementary Figure S8**–right) showed a higher degree of irregular cluster membership patterns for SAS and ME. Optimal K was selected at 5, taking into account the plateau at the mean of estimated Ln probability of data. However, the Delta K graph showed a peak at $K = 4$, that arguably points to a slightly lower degree of differentiation between NAF and ME, as might be expected given their almost continuous regional distribution in the southern Mediterranean. These two population groups are often considered together for BGA analysis, but further expansion of the reference data, especially for NAF, could enhance the somewhat low levels of contrast found in our analyses. For the MDS analyses, a higher dispersion of the clusters was observed in comparison with the analysis at continental level, with some overlap between NAF and ME. The NJ tree plot shows a distinct EAS branch and a complex hierarchical pattern for SAS, EUR, ME and NAF branches. As expected, cumulative PSD and pairwise Divergence (**Supplementary Figure S9**–bottom) showed lower values and

higher imbalance in these sets of populations in comparison to the more balanced continental differentiation. Pairwise Divergence increased accordingly to geographic distance, with comparisons including EAS reaching the highest values and the lowest values recorded for the closest pairs of NAF-ME, ME-EUR, and EUR-SAS. **Supplementary Figure S11** assembles analysis including non-reference populations at the NAF-Eurasia level. All the tested unadmixed populations showed similar behavior to their reference populations.

Supplementary Figure S12 shows the population assignment analysis of the 42 Sub-Saharan African, European, Central and South Asian and East Asian populations from HGDP WGS against the continental and NAF-Eurasian reference populations, indicating the expected patterns. Central and South Asian populations show a clear frequency cline of admixture between European and East Asian ancestries at the continental level that can also be observed in both the MDS and NJ graphical summaries. At the Eurasian level, these populations show in STRUCTURE a complex mixture of ancestries with a predominant SAS component, despite the fact that none of these populations are located in the Indian sub-continent (unlike the reference populations), and this is reflected in the MDS plot showing a widely distributed set of points centered in the SAS cluster and extending to the NAF, ME, EUR, and EAS clusters.

BGA Inference From Mixtures

Simulated profiles from mixtures 1, 2, and 3 are shown in **Supplementary File S2**, while **Supplementary Table S8**



contains information on both the individual profiles forming the mixtures and the deconvoluted major and minor components. All the haplotypes were assigned correctly to the previously known mixture contributors. Discrepancies between analysts were observed only for the more balanced ratio of 1:3 and were consistent with differences on the degree of risk assumed when assigning the alleles. For example, for MH 2pA analyst 1 assigned haplotypes TAAT/TAAT for the major component and TAGT/— for the minor, considering a possible drop-out of a second allele of the minor component; while analyst two assigned TAAT/— for the major and no haplotypes to the minor —/—; taking into account that it cannot be completely discounted that the TAGT haplotype was from the major component that was showing a high heterozygote imbalance. The most conservative approach—the one from analyst 2 in the example—was used for mixture component BGA inference analysis.

For mixture 1, with the most balanced ratio of 1:3, both the major and minor components resulted in partial profiles after deconvolution, reaching profile completeness percentages of 42.16 and 63.23% respectively. For mixtures 2 and 3, the higher imbalance of the components at ratios 1:5 and 1:7 allowed a full differentiation of the major component. The minor components of mixtures 2 and 3 reached a similar completeness level to that observed in mixture 1 of 42.16 and 43.63%, respectively, despite the fact that ancestry of the individuals contributing to these two mixtures are totally (mixture 2), or partially shared (EUR

component in mixture 3). This is not unexpected as the panel was designed for identification purposes.

Figure 5 shows BGA results for the deconvoluted minor and major components of the mixtures. STRUCTURE analysis revealed the expected ancestry for all deconvoluted profiles. Moreover, estimated co-ancestry proportions of the EUR and AMR for the minor component reached similar levels to the complete profile of the admixed PEL component sample, with a 56.8 and a 55.7% of AMR component, respectively. For MDS, partial profiles from unadmixed samples tended to be spread more away from the reference population cluster, but consistently pointed to the expected ancestry. Admixed partial profile from minor component of mixture 3 appeared almost equidistant from the EUR and AMR clusters, inconsonance with expected.

DISCUSSION

In this study, designed to evaluate extended functionality of MH loci for mixed DNA analysis and compile the necessary population reference data for this purpose, a total of 240 reference HGDP-CEPH individuals of native populations from NAF, ME, OCE, and AMR were analyzed with the panel of 102 autosomal and 11 X-chromosome MHs. Most MHs (109/113) performed well in MPS tests, even when chips were loaded with ~40 sample libraries. Moreover, 99% concordance was achieved between the MH alleles obtained through MPS and the SGDP phased data

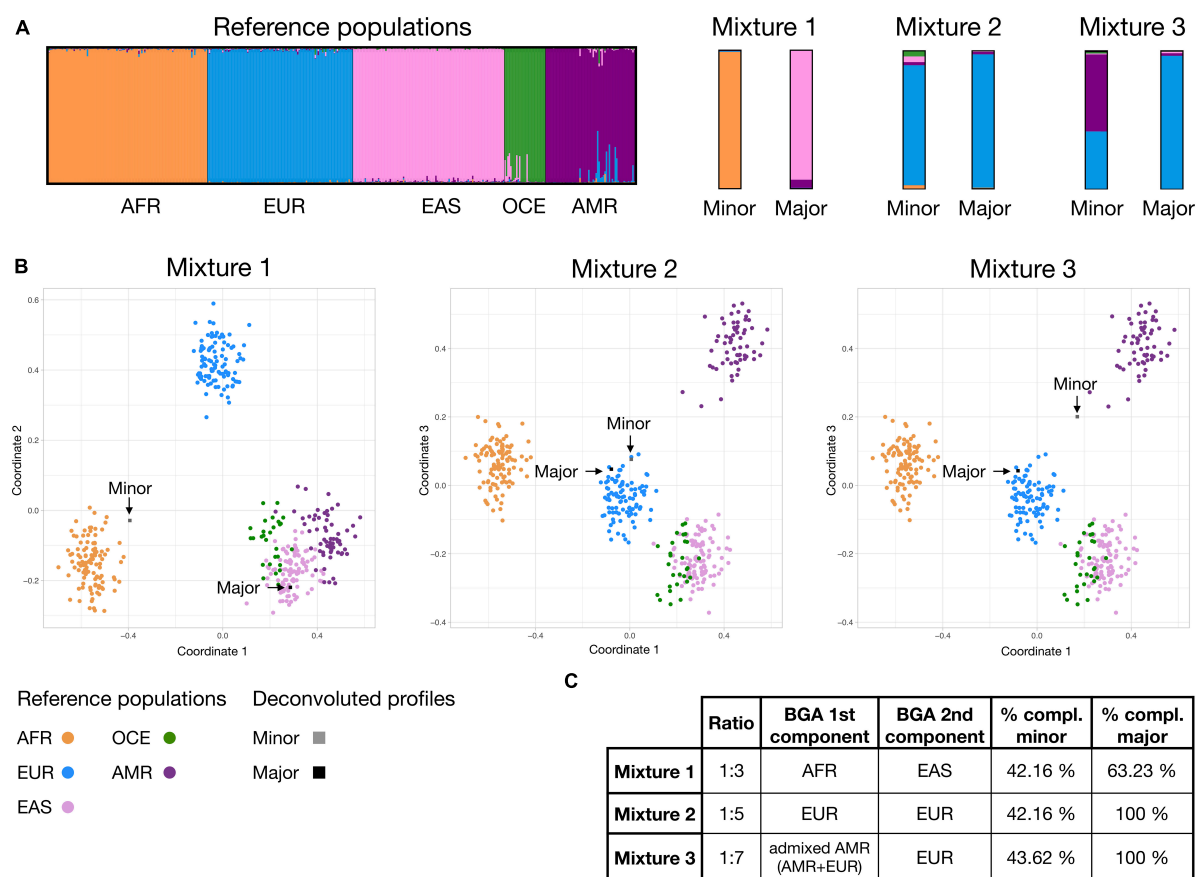


FIGURE 5 | Bio-geographical ancestry inference for the major and minor mixture components in mixtures 1, 2, and 3; classified using the continental reference set presented on **Figure 3**. **(A)** STRUCTURE results of ancestry proportions at $K = 5$. Each bar represents an individual and is colored in segments whose lengths correspond to their genetic cluster membership coefficients in up to five inferred population groups. **(B)** Three dimensional MDS analysis showing coordinates 1 and 2 (for mixture 1) or 1 and 3 (for mixtures 2 and 3). Populations and major and minor components are colored according to the legend. **(C)** Table showing, for each mixture ratio, the expected ancestry of the known components and % of completeness (compl.) of the minor and major deconvoluted MH profiles. Details of the simulated profiles and deconvolution results can be found in **Supplementary File S2** and **Supplementary Table S8**.

used for direct comparisons, while reaching 99.75% concordance with HGDP WGS data. The concordance study revealed some inconsistencies due to the probabilistic phasing algorithm used by both datasets, emphasizing the idea that the phase of the SNPs forming the haplotypes is more accurately derived when detected directly from sequence reads of individual strands, which will encompass all the SNPs in the MH in the same amplicon and using the pipeline developed for the forensic use of the panel. This pipeline outputs the depth coverage of each haplotype and produces profiles similar to those from STRs. Moreover, the pipeline allows for customization of minimum allele frequency and minimum coverage parameters, analogous to the analytical and interpretation thresholds used in capillary electrophoresis analysis. These characteristics aid the interpretation of MH results by forensic experts, especially for mixture analysis, and enhances the utility of the MHs panel we have developed.

Despite the fact that some populations had limited numbers of samples, MHs showed similar degrees of polymorphism to those encountered in the extensive 1 KG dataset. This endorses

the use of the panel for individual identification or kinship testing in the additional worldwide populations analyzed. For this purpose, one of the major advantages of the panel is the small size of the amplicons, that previously outperformed standard STR analysis when dealing with degraded DNA (De La Puente et al., 2019). Compared to SNaPshot (Sánchez et al., 2006; Freire-Aradas et al., 2012; Wang et al., 2016) or commercial MPS SNP panels (Precision ID Identity Panel from TFS, ForenSeq DNA Signature Prep Kit from Verogen) commonly used as supplementary kinship markers, or for degraded DNA analysis, the MHs panel offers a much higher discrimination power due to the increased levels of polymorphism of the markers, while maintaining sensitivity to low level DNA.

At the same time, the new population data we report is a valuable addition to BGA analyses using the panel. The results demonstrate the ability of the panel to differentiate the five major continental groups (AFR, EUR, EAS, OCE, and AMR) and, to a lesser extent, the main sets of populations within Eurasia (NAF, ME, EUR, SAS, EAS). Populations NAF,

ME, and SAS are sited in the middle of variation clines and therefore their differentiation is challenging, especially for NAF and ME regions. To address such challenges, MPS capabilities support much bigger multiplex scales than a typical SNaPshot multiplex assay for SNP genotyping while maintaining forensic sensitivity, allowing a more fine scale geographic resolution in BGA analyses. The MHs panel takes advantage of the higher multiplex capabilities while of MPS using highly polymorphic markers giving high heterozygosity values within populations (allowing individual identification) and high between population differentiation (allowing BGA inference). For these reasons, although not an original criterion for the selection of the component MHs of the panel, the degree of BGA information is similar or superior to that achieved with other custom (Eduardoff et al., 2016; Pereira et al., 2019; Phillips et al., 2019) or commercial MPS panels (Precision ID Ancestry Panel from TFS, ForenSeq DNA Signature Prep Kit from Verogen). The MHs panel considerably exceeds the capabilities of dedicated forensic SNaPshot assays for BGA in use before the advent of MPS (Phillips et al., 2007; Daga-Roszak et al., 2016; De La Puente et al., 2016).

Finally, in this study we began to explore the scope for BGA inference from deconvoluted mixed DNA contributors. Preliminary studies by Oldoni et al. (2017), based on likelihood ratios of profile likelihoods from each population indicated that it is feasible to deconvolute simple two-donor mixtures with skewed mixture ratios, by assigning haplotypes to a major and a minor component and then to infer their ancestry. Here, we confirmed this form of analysis is effective, because it can take advantage of the fact that both the MDS and STRUCTURE methodologies can handle partial profiles. However, extra caution must be used when inferring ancestry for investigative leads when the inferences are made from profiles with high levels of incompleteness. Despite profile deconvolution being both laborious and error-prone, in the near future it is likely that probabilistic genotyping software will be adapted for BGA inference purposes.

Deconvolution of mixed MH profiles is simplified by the absence of stutter artifacts and probabilistic genotyping software can be readily adapted and used for individual identification of the mixture contributors. The ability of the panel to identify the contributors is supported by the fact that, assuming a similar level of informativeness for all MHs [and as shown by the consistent gradient of the RMP slope from Figure 4 in De La Puente et al. (2019)], a ~60% locus completeness of the panel (comparable to the completeness levels shown for mixture 1 deconvolution of the major component) reaches a mean cumulative power of discrimination value across all populations (data from Figure 2) of ~E-39 while a ~40% completeness of the panel (comparable to the minor component) reaches levels of ~E-30 (i.e., comparable to 21 autosomal STRs using GlobalFiler).

CONCLUSION

The MHs panel we have previously developed is found to be even more of a multi-purpose tool for forensic applications than

originally proposed. It is applicable in those forensic cases in which regular STR analysis by itself does not provide an answer or supplementary information is needed. The same component loci of the MHs panel prove to be highly informative for: individual identification with a focus on highly degraded DNA, especially since all amplicon sizes are less than 175 bp; kinship testing; mixed DNA analysis and BGA inference—with indications from our studies that the latter two functions can be combined in simple mixtures. With this in mind, the panel could help to improve identifications in disaster victim identification programs that involve multiple nationalities, where BGA can assist in the first triage of the victims and the selection of the correct allele frequencies for identification through comparisons to surviving relatives. The panel has been fully validated for forensic purposes and can be implemented with both the two main MPS platforms in common use in forensic laboratories: MiSeq and Ion S5, with the latter allowing automated library construction.

DATA AVAILABILITY STATEMENT

The data generated in this manuscript has been deposited at the following public repositories:

- raw reads as fastq were submitted to the European Nucleotide Archive (ENA) under accession number PRJEB39413
- vcf files from Torrent Suite were submitted to the European Variation Archive (EVA) under accession number PRJEB39574.

AUTHOR CONTRIBUTIONS

All authors listed made substantial and direct contributions to the work and approved it for publication. MVL, WP, CP, AC, and MdLP designed the study, developed the ideas and obtained funding for the project. JR-R and AA-C conducted the DNA analysis. MdLP, CP, CX, JA, MACdC, and AG-T analyzed the results. MdLP and CP wrote the manuscript. All authors discussed the results and contributed to the revision of the manuscript.

FUNDING

The studies reported here are supported by MAPA: Multiple Allele Polymorphism Analysis (BIO2016-78525-R), a research project funded by the Spanish Research State Agency (AEI), and co-financed with ERDF funds; and by the European Union's Horizon 2020 Research and Innovation Program under grant agreement no. 740580 within the framework of the Visible Attributes through Genomics (VISAGE) Project. MdLP is supported by a postdoctoral fellowship awarded by the Consellería de Cultura, Educación e Ordenación Universitaria and the Consellería de Economía, Emprego e Industria of the Xunta de Galicia (ED481B 2017/088).

ACKNOWLEDGMENTS

STRUCTURE runs were performed in the FinisTerae II supercomputer from the Centro de Supercomputación de Galicia (CESGA).

REFERENCES

- Almarri, M. A., Bergstrom, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A. S., et al. (2020). Population structure, stratification, and introgression of human structural variation. *Cell* 182, 189–199.e115. doi: 10.1016/j.cell.2020.05.024
- Bennett, L., Oldoni, F., Long, K., Cisana, S., Maddela, K., Wootton, S., et al. (2019). Mixture deconvolution by massively parallel sequencing of microhaplotypes. *Int. J. Legal. Med.* 133, 719–729. doi: 10.1007/s00414-019-02031-2
- Bergstrom, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecsek, P., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science* 367:5012. doi: 10.1126/science.aay5012
- Cann, H. M., de Toma, C., Cazes, L., Legrand, M. F., Morel, V., Piouffre, L., et al. (2002). A human genome diversity cell line panel. *Science* 296, 261–262. doi: 10.1126/science.296.5566.261b
- Chen, P., Deng, C., Li, Z., Pu, Y., Yang, J., Yu, Y., et al. (2019a). A microhaplotypes panel for massively parallel sequencing analysis of DNA mixtures. *Forensic Sci. Int. Genet.* 40, 140–149. doi: 10.1016/j.fsigen.2019.02.018
- Chen, P., Yin, C., Li, Z., Pu, Y., Yu, Y., Zhao, P., et al. (2018). Evaluation of the Microhaplotypes panel for DNA mixture analyses. *Forensic. Sci. Int. Genet.* 35, 149–155. doi: 10.1016/j.fsigen.2018.05.003
- Chen, P., Zhu, W., Tong, F., Pu, Y., Yu, Y., Huang, S., et al. (2019b). Identifying novel microhaplotypes for ancestry inference. *Int. J. Legal. Med.* 133, 983–988. doi: 10.1007/s00414-018-1881-x
- Cheung, E. Y. Y., Phillips, C., Eduardoff, M., Lareu, M. V., and McNeven, D. (2019). Performance of ancestry-informative SNP and microhaplotype markers. *Forensic. Sci. Int. Genet.* 43:102141. doi: 10.1016/j.fsigen.2019.102141
- Daca-Roszak, P., Pfeifer, A., Zebracka-Gala, J., Jarzab, B., Witt, M., and Zietkiewicz, E. (2016). EurEAs_Gplex—A new SNaPshot assay for continental population discrimination and gender identification. *Forensic. Sci. Int. Genet.* 20, 89–100. doi: 10.1016/j.fsigen.2015.10.004
- De La Puente, M., Phillips, C., Xavier, C., Amigo, J., Carracedo, A., Parson, W., et al. (2019). Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems. *Forensic. Sci. Int. Genet.* 45:102213. doi: 10.1016/j.fsigen.2019.102213
- De La Puente, M., Santos, C., Fondevila, M., Manzo, L., Carracedo, A., Lareu, M. V., et al. (2016). The Global AIMS Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs. *Forensic. Sci. Int. Genet.* 22, 81–88. doi: 10.1016/j.fsigen.2016.01.015
- Earl, D., and Von Holdt, B. (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
- Eduardoff, M., Gross, T. E., Santos, C., De La Puente, M., Ballard, D., Strobl, C., et al. (2016). Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM. *Forensic. Sci. Int. Genet.* 23, 178–189. doi: 10.1016/j.fsigen.2016.04.008
- 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. (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
- Freire-Aradas, A., Fondevila, M., Kriegl, A. K., Phillips, C., Gill, P., Prieto, L., et al. (2012). A new SNP assay for identification of highly degraded human DNA. *Forensic. Sci. Int. Genet.* 6, 341–349. doi: 10.1016/j.fsigen.2011.07.010
- Gandotra, N., Speed, W. C., Qin, W., Tang, Y., Pakstis, A. J., Kidd, K. K., et al. (2020). Validation of novel forensic DNA markers using multiplex microhaplotype sequencing. *Forensic. Sci. Int. Genet.* 47:102275. doi: 10.1016/j.fsigen.2020.102275
- Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529
- Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., et al. (2014). Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Forensic. Sci. Int. Genet.* 12, 215–224. doi: 10.1016/j.fsigen.2014.06.014
- Kidd, K. K., and Speed, W. C. (2015). Criteria for selecting microhaplotypes: mixture detection and deconvolution. *Investig. Genet.* 6:1. doi: 10.1186/s13323-014-0018-3
- Kidd, K. K., Speed, W. C., Pakstis, A. J., Podini, D. S., Lagace, R., Chang, J., et al. (2017). Evaluating 130 microhaplotypes across a global set of 83 populations. *Forensic. Sci. Int. Genet.* 29, 29–37. doi: 10.1016/j.fsigen.2017.03.014
- 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 across K. *Mol. Ecol. Resour.* 15, 1179–1191. doi: 10.1111/1755-0998.12387
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., et al. (2016). The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206. doi: 10.1038/nature18964
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110
- Oldoni, F., Hart, R., Long, K., Maddela, K., Cisana, S., Schanfield, M., et al. (2017). Microhaplotypes for ancestry prediction. *Forensic. Sci. Int.* 6, e513–e515. doi: 10.1016/j.fsigen.2017.09.209
- Oldoni, F., Kidd, K. K., and Podini, D. (2018). Microhaplotypes in forensic genetics. *Forensic. Sci. Int. Genet.* 38, 54–69. doi: 10.1016/j.fsigen.2018.09.009
- Onogi, A., Nurimoto, M., and Morita, M. (2011). Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. *BMC Bioinformatics* 12:263. doi: 10.1186/1471-2105-12-263
- Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 26, 419–420. doi: 10.1093/bioinformatics/btp696
- Pereira, V., Freire-Aradas, A., Ballard, D., Borsting, 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., McNeven, D., Kidd, K. K., Lagace, 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
- Phillips, C., Salas, A., Sánchez, J. J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., et al. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic. Sci. Int. Genet.* 1, 273–280. doi: 10.1016/j.fsigen.2007.06.008
- Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van Der Auwera, G. A., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. doi: 10.1101/201178
- Porrás-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., 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

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.581041/full#supplementary-material>

- Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.
- 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, 1402–1422. doi: 10.1086/380416
- Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., et al. (2002). Genetic structure of human populations. *Science* 298, 2381–2385. doi: 10.1126/science.1078311
- Sánchez, J. J., Phillips, C., Børsting, C., Balogh, K., Bogus, M., Fondevila, M., et al. (2006). A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* 27, 1713–1724. doi: 10.1002/elps.200500671
- Sobrinho, B., Brion, M., and Carracedo, A. (2005). SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic. Sci. Int.* 154, 181–194. doi: 10.1016/j.forsciint.2004.10.020
- Sun, S., Liu, Y., Li, J., Yang, Z., Wen, D., Liang, W., et al. (2020). Development and application of a nonbinary SNP-based microhaplotype panel for paternity testing involving close relatives. *Forensic. Sci. Int. Genet.* 46:102255. doi: 10.1016/j.fsigen.2020.102255
- The Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393
- Thomas, N. (2019). microhaplot: Microhaplotype Constructor and Visualizer. Available online at: <https://github.com/ngthomas/microhaplot>
- Turchi, C., Melchionda, F., Pesaresi, M., and Tagliabracci, A. (2019). Evaluation of a microhaplotypes panel for forensic genetics using massive parallel sequencing technology. *Forensic. Sci. Int. Genet.* 41, 120–127. doi: 10.1016/j.fsigen.2019.04.009
- van der Gaag, K. J., de Leeuw, R. H., Laros, J. F. J., den Dunnen, J. T., and de Knijff, P. (2018). Short hypervariable microhaplotypes: a novel set of very short high discriminating power loci without stutter artefacts. *Forensic. Sci. Int. Genet.* 35, 169–175. doi: 10.1016/j.fsigen.2018.05.008
- Voskoboinik, L., Motro, U., and Darvasi, A. (2018). Facilitating complex DNA mixture interpretation by sequencing highly polymorphic haplotypes. *Forensic. Sci. Int. Genet.* 35, 136–140. doi: 10.1016/j.fsigen.2018.05.001
- Wang, Q., Fu, L., Zhang, X., Dai, X., Bai, M., Fu, G., et al. (2016). Expansion of a SNaPshot assay to a 55-SNP multiplex: assay enhancements, validation, and power in forensic science. *Electrophoresis* 37, 1310–1317. doi: 10.1002/elps.201500353

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

Copyright © 2020 de la Puente, Ruiz-Ramírez, Ambroa-Conde, Xavier, Amigo, Casares de Cal, Gómez-Tato, Carracedo, Parson, Phillips and Lareu. 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.



Multi-Indel: A Microhaplotype Marker Can Be Typed Using Capillary Electrophoresis Platforms

Shengqiu Qu^{1†}, Meili Lv^{2†}, Jiaming Xue¹, Jing Zhu^{1,3}, Li Wang⁴, Hui Jian¹, Yuqing Liu¹, Ranran Zhang¹, Lagabaiyila Zha⁵, Weibo Liang^{1*} and Lin Zhang^{1*}

¹ Department of Forensic Genetics, West China School of Basic Medical Sciences and Forensic Medicine, Sichuan University, Chengdu, China, ² Department of Immunology, West China School of Basic Medical Sciences and Forensic Medicine, Sichuan University, Chengdu, China, ³ Department of Forensic Medicine, Sichuan Police College, Luzhou, China, ⁴ Department of Obstetrics and Gynecology, West China Second University Hospital, Sichuan University, Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, Chengdu, China, ⁵ Department of Forensic Medicine, School of Basic Medical Sciences, Central South University, Changsha, China

OPEN ACCESS

Edited by:

Kenneth K. Kidd,
Yale University, United States

Reviewed by:

Chiara Turchi,
Legal Medicine, Italy
Pankaj Shrivastava,
Forensic Science Laboratory, Sagar,
India

*Correspondence:

Weibo Liang
liangweibo@scu.edu.cn;
liangweibo@gmail.com
Lin Zhang
zhanglin@scu.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: 29 May 2020

Accepted: 06 October 2020

Published: 23 October 2020

Citation:

Qu S, Lv M, Xue J, Zhu J,
Wang L, Jian H, Liu Y, Zhang R,
Zha L, Liang W and Zhang L (2020)
Multi-Indel: A Microhaplotype Marker
Can Be Typed Using Capillary
Electrophoresis Platforms.
Front. Genet. 11:567082.
doi: 10.3389/fgene.2020.567082

Since the concept of microhaplotypes was proposed by Kidd in 2013, various microhaplotype markers have been investigated for various forensic purposes, such as individual identification, deconvolution of DNA mixtures, or forensic ancestry inference. In our opinion, various compound markers are also regarded as generalized microhaplotypes, encompassing two or more variants in a short segment of DNA (e.g., 200 bp). That is, a set of variants (referred to herein as multi-variants) within a certain length includes single nucleotide polymorphisms (SNP), insertion/deletion polymorphisms (Indels), or short tandem repeat polymorphisms (STRs). At present, multi-variant is mainly aimed at multi-SNPs. However, the haplotype genotyping of multi-variants relies on single-strand analysis, mainly using massively parallel sequencing (MPS). Here, we describe a method based on a capillary electrophoresis (CE) platform that can directly obtain haplotypes of individuals. Several microhaplotypes consisting of three or more Indels with different insertion or deletion lengths in the range of less than 200 bp were screened out, each of which had at least three haplotypes. As a result, the haplotype of an individual was reflected by the length of its polymorphism. Finally, we established a multiplex amplification system containing 18 multi-Indel markers that could identify haplotypes on each chromosome of an individual. The combined power of discrimination (CPD) and the cumulative probability of exclusion (CPE) were 0.999999999997234 and 0.9984, respectively.

Keywords: multi-indel, microhaplotype, capillary electrophoresis, forensic genetics, paternity tests

INTRODUCTION

Owing to various forensic cases encountered in practice, compound markers have attracted the interest of forensic DNA scientists. Compound biomarkers consisting of two or more variants that occur in short DNA segments of ~200 bp for example, can be regarded as generalized microhaplotypes, including insertion and deletion polymorphisms (Indels) closely linked to short tandem repeat polymorphisms (STRs) (DIP-STR), single nucleotide polymorphisms (SNP)

closely linked to STR (SNP-STR), Indel polymorphisms closely linked to SNP (DIP-SNP), and several Indel polymorphisms linked very tightly in physical positions (multi-Indels) (Castella et al., 2013; Wang et al., 2015; Wendt et al., 2016; Tan et al., 2017; Tan et al., 2018; Oldoni and Podini, 2019).

Haplotypes are presently interpreted in three ways. A statistical inference method was used after separately genotyping each locus, but it could not reflect the true haplotype of individuals (such as PHASE) (Kong et al., 2008; Kidd et al., 2013, 2014). Other ways to interpret include the use of DIP-STR, SNP-STR, DIP-SNP, SNP-SNP, or other compound markers for detection. By designing allele-specific PCR primers, the 3' end of a PCR primer is paired with upstream DIP or SNP alleles. A shared reverse primer is then designed downstream of other STR or SNP markers. Thereafter, two allele-specific sequences are obtained using PCR. The genotype of haplotype markers from an individual can be determined using a capillary electrophoresis (CE) platform and a two-step detection method, but the phase of a haplotype can be unambiguously determined only when the microhaplotypes include two variants (Castella et al., 2013; Cereda et al., 2014; Oldoni et al., 2015; Tan et al., 2017; Liu et al., 2018, 2019; Moriot and Hall, 2019; Oldoni and Podini, 2019; Zhang et al., 2020). Additionally, the main limitation of microhaplotype markers comprising only two variants is the difficulty with increasing polymorphism. A third method relies on single-stranded haplotypes that are resolved by experimental analyzes such as massively parallel sequencing (MPS), which can directly detect the phases of haplotypes on sequenced strands (Borsting and Morling, 2015; Snyder et al., 2015; Wang et al., 2015; Wendt et al., 2016; Chen et al., 2019; Turchi et al., 2019; Zhu et al., 2019; de la Puente et al., 2020; Pang et al., 2020; Sun et al., 2020). However, forensic scientists face many practical challenges due to the complexity of MPS, extensive data processing requirements, and higher costs.

Since the discovery and identification of 2,000 human diallelic Indels in 2002, many studies have found that Indels can serve as important complements to forensic genetic markers in addition to STR and SNP (Weber et al., 2002). Compared to STR, Indel amplicon fragments are shorter, and mutation rates are lower. Compared to SNP, Indels have length polymorphism, which can be directly detected by CE of PCR products. This can be easily achieved in most forensic DNA laboratories without complex detection methods. However, most Indels have only two alleles, the polymorphisms are relatively poor and the discriminatory power is relatively lower than that of STR. The present study considered a marker containing at least two Indel loci in a short segment of DNA (namely multi-Indel), as a new microhaplotype. This marker not only increased Indel polymorphism, but also retained the advantages of SNP and STR. Since Indels are markers with length polymorphism, we selected Indel loci with different allele lengths to form a microhaplotype that was directly detectable by CE. According to length polymorphism, it unambiguously reflected the phases of haplotypes from individuals.

Previous studies of multi-Indels have been limited to increasing polymorphism. However, as the length of an insertion or deletion in alleles of an Indel is not specific, some polymorphism information is lost (Huang et al., 2014; Sun et al., 2016). Additionally, individual haplotypes have been statistically inferred after genotyping each Indel locus, which does not reflect the true haplotype of an individual (Zhao et al., 2018). In the present study, we proposed a strategy based on a CE platform to obtain accurate haplotypes of individuals, and constructed a multiplex amplification system containing 18 multi-Indel markers to improve the discrimination power of Indels.

MATERIALS AND METHODS

Ethics

The participants provided their written informed consent to participate in this study and for participants under the age of 16, the legal guardian provided written informed consent to participate. All samples were obtained under the supervision of the Ethical Committee of the Sichuan University (KS2019042).

Samples and DNA Extraction

This study included 335 samples of EDTA blood collected from the Sichuan Province, China. The samples were collected under written informed consent from 170 unrelated Sichuan Han individuals, 30 unrelated Sichuan Yi individuals, and 83 parent-child pairs. Notably, 134 samples were from 17 unrelated extended families that descended from 83 parent-child pairs; thus, some parent-child pairs had the same alleged parent or alleged child. We extracted DNA using the BioTeke DNA kits (BioTeke Corp., Beijing, China) as described by the manufacturer. The collected DNA was quantified using the NanoDropTM 1000 spectrophotometer (Thermo Fisher Scientific Inc., Waltham, MA, United States).

Selection of Multi-Indel Markers

Candidate Indels were selected from 208 samples including 103 Han Chinese in Beijing, China (CHB) and 105 Southern Han Chinese in China (CHS) in the 1000 Genomes Project phase 3 using VCFtools¹ (Sudmant et al., 2015) that met the following criteria: being biallelic, minor allele frequency (MAF) >0.1, located in a non-coding region or intron, allele length of each Indel ranged from 1 to 30 bp; one multi-Indel comprised at least three Indels, physical distance between selected Indels in one multi-Indel marker was <200 bp, alleles had different lengths, and the length of any allele was not equal to the sum of the lengths of the other two or more alleles (each theoretical haplotype has a unique amplicon length), different multi-Indel markers were >10 Mb apart if on the same chromosomal arm, no other Indel variation had MAF >0.005 within this range, the haplotype frequency calculated by Haploview was ≥ 3 , and at least three haplotypes had a frequency of ≥ 0.05 (Barrett et al., 2005).

¹https://vcftools.github.io/man_latest.html

Genotyping Multi-Indels

Primer Design and Optimization

We designed PCR primers using the online tool Primer3web² according to the following criteria: PCR product size, 70–250 bp; T_m values, 55–62°C, and GC content, 30–60%. Potential secondary structures between obtained primer pairs (including formation of primer dimers and hairpin structures, were examined using AutoDimer³, and specific primers were identified using Primer-BLAST⁴. All primer pairs were then assigned according to the predicted amplicon length, and one of the primer pairs was labeled with the fluorochromes, FAM, HEX, TAMRA, and ROX. All primers were synthesized (Thermo Fisher Scientific Inc.) then purified using high performance liquid chromatography (HPLC). Subsequently, we used 1–5 samples to perform a singleplex PCR reaction for each microhaplotype locus. CE was used to detect the PCR products of each microhaplotype locus. And the homozygous samples were amplified using the corresponding primers that are not labeled with fluorescent dyes for Sanger sequencing verification. The size of each locus was examined and compared with the size of CE to determine the electrophoretic mobility of each allele.

Multiplex PCR Amplification

In multiplex PCR amplification, the initial each primer concentration was 0.2 μM. Then this multiplex amplification system was then optimized based on primer concentrations and peak heights. We programmed the thermal cycler according to the manufacturer's instructions. In order to minimize the influence of the annealing temperature of the multiplex system, 18 multi-Indel markers were multiplex amplified under different annealing temperature gradients (56.9, 57.6, 58.4, and 59.1°C) and different PCR cycles (25, 27, 29 and 32) with 1 ng of control DNA F312. According to the optimized and relatively balanced genotyping profiles, the optimal annealing temperature and optimal cycle number of our system were finally determined. The final reaction volume of 10 μL included 5 μL of 2× Multiplex PCR Master Mix (Qiagen GmbH, Hilden, Germany), 2 μL of primer mixture, 1 μL of target DNA (1 ng/μL), and 2 μL of RNase-free water. The samples were amplified by PCR using the GeneAmp 9700 PCR System (Applied Biosystems, Foster City, CA, United States) under the following cycle conditions: 95°C for 15 min, then 27 cycles of 30 s at 94°C, 90 s at 58.4°C, 60 s at 72°C, and hold at 60°C for 60 min. All 335 samples were genotyped using the 18 multi-Indel markers in one multiplex PCR reaction.

Detection and Analysis

The PCR products were detected using the ABI 3500 Genetic Analyzer (Applied Biosystems) and a preloaded AGCU E5 dye fragment analysis run module. Samples were prepared for CE by mixing 1 μL of the PCR products with 8.9 μL of Hi-Di formamide (Applied Biosystems) and 0.1 μL of SIZ500 size standard (AGCU ScienTech, Jiangsu, China). Samples were

injected at 1.2 kV for 5 s and resolved by electrophoresis at 15 kV for 1,310 s in Performance Optimized Polymer-4 (POP-4 polymer) (Applied Biosystems). Genotyping data were then analyzed using the GeneMapper™ ID Software v3.2.1 (Applied Biosystems), with an allele peak threshold of 100 relative fluorescence units (RFU).

Allele Nomenclature

Since a nomenclature system for multi-Indel markers has not been standardized and they are essentially a type of microhaplotype, we named the multi-Indel markers in this study according to those suggested by Kidd (2016). We labeled the smallest of their alleles as 0 according to the size of the amplicon in each multi-Indel marker, and if other alleles were N bp larger than the smallest allele, these were called N. New alleles identified in this study were also named according to their length (Huang et al., 2014).

Sensitivity Study

We evaluated the sensitivity of our multiplex system. Serially diluted control DNA F312 (2 ng μL⁻¹ stock) (Beijing Microread Genetics, Beijing, China) was amplified in triplicate with quantities of 1, 0.5, 0.25, 0.125, and 0.0625 ng. These samples were processed under the same reaction conditions described above.

Mixture Studies

We assessed the ability of our multiple system to detect DNA in mixtures of several ratios of female and male DNA. Mixtures of female/male DNA samples (control DNA F312 and M308, Beijing Microread Genetics, Beijing, China) in ratios of 19:1, 9:1, 4:1, 3:1, 1:1, 1:3, 1:4, 1:9, and 1:19 ng were amplified in our multiplex assay in triplicate.

Degradation Study

We simulated several degraded samples that were amplified and resolved by electrophoresis as described above to evaluate the ability of our multiple system to detect DNA in degraded samples. The control DNA M308 was ultrasonically degraded by 0, 100, 200, 300, or 400 cycles of 200 W for 10 s per cycle with 4-s intervals between cycles. The extracted DNA from the EDTA blood was ultrasonically degraded by 0, 200, 400, and 600 cycles of 400 W for 10 s per cycle, with 4-s intervals between cycles.

Statistical Analysis

Each allele was considered as one haplotype. The allele frequency was the available haplotype frequency. The forensic parameters allele frequencies, power of discrimination (PD), power of exclusion (PE), typical paternity index (TPI), and observed heterozygosity (Ho), and the exact tests of the Hardy–Weinberg equilibrium (HWE) were calculated using a modified spreadsheet within PowerStat v1.2 (Promega Corp., Madison, WI, United States) (Zhao et al., 2003). Linkage disequilibrium (LD) in pairwise loci were analyzed using GENEPop (Rousset, 2008). The effective number of alleles (A_e) was calculated based on the formula proposed by Kidd and Speed (2015).

The paternity index (PI) is the likelihood ratio of the probability that an alleged father with the DNA result is the

²<http://primer3.ut.ee/>

³<http://www.csl.nist.gov/biotech/strbase/AutoDimerHomepage/AutoDimerProgramHomepage.htm>

⁴<https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi>

biological father of the child and the probability that the random man is the biological father of the child. The PI was calculated based on LR principles according to the International Society for Forensic Genetics (ISFG) (Gjertson et al., 2007). The combined paternity index (CPI) was equivalent to the product of PI for all multi-Indel markers tested in each parent–child pair.

RESULTS

Marker Selection and General Information

We screened candidate Indels that met the inclusion criteria from the 1000 Genomes Project database. The filter of biallelic Indels with $MAF > 0.1$ and the allele length variation of each Indel from 1 to 30 bp resulted in 629,402 candidates, which were then filtered according to differences between allele lengths of all loci within a physical distance of <200 bp, and 26,092 potential haplotype markers remained. These were filtered according to each haplotype containing at least 3 Indels, which left 1,642 candidates. Loci in gene coding regions and those positioned <10 Mb apart on the same chromosomal arm were excluded. According to the number and the frequency of haplotypes calculated by Haploview and filtering according to our primer design criteria, only 52 candidates remained. Finally, 18 candidate multi-Indel markers containing 54 Indel loci were genotyped in one multiplex panel after removing loci for which correct genotype results could not be obtained due to long homopolymer structures or 2–15 nucleotide tandem repeats. **Table 1** shows the general information of the 18 multi-Indel markers, and **Supplementary Table S1** shows the haplotype frequency calculated by Haploview.

Multiplex Assays

Before performing the multiplex amplification, we verified the amplification of the primer pairs at each marker by performing singleplex PCR reaction and detection by CE. The size of the allele was determined based on the results of Sanger sequencing of the corresponding homozygous samples. The CE detection results of the singleplex PCR reaction and the Sanger sequencing of the corresponding markers were shown in the **Supplementary Data**. After the development and optimization of this multiplex panel, 18 microhaplotype markers were successfully amplified in a single PCR reaction, and the optimal temperature was determined as 58.4°C , the optimal cycle number was determined as 27, following the optimized PCR conditions presented in section “Multiplex PCR Amplification.” After one PCR reaction and the next CE run, 18 multi-Indel markers containing 54 Indel loci were genotyped per DNA sample. The results showed that 18 complete profiles were detected in each test sample. **Figure 1** shows an example of capillary electropherogram obtained by genotyping the control DNA F312. **Supplementary Figure S1** shows a capillary electropherogram of the control DNA M308, and **Table 1** includes information about the sequences and concentrations of all primers in the system.

Sensitivity Study

The sensitivity of our multiplex assay was tested with control DNA F312 serially diluted to template amounts of 1, 0.5, 0.25, 0.125, and 0.0625 ng. Each template amount was amplified three times. Sample inputs >0.125 ng consistently generated full profiles (**Figure 2**) when amplified for 27 PCR cycles and when the threshold for allele calls was 100 RFU. As the template DNA concentration was gradually reduced from 1 to 0.125 ng, the average detected peak height shifted from 4,144 to 351 RFU. When the template DNA F312 decreased to 0.0625 ng, profiles were partial and an average 91.36% of the allele was detected with an average peak height of 212 RFU. Therefore, our multiplex system obtained reliable profiles at a threshold of 100 RFU above a DNA concentration of 0.125 ng.

Mixture Studies

Template DNA (1 ng) comprising a mixture of control DNA F312 and 19:1, 9:1, 4:1, 3:1, 1:1, 1:3, 1:4, 1:9, and 1:19 ratios of M308 was tested in triplicate. All unique minor profiles were called at ratios of 4:1, 3:1, 1:1, 1:3, and 1:4 (**Figure 3**), and minor alleles were called at averages of 80.56% and 91.67% at ratios of 9:1 and 1:9, respectively. The unique minor profile in the mixture was called at averages of 69.44 and 94.44% at ratios of 19:1 and 1:19, respectively.

Degradation Study

We simulated the degradation of the control DNA M308, and DNA extracted from fresh EDTA blood to determine the effects of sample degradation. After the control DNA M308 was disrupted using 0–400 ultrasound cycles of 200 W, full profiles were obtained using a peak height analysis threshold of 100 RFU. However, the average peak height gradually decreased as the number of cycles increased (**Figure 4**). Only 83% of the alleles were called from the DNA sample extracted from fresh EDTA blood (a conventional case sample), after 200 ultrasound cycles at 400 W, and after 400 and 600 cycles, 33.33 and 23.33% of alleles were called, respectively (**Figure 5**).

Statistical Analysis

We genotyped 200 unrelated individuals from Sichuan using our panel of 18 multi-Indel markers containing 54 Indel loci multiplex systems. **Supplementary Table S2** shows their genotype profiles. The mean distance between the outermost Indels of each multi-Indel was 58 (5–142) bp. The average amplicon length was 182 (107–326) bp. The actual and theoretical amplicon sizes differed in seven multi-Indel markers. Our multiplex detected 77 specific amplicons (that is, 77 haplotypes) in 200 Sichuan individuals. One of these, mh01zl001, was monomorphic in the surveyed population, so we excluded this locus from further statistical analysis. We found 2, 3, 4, 5, 7, 9, and 10 haplotypes in 3, 4, 3, 4, 1, 1, and 1 multi-Indel markers respectively. **Supplementary Table S3** lists the alleles of 17 multi-Indel markers and their frequencies. The mean and median values of A_e for these 17 loci were 2.83 and 2.92, respectively (**Figure 6**).

TABLE 1 | The general information of 18 multi-Indel markers.

Microhaplotype	GRCh37	rs-Number dbSNP	Extent in bp	Allele1/Allele2	Insertion allele length	Primer sequences (label)	Primer concentration (μ M)	Theoretical amplicon size (bp)
mh01zl001	2029533	rs368828322	16	A/AG	1	GGCGGGGTGAATAGTTTGAC (ROX)	1.569	150, 151, 160, 161, 179, 180, 170, 169
	2029539	rs372567620		A/AAGGTCAGAGC	10	TCAGTAAACAACCCCTGCCT		
	2029549	rs148361309		C/CAGGTGACCAGGAGTGACTA	19			
mh01zl002	100194878	rs55796544	25	C/CCT	2	TGTGCTCCTCTTTCTCACTAGT (TAMRA)	0.392	106, 107, 109, 110, 103, 104, 105, 108
	100194896	rs67810269		C/CTGTA	4	TTAAGATGGTCAGGGCATCAG		
	100194903	rs71075445		C/CT	1			
mh02zl001	30981778	rs142363578	142	C/CTTCT	4	CCCTTACTCCCTCTCGTCTTC (TAMRA)	0.196	196, 198, 200, 202, 215, 217, 213, 211
	30981829	rs144117237		T/TTC	2	GGAGGGATGAAGGGAGGC		
	30981920	rs148016741		C/CCCTCCCTCCCTCCCT	15			
mh02zl003	212161558	rs575990766	85	A/ACATATGTATG	10	ACTAAAGCCTGTATATGTAGCCT (ROX)	1.569	216, 232, 238, 242, 212, 222, 226, 228
	212161604	rs141442566		A/ATACATATGTATGTATG	16	CCCAGTATCATTCTCTATCTCTGC		
	212161643	rs66617012		A/ATAAG	4			
mh03zl001	73878996	rs34404453	47	T/TC	1	TGATTCTTCTTACTCCTCCAAAG (HEX)	0.196	124, 125, 126, 129, 120, 121, 123, 128
	73879030	rs149171688		A/AAATAT	5	GGCAACAGAATAAGACTCCGTT		
	73879043	rs34483288		A/AATT	3			
mh03zl002	87352688	rs200679094	48	G/GAAATCTAAATAT	12	ACCATCTACATTTTCCCTGTAAA	0.588	118, 123, 130, 131, 119, 122, 134, 135
	87352693	rs370444413		A/AGGTG	4	GCTGGGTCATCGCCATTTT (TAMRA)		
	87352736	rs74604190		T/TA	1			
mh03zl003	163670527	rs80013016	102	T/TG	1	AGCTAGAGGTGTGTAGGCAA (FAM)	0.196	183, 184, 194, 195, 199, 210, 211, 200
	163670601	rs372207681		C/CAGGTGCCAGCT	11	GCTTCTGCGTGACACTGC		
	163670629	rs111507567		G/GCTGCTGCTTTGGGCAA	16			
mh04zl001	18391312	rs11282557	17	G/GACAGTATTT	9	CCTTGTTGCTGCAGTAGAAAAT (ROX)	0.147	135, 144, 146, 155, 158, 147, 156, 167
	18391316	rs58595156		G/GGAAAAATTGCT	11	TGATCACTTAAGTTCGATGAAAGAA		
	18391329	rs372089291		A/ATTCTCCTAAATT	12			
mh04zl003	187124231	rs66502037	82	C/CAT	2	TGCGCATATACACATACATAGATG	0.098	150, 153, 155, 157, 151, 146, 152, 148
	187124238	rs77222977		G/GCACA	4	ATGTGTATGGGGTTGTGCAC (TAMRA)		
	187124313	rs71871946		T/TCATA	5			

(Continued)

TABLE 1 | Continued

Microhaplotype	GRCh37	rs-Number dbSNP	Extent in bp	Allele1/Allele2	Insertion allele length	Primer sequences (label)	Primer concentration (μ M)	Theoretical amplicon size (bp)
mh07zl001	57322877	rs71053237	103	C/CTAAATGAT	8	TTGTGGGGTGGCGGAAG	0.392	172, 179, 180, 182, 169, 171, 174, 177
	57322974	rs72447238		T/TATA	3	GCACTGGATGGCACTCTTTT (HEX)		
	57322980	rs71053238		A/AAT	2			
mh10zl001	7140235	rs145059123	72	C/CA	1	ACACATTCACACATTCATTAGACA (ROX)	1.569	215, 216, 218, 221, 224, 217, 222, 223
	7140259	rs539040996		T/TAGACAC	6	TGGTGTGTGTGTATGCTAGTG		
	7140307	rs34860860		T/TCA	2			
mh10zl002	113582580	rs143378119	38	G/GAGAATACATTA	11	GAACAGAGTGTATCCATTTTCT	0.098	110, 112, 115, 121, 126, 137, 143, 148
	113582616	rs370632025		T/TTATGG	5	TCAGGCCAATCACACGTG (FAM)		
	113582618	rs571358799		C/CAGGACTGGAAGGAGAATACAAT	22			
mh13zl001	25442007	rs58303500	26	G/GCAT	3	GCCGTGATCTTCTCTGGGAA (FAM)	0.196	217, 218, 220, 224, 214, 215, 221, 223
	25442012	rs34156563		A/AT	1	TAATGAGGGCTGGGGTGTTT		
	25442033	rs67523118		A/ACTTATT	6			
mh18zl001	61672654	rs377195018	16	A/AGAGGTGGGACC	11	GAATGCCGTCTTCCACCAAA (FAM)	0.196	147, 153, 162, 164, 149, 151, 158, 160
	61672667	rs59925455		T/TTGGG	4	AGGGGCAAGGTAGTTCTCTG		
	61672670	rs201823781		G/GAT	2			
mh19zl001	490362	rs138906215	87	A/AAAAG	4	GGCGACAAGAGTGAAACTC	0.784	155, 157, 159, 162, 153, 160, 164, 166
	490414	rs35679623		A/ACT	2	CTCAGCGTGAACAAAGAGTG (HEX)		
	490449	rs11283323		C/CAATACTG	7			
mh19zl002	56039183	rs543507020	85	A/ATGCACACACTCACAATTGCACACACG	26	TGGACACACGCACACTGG (FAM)	0.588	177, 181, 201 203, 175, 179, 205, 207
	56039213	rs559033587		A/ACACT	4	TCTGTGCAAGTGTGAATGTCTG		
	56039268	rs36127315		G/GCA	2			
mh21zl001	21261329	rs562677243	5	T/TTA	2	AGCAATGTGTTTACAGATACCA	0.686	164, 167, 174, 176, 165, 166, 173, 175
	21261333	rs576365249		C/CG	1	AGGCCATGGAGAGGAGTAGA (TAMRA)		
	21261334	rs373980639		G/GGGGACATT	9			
mh21zl002	42384738	rs147144385	55	T/TA	1	ACACATTCTCAAGCACTCACA (HEX)	0.049	147, 149, 150, 151, 144, 145, 146, 148
	42384760	rs200218606		T/TCACA	4	GGTGGGAGATGTGAATGTGT		
	42384793	rs138205093		A/ACT	2			

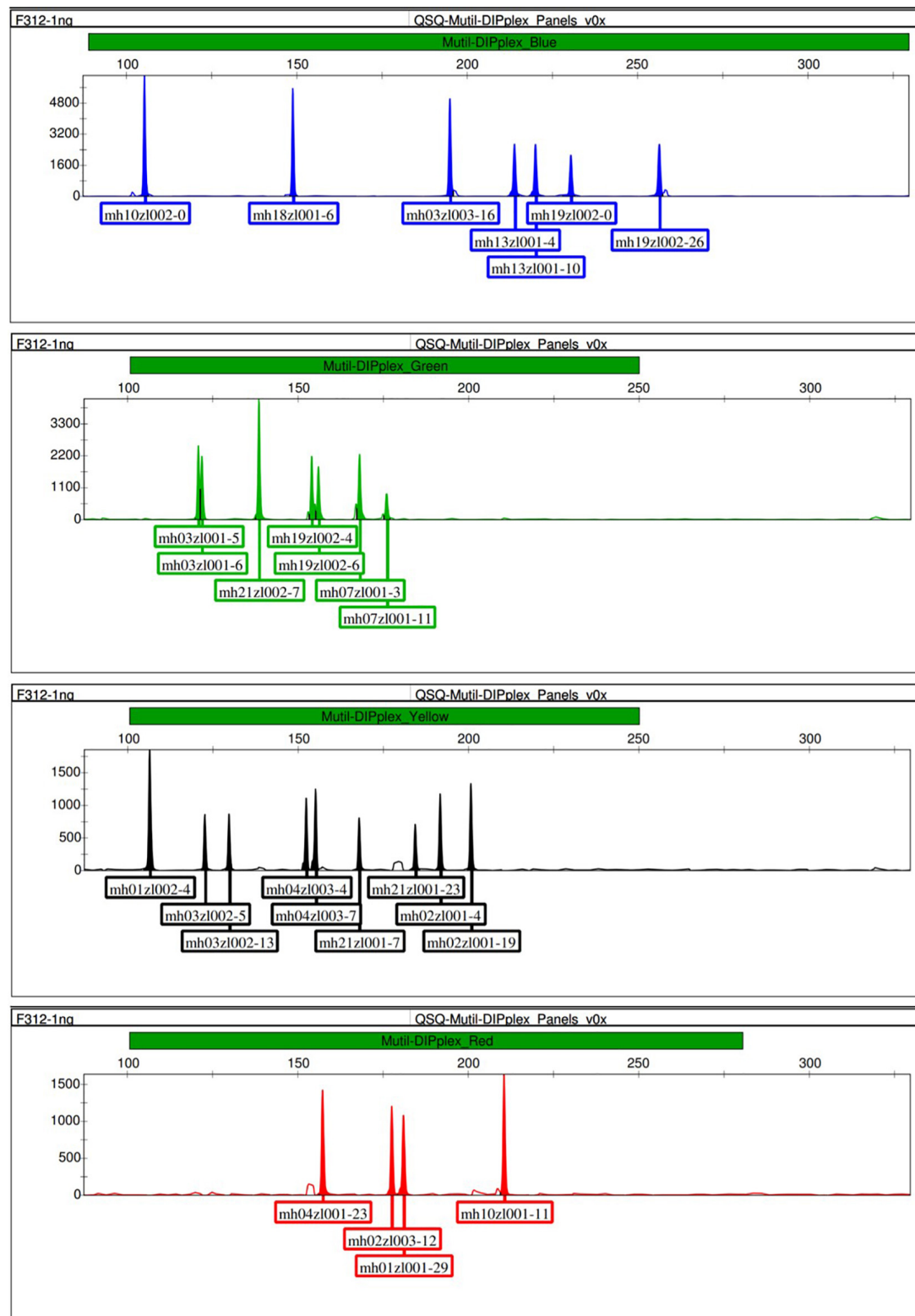


FIGURE 1 | Representative electropherogram of control DNA F312 amplified at 1 ng.

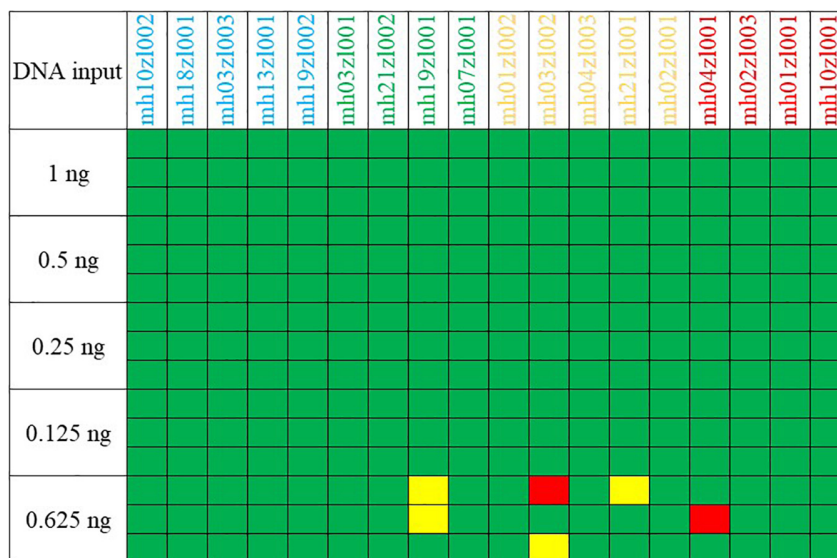


FIGURE 2 | Results of a sensitivity study using serially diluted control DNA F312. Green boxes, no allele drop-out; red boxes, no alleles recovered; yellow boxes, only one of two expected heterozygote alleles was called.

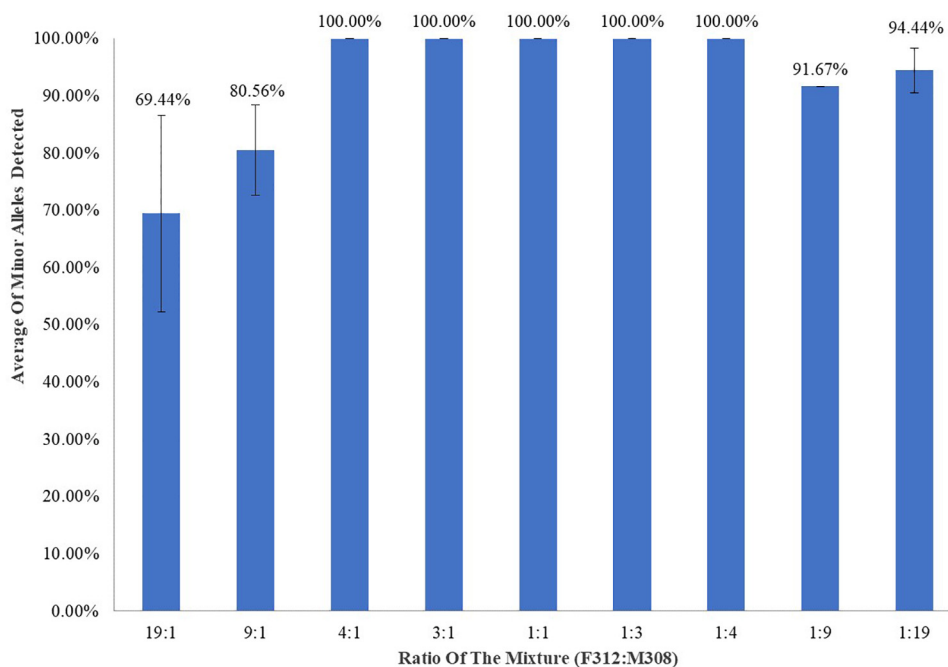


FIGURE 3 | F312/M308 DNA amplification using 1 ng of total DNA and assessed in triplicate.

We also tested each locus for conformity to the HWE model and for potential LD. The threshold p value for the HWE test was set at 0.00037 after the Bonferroni correction, and no deviations from linkage equilibrium were significant between pairwise loci after the Bonferroni correction ($p > 3.68 \times 10^{-4}$; **Supplementary Table S4**). **Table 2** lists the PD, PE, Ho, PM, PIC, TPI, and p values for HWE of the 17 multi-Indel markers. The average PD value was 0.7585 (range, 0.5146–0.9469). The average PE

value for the 17 loci was 0.591 (range, 0.0888–0.5535). The Ho was 0.355 to 0.775, and combined PD and combined PE were 0.99999999997234 and 0.998414249965817, respectively.

Application in Paternity Testing

We analyzed 83 parent–child pairs and calculated PI using genotype data using the multi-Indel multiplex panel. **Supplementary Table S5** shows the genotypes of 83 parent–child

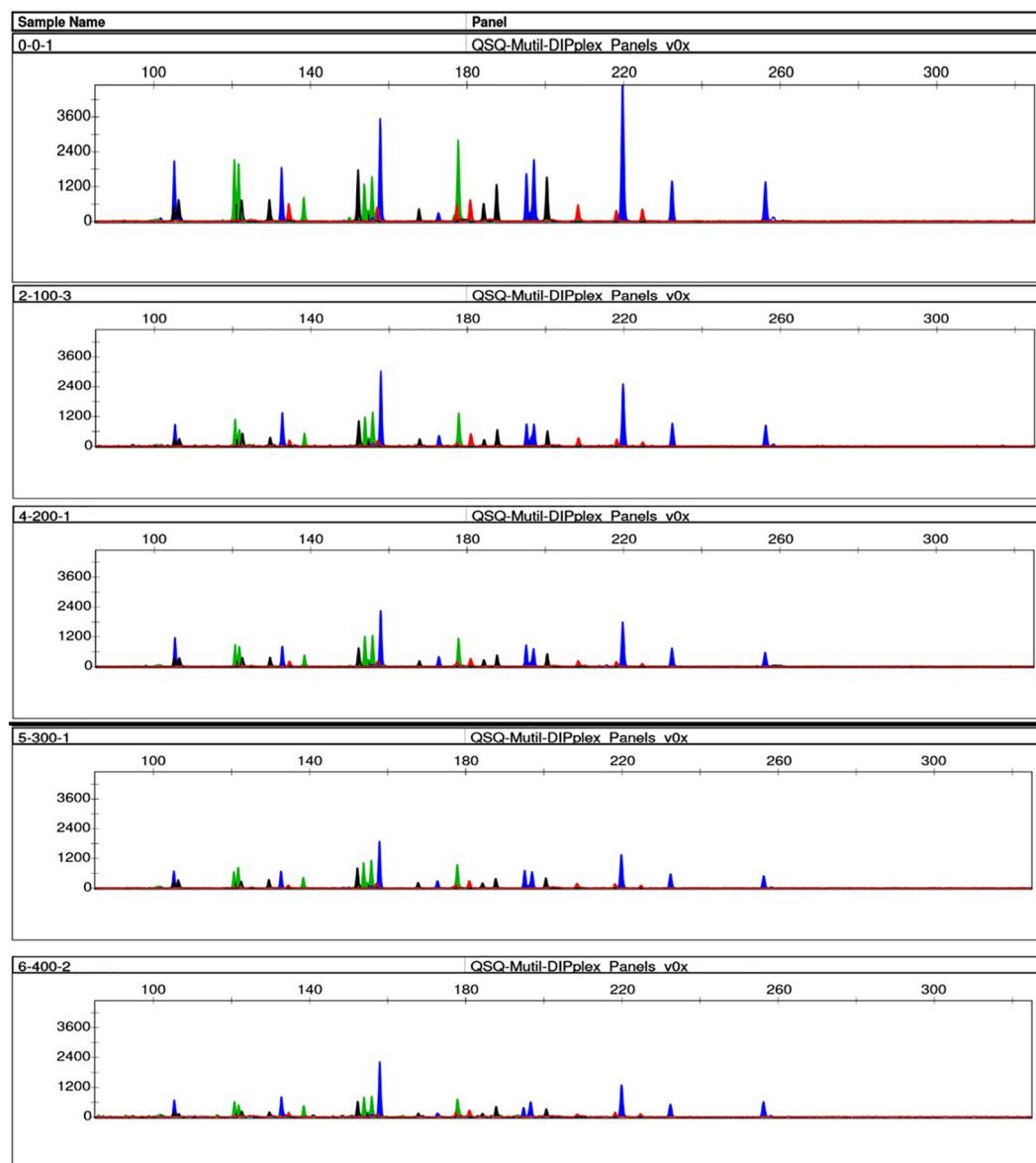


FIGURE 4 | Effects of ultrasound on degraded control DNA M308. We degraded control DNA using 0 (**top**), 100, 200, 300, and 400 (**bottom**) ultrasonic cycles of 200 W for 10 s per cycle, with 4 s between cycles.

pairs and the specific PI per locus and CPI per parent–child pair. The allele frequency of 17 multi-Indel markers was obtained separately from the 200 unrelated individuals. All the parent–child pairs conformed to the Mendelian laws of inheritance. No mutation or recombination was found in any of the multi-Indel markers from 83 parent–child pairs. Overall, the CPI in 83 parent–child pairs determined by the panel of 17 multi-Indel markers averaged $2.82066955485148 \times 10^6$ (range, $0.58394420522483 \times 10^3$ to $5.06111014257473 \times 10^7$). Fourteen parent–child pairs had a CPI below 10,000, which did not support a biological parent–child relationship between them. However, their CPI were >0.0001 , so a biological parent–child relationship cannot be excluded. The number of loci

would need to be increased, or combined with STR kits to clarify this situation.

DISCUSSION

Multi-variant is slightly different from the traditional microhaplotype. We believe that a set of all variants including SNP, Indel and STR within a specifically short length can be considered as generalized microhaplotypes. Only microhaplotypes containing two SNP can presently be genotyped on the CE platform due to limitations of the system (Zhang et al., 2020). Therefore, we selected Indels from the 1000 Genomes

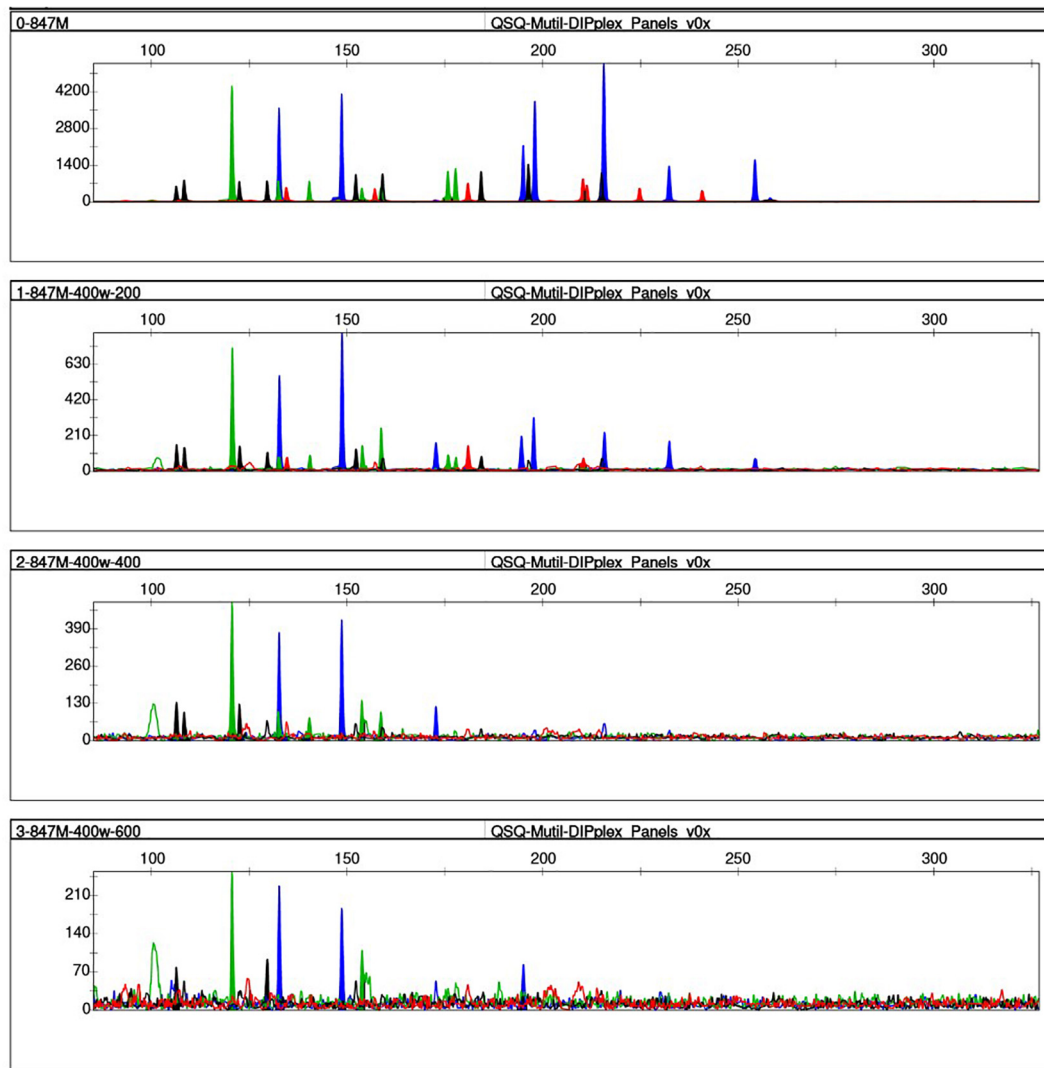


FIGURE 5 | Effects of ultrasound on DNA extracted from fresh EDTA blood. We degraded DNA using 0 (**top**), 100, 200, 300, and 400 (**bottom**) ultrasonic cycles of 200 W for 10 s per cycle, with 4 s between cycles.

Project as the basis for constructing microhaplotypes that could be analyzed using this platform. The human Indel mutation rate ranges from 0.53 to 1.5×10^{-9} per base per generation (Kondrashov, 2003; Lynch, 2010; Campbell and Eichler, 2013; Ramu et al., 2013; Besenbacher et al., 2015; Zhao et al., 2018). This mutation rate is one order of magnitude lower than that for SNP and five orders of magnitude lower than that for STR. Therefore, Indels combine the advantages of SNP and STR. Multi-Indels increase their polymorphism while retaining the advantages of Indels. We used Haploview to initially screen haplotype frequency. Since Haploview can only recognize biallelic alleles and biallelic loci are the most prevalent in Indels, this study investigated only biallelic Indels. We extracted 2,052,970 biallelic Indels from 22 autosomes in the 1000 Genomes Project using VCFtools. We further restricted the alleles according to their length. In theory, different amplicon lengths represent different

haplotypes, so haplotype polymorphism can be determined according to allele frequency. In addition, the allele frequencies of SNP/InDel vary significantly among different populations. When applied to individual identification in forensic cases, population-specific allele frequencies are necessary (Oldoni et al., 2018). In our study, the application in the Chinese population is temporarily considered, so only the CHB and CHS population in the 1000 Genomes Project phase 3 are used as the source of screening candidate markers.

As a result, the frequency of some multi-Indel markers differed from the theoretical data obtained by the 1000 genome project database using Haploview (**Supplementary Tables S1, S3**). According to the law of free combination, three single markers with linkage equilibrium should display eight different haplotypes. A haplotype with a minimum frequency of 0.001 can be obtained using Haploview calculations. However,

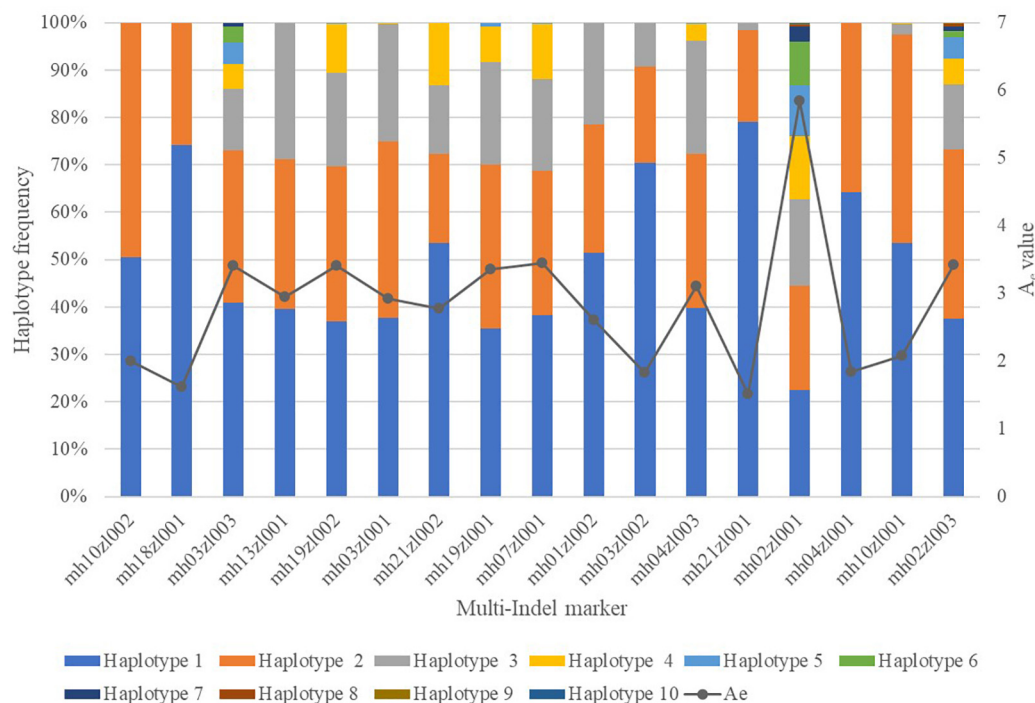


FIGURE 6 | Effective numbers of alleles (A_e) and haplotype frequencies of 17 multi-Indel markers.

TABLE 2 | Values for PD, PE, Ho, PM, PIC, TPI, and HWE of a 17 multi-Indel marker.

Microhaplotype	PD	PE	Ho	PM	PIC	TPI	HWE (p)
mh01zl002	0.78135	0.315571	0.62	0.21865	0.545721	1.315789	0.934909564
mh02zl001	0.9469	0.553495	0.775	0.0531	0.806189	2.222222	0.034566771
mh02zl003	0.8668	0.36213	0.655	0.1332	0.658147	1.449275	0.092087668
mh03zl001	0.7905	0.459875	0.72	0.2095	0.583465	1.785714	0.069324435
mh03zl002	0.64445	0.174709	0.485	0.35555	0.403443	0.970874	0.387110874
mh03zl003	0.8471	0.36213	0.655	0.1529	0.659459	1.449275	0.09720049
mh04zl001	0.60385	0.151068	0.455	0.39615	0.353869	0.917431	0.875124703
mh04zl003	0.83395	0.369131	0.66	0.16605	0.612997	1.470588	0.560030079
mh07zl001	0.8645	0.436037	0.705	0.1355	0.657207	1.694915	0.837279591
mh10zl001	0.6726	0.174709	0.485	0.3274	0.408344	0.970874	0.308612984
mh10zl002	0.6298	0.178899	0.49	0.3702	0.374975	0.980392	0.751342383
mh13zl001	0.8158	0.315571	0.62	0.1842	0.586598	1.315789	0.20736369
mh18zl001	0.54335	0.110909	0.395	0.45665	0.309277	0.826446	0.734621995
mh19zl001	0.84985	0.42826	0.7	0.15015	0.645455	1.666667	0.908230714
mh19zl002	0.8661	0.383394	0.67	0.1339	0.651957	1.515152	0.236016462
mh21zl001	0.5146	0.0888	0.355	0.4854	0.289889	0.775194	0.621845819
mh21zl002	0.8229	0.284923	0.595	0.1771	0.594375	1.234568	0.168954084

we found three multi-Indel markers (mh04zl001, mh10zl002, and mh18zl001) with three Indels having only two different haplotypes as two alleles, which might be related to the complete LD between closely adjacent markers (distances were 17, 38, and 16 bp, respectively). Additionally, seven multi-Indel markers were inconsistent with the theoretical amplicon length, and the haplotype frequency was also different. We verified the homozygous samples of each marker by Sanger

sequencing, especially each amplicon that was inconsistent with the theoretical length. Although our screen limited the existence of other Indels with $MAF > 0.005$ in this range, mh02zl001 and mh02zl003 had 10 and 9 haplotypes, respectively, because additional Indels were detected in this range. In addition, according to the Sanger sequencing results, novel mutations were also found in the mh10zl001 and mh21zl001 loci, which caused the actual allele size and frequency to be inconsistent

with the theoretical value. For the other two loci, mh03zl003 and mh21zl002, we did not find redundant mutations in homozygous samples that have been sequenced by Sanger, but there are also inconsistencies with alleles. These Indels were not included in the database because the goal of the 1000 Genomes Project is to capture the most common human genetic variations (Bergstrom et al., 2020). The development and progress of sequencing technology allows the collection of more varied information.

Our multi-Indel multiplex panel has many advantages. We designed one pair of primers for each multi-Indel marker and one PCR amplicon and one CE run for genotyping. The elimination of sequences with 2–15 nucleotide tandem repeats improved genotyping accuracy and avoided stutter, which is a benefit when analyzing mixtures. Low mutation rates are highly significant in paternity testing, but our results showed that our panel could only serve as an effective supplement to STR, because the PE was not high enough (Huang et al., 2014; Gao et al., 2015; Zhao et al., 2018).

A generalized microhaplotype is essentially a set of all variants in a short fragment, namely multi-variants, which have higher polymorphism. The MPS technology can directly obtain sequences within the read length range, and thus directly determine the phase of a haplotype. Currently, the CE platform is more prevalent in forensic laboratories, so multi-Indels have other potential applications. With the future popularization of MPS, the application of generalized microhaplotypes will become more widespread.

CONCLUSION

In our research, we proposed that the generalized microhaplotype is essentially a collection of all variants in a very short fragment (200 bp), that is, multi-variants with high polymorphism. At present, as the CE platform was widely used in all forensic genetic laboratories, a method based on the CE platform is described in this study. This method can simultaneously detect 18 microhaplotype markers consisting of three or more Indels with different insertion or deletion lengths in the range of less than 200 bp. Our multi-InDel microhaplotypes panel have shorter fragments than conventional STR markers, and have more potential in forensics considering the degraded DNA. In addition, multi-InDel microhaplotypes do not generate stutter involved with PCR amplification, which have more potential in forensics considering the mixture of DNA from two or more individuals. Finally, multi-InDel microhaplotypes offer a much lower mutation rate than STR markers, and it can be

used as supplementary in paternity cases with STR mutation. And the results of combined power of discrimination (CPD) (0.99999999997234) certified the usefulness of our panel for forensic personal identification. But our results also showed that our panel can only be used as an effective supplement to STR, because the CPE (0.9984) is not high enough. Therefore, microhaplotypes consisting of three or more Indels which can be resolved by CE platform have great application potential in forensic genetics.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the Figshare: <https://doi.org/10.6084/m9.figshare.12924551.v2>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethical Committee of the Sichuan University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SQ, WL, and LZ designed this study. SQ, ML, and JX wrote the manuscript. YL conducted sample collection. SQ, JZ, and RZ conducted the experiment. SQ, LW, HJ, and LaZ analyzed the results. All authors reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by grants from the National Natural Science Foundation of China (No. 81971799).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.567082/full#supplementary-material>

Supplementary Figure 1 | Representative electropherogram of control DNA M308 amplified at 1 ng.

REFERENCES

- Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457
- Bergstrom, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecsek, P., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science* 367:eaay5012. doi: 10.1126/science.aay5012
- Besenbacher, S., Liu, S., Izarzugaza, J. M. G., Grove, J., Belling, K., Bork-Jensen, J., et al. (2015). Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.* 6:5969. doi: 10.1038/ncomms6969
- Borsting, C., and Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. *Foren. Sci. Int. Genet.* 18, 78–89. doi: 10.1016/j.fsigen.2015.02.002
- Campbell, C. D., and Eichler, E. E. (2013). Properties and rates of germline mutations in humans. *Trends Genet.* 29, 575–584. doi: 10.1016/j.tig.2013.04.005

- Castella, V., Gervais, J., and Hall, D. (2013). DIP-STR: highly sensitive markers for the analysis of unbalanced genomic mixtures. *Hum. Mutat.* 34, 644–654. doi: 10.1002/humu.22280
- Cereda, G., Biedermann, A., Hall, D., and Taroni, F. (2014). An investigation of the potential of DIP-STR markers for DNA mixture analyses. *Foren. Sci. Intern. Genet.* 11, 229–240. doi: 10.1016/j.fsigen.2014.04.001
- Chen, P., Deng, C., Li, Z., Pu, Y., Yang, J., Yu, Y., et al. (2019). A microhaplotypes panel for massively parallel sequencing analysis of DNA mixtures. *Foren. Sci. Int. Genet.* 40, 140–149. doi: 10.1016/j.fsigen.2019.02.018
- de la Puente, M., Phillips, C., Xavier, C., Amigo, J., Carracedo, A., Parson, W., et al. (2020). Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems. *Foren. Sci. Int. Genet.* 45:102213. doi: 10.1016/j.fsigen.2019.102213
- Gao, T. Z., Yun, L. B., He, W., Gu, Y., and Hou, Y. P. (2015). The application of multi-InDel as supplementary in paternity cases with STR mutation. *Foren. Sci. Intern. Genet. Suppl. Ser. 5*, e218–e219. doi: 10.1016/j.fsigss.2015.09.087
- Gjertson, D. W., Brenner, C. H., Baur, M. P., Carracedo, A., Guidet, F., Luque, J. A., et al. (2007). ISFG: recommendations on biostatistics in paternity testing. *Foren. Sci. Intern. Genet.* 1, 223–231. doi: 10.1016/j.fsigen.2007.06.006
- 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
- Kidd, K. K. (2016). Proposed nomenclature for microhaplotypes. *Hum. Genom.* 10:16. doi: 10.1186/s40246-016-0078-y
- Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., et al. (2013). Microhaplotype loci are a powerful new type of forensic marker. *Foren. Sci. Intern. Genet. Suppl. Ser. 4*, e123–e124. doi: 10.1016/j.fsigss.2013.10.063
- Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., et al. (2014). Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Foren. Sci. Int. Genet.* 12, 215–224. doi: 10.1016/j.fsigen.2014.06.014
- Kidd, K. K., and Speed, W. C. (2015). Criteria for selecting microhaplotypes: mixture detection and deconvolution. *Investig. Genet.* 6:1. doi: 10.1186/s13323-014-0018-3
- Kondrashov, A. S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Hum. Mutat.* 21, 12–27. doi: 10.1002/humu.10147
- Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* 40, 1068–1075. doi: 10.1038/ng.216
- Liu, J., Li, W., Wang, J., Chen, D., Liu, Z., Shi, J., et al. (2019). A new set of DIP-SNP markers for detection of unbalanced and degraded DNA mixtures. *Electrophoresis* 40, 1795–1804. doi: 10.1002/elps.201900017
- Liu, Z. Z., Liu, J. D., Wang, J. Q., Chen, D. Q., Liu, Z. D., Shi, J., et al. (2018). A set of 14 DIP-SNP markers to detect unbalanced DNA mixtures. *Biochem. Biophys. Res. Commun.* 497, 591–596. doi: 10.1016/j.bbrc.2018.02.109
- Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U.S.A.* 107:961. doi: 10.1073/pnas.0912629107
- Moriot, A., and Hall, D. (2019). Analysis of fetal DNA in maternal plasma with markers designed for forensic DNA mixture resolution. *Genet. Med.* 21, 613–621. doi: 10.1038/s41436-018-0102-9
- Oldoni, F., Castella, V., and Hall, D. (2015). A novel set of DIP-STR markers for improved analysis of challenging DNA mixtures. *Foren. Sci. Int. Genet.* 19, 156–164. doi: 10.1016/j.fsigen.2015.07.012
- Oldoni, F., Kidd, K. K., and Podini, D. (2018). Microhaplotypes in forensic genetics. *Foren. Sci. Int. Genet.* 38, 54–69. doi: 10.1016/j.fsigen.2018.09.009
- Oldoni, F., and Podini, D. (2019). Forensic molecular biomarkers for mixture analysis. *Foren. Sci. Int. Genet.* 41, 107–119. doi: 10.1016/j.fsigen.2019.04.003
- Pang, J. B., Rao, M., Chen, Q. F., Ji, A. Q., Zhang, C., Kang, K. L., et al. (2020). A 124-plex microhaplotype panel based on next-generation sequencing developed for forensic applications. *Sci. Rep.* 10:1945. doi: 10.1038/s41598-020-58980-x
- Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., Cartwright, R. A., et al. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods* 10, 985–987. doi: 10.1038/nmeth.2611
- Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* 8, 103–106. doi: 10.1111/j.1471-8286.2007.01931.x
- Snyder, M. W., Adey, A., Kitzman, J. O., and Shendure, J. (2015). Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* 16, 344–358. doi: 10.1038/nrg3903
- Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81. doi: 10.1038/nature15394
- Sun, K., Ye, Y., Luo, T., and Hou, Y. (2016). Multi-InDel analysis for ancestry inference of sub-populations in China. *Sci. Rep.* 6:39797. doi: 10.1038/srep39797
- Sun, S., Liu, Y., Li, J., Yang, Z., Wen, D., Liang, W., et al. (2020). Development and application of a nonbinary SNP-based microhaplotype panel for paternity testing involving close relatives. *Foren. Sci. Int. Genet.* 46:102255. doi: 10.1016/j.fsigen.2020.102255
- Tan, Y., Bai, P., Wang, L., Wang, H., Tian, H., Jian, H., et al. (2018). Two-person DNA mixture interpretation based on a novel set of SNP-STR markers. *Foren. Sci. Int. Genet.* 37, 37–45. doi: 10.1016/j.fsigen.2018.07.021
- Tan, Y., Wang, L., Wang, H., Tian, H., Li, Z. L., Wang, Q., et al. (2017). An investigation of a set of DIP-STR markers to detect unbalanced DNA mixtures among the southwest Chinese Han population. *Foren. Sci. Int. Genet.* 31, 34–39. doi: 10.1016/j.fsigen.2017.08.014
- Turchi, C., Melchionda, F., Pesaresi, M., and Tagliabracci, A. (2019). Evaluation of a microhaplotypes panel for forensic genetics using massive parallel sequencing technology. *Foren. Sci. Int. Genet.* 41, 120–127. doi: 10.1016/j.fsigen.2019.04.009
- Wang, L., He, W., Mao, J., Wang, H., Jin, B., Luo, H. B., et al. (2015). Development of a SNP-STRs multiplex for forensic identification. *Foren. Sci. Intern. Genet. Suppl. Ser. 5*, e598–e600. doi: 10.1016/j.fsigss.2015.09.236
- 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
- Wendt, F. R., Warshawer, D. H., Zeng, X., Churchill, J. D., Novroski, N. M. M., Song, B., et al. (2016). Massively parallel sequencing of 68 insertion/deletion markers identifies novel microhaplotypes for utility in human identity testing. *Foren. Sci. Int. Genet.* 25, 198–209. doi: 10.1016/j.fsigen.2016.09.005
- Zhang, R., Tan, Y., Jian, H., Qu, S., Liu, Y., Zhu, J., et al. (2020). A new approach to detect a set of SNP-SNP markers: combining ARMS-PCR with SNaPshot technology. *Electrophoresis* 7, 150–151. doi: 10.1002/elps.2020.00009
- Zhao, F., Wu, X., Cai, G., and Xu, C. (2003). The application of modified-powerstates software in forensic biostatistics. *Chin. J. Foren. Med.* 18, 297–298.
- Zhao, X. H., Chen, X. G., Zhao, Y. C., Zhang, S., Gao, Z. H., Yang, Y. W., et al. (2018). Construction and forensic genetic characterization of 11 autosomal haplotypes consisting of 22 tri-allelic indels. *Foren. Sci. Int. Genet.* 34, 71–80. doi: 10.1016/j.fsigen.2018.02.001
- Zhu, J., Lv, M., Zhou, N., Chen, D., Jiang, Y., Wang, L., et al. (2019). Genotyping polymorphic microhaplotype markers through the Illumina(R) MiSeq platform for forensics. *Foren. Sci. Int. Genet.* 39, 1–7. doi: 10.1016/j.fsigen.2018.11.005

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

Copyright © 2020 Qu, Lv, Xue, Zhu, Wang, Jian, Liu, Zhang, Zha, Liang 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.



From Identification to Intelligence: An Assessment of the Suitability of Forensic DNA Phenotyping Service Providers for Use in Australian Law Enforcement Casework

Lauren Atwood^{1*}, Jennifer Raymond¹, Alison Sears^{1,2}, Michael Bell¹ and Runa Daniel³

¹ Science and Research Unit, Forensic Evidence and Technical Services Command, New South Wales Police Force, Sydney, NSW, Australia, ² Forensic Analytical and Science Service, New South Wales Health Pathology, Sydney, NSW, Australia,

³ Office of the Chief Forensic Scientist, Victoria Police Forensic Services Department, Melbourne, VIC, Australia

OPEN ACCESS

Edited by:

Ozlem Bulbul,
Istanbul University-Cerrahpasa,
Turkey

Reviewed by:

Horolma Pamjav,
Hungarian Institute for Forensic
Sciences, Hungary
Claus Borsting,
University of Copenhagen, Denmark

*Correspondence:

Lauren Atwood
atwo1lau@police.nsw.gov.au

Specialty section:

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

Received: 01 June 2020

Accepted: 20 November 2020

Published: 12 January 2021

Citation:

Atwood L, Raymond J, Sears A,
Bell M and Daniel R (2021) From
Identification to Intelligence: An
Assessment of the Suitability
of Forensic DNA Phenotyping Service
Providers for Use in Australian Law
Enforcement Casework.
Front. Genet. 11:568701.
doi: 10.3389/fgene.2020.568701

Forensic DNA Phenotyping (FDP) is an established but evolving field of DNA testing. It provides intelligence regarding the appearance (externally visible characteristics), biogeographical ancestry and age of an unknown donor and, although not necessarily a requirement for its casework application, has been previously used as a method of last resort in New South Wales (NSW) Police Force investigations. FDP can further assist law enforcement agencies by re-prioritising an existing pool of suspects or generating a new pool of suspects. In recent years, this capability has become ubiquitous with a wide range of service providers offering their expertise to law enforcement and the public. With the increase in the number of providers offering FDP and its potential to direct and target law enforcement resources, a thorough assessment of the applicability of these services was undertaken. Six service providers of FDP were assessed for suitability for NSW Police Force casework based on prediction accuracy, clarity of reporting, limitations of testing, cost and turnaround times. From these assessment criteria, a service provider for the prediction of biogeographical ancestry, hair and eye colour was deemed suitable for use in NSW Police Force casework. Importantly, the study highlighted the need for standardisation of terminology and reporting in this evolving field, and the requirement for interpretation by biologists with specialist expertise to translate the scientific data to intelligence for police investigators.

Keywords: forensic DNA phenotyping, intelligence, casework, law enforcement, massively parallel sequencing

INTRODUCTION

Since the application of DNA analysis in forensic casework in the late 1980s, considerable technological advancements have resulted in an expansion of forensic DNA analysis capabilities. Currently, in the majority of operational forensic laboratories, the use of DNA evidence is heavily focused on identification using STRs, limited by the reliance on comparison to other STR-generated profiles stored in a DNA database or to a reference sample from a known suspect. A notable difference with inferring biogeographical ancestry (BGA) and externally visible characteristics

(EVCs) of an unknown individual—referred to as Forensic DNA Phenotyping (FDP) or DNA Intelligence—is the capacity to provide DNA information in an investigation to assist with individual identification by generating leads without reliance on the availability of a comparison sample. FDP enables investigators to generate or re-prioritise a suspect pool based on an unknown sample, thereby providing investigative leads that could assist with the identification of the DNA donor using STR profiling (or other) techniques. Such intelligence can be applied to cold cases, unidentified human remains cases and disaster victim identification; all scenarios where the success of STR identification can often have additional limitations due to degraded, or poor quality, biological evidence. This methodology has been applied successfully in casework for approximately 15 years with some of the earliest reported cases being the Louisiana Serial Killer case (2004) (Touchette, 2003) and the 11M Madrid bomb attack (Phillips et al., 2009).

Prediction accuracy is essential for confidence in result outcomes when applying FDP to casework. The use of relevant and informative DNA markers for the traits of interest is of paramount importance. Secondly, the composition of the reference set that is used to train the analysis algorithms must be appropriate and relevant for the predictive trait. The populations contained within these datasets are often unknown to the user or may vary considerably in their representative construct applicable to the trait being tested (Cheung et al., 2018). In addition, the accuracy of the prediction is dependent on the prediction algorithm used. Admixture is an additional challenge in the prediction of BGA, and ongoing research continues to address interpretation and reporting for operational application (Jin et al., 2018). The technical limitations of BGA and EVC prediction, including the availability of a quality sample, genetic admixture, and available reference datasets, have been discussed at various lengths (Kayser, 2015; Schneider et al., 2019).

A number of forensically relevant panels have been developed to provide accurate predictions of an unknown individual's EVCs and BGA (Walsh et al., 2011b, 2013; Al-Asfi et al., 2017; Phillips et al., 2019). However, service providers differ in their testing approach and reference sets used, which may be reflected in the result outcome and, ultimately, the prediction accuracy. From an operational perspective, confidence in results and outcomes stems not only from a technically acceptable prediction, but a result that also clearly defines the reliability of the conclusion, whilst considering the above limitations of testing and reporting outcomes.

In addition to prediction accuracy, and contextualising testing limitations, an operational need is for a service provider to generate a report that is appropriate for direct release to a non-scientific/non-specialist audience (hereafter referred to as a lay audience). Reporting of STR profiles uses statistics to demonstrate the strength of a match whereas the use of statistical analysis for FDP reporting is to demonstrate the confidence in the prediction. Therefore, translation of scientific outcomes of FDP to lay audiences has been shown to be variable, particularly compared to STR profile reporting (Scudder et al., 2020). However, it has been proposed that ongoing education is beneficial for lay audiences to gain an understanding

and awareness of the method and its application in casework (Daniel, 2016; Raymond et al., 2017). As FDP is a new and developing technology to be embedded in operational use within the New South Wales (NSW) Police Force, it is pertinent that considerable attention is focused on ensuring accurate comprehension by investigators.

The aim of this study (conducted in 2017/2018) was to compare results obtained from six providers of FDP services for BGA and EVCs (hair, eye, skin colour, and age) to determine suitability for operational application to NSW Police Force casework. Established service providers of FDP, with recognised expertise within the forensic community, were canvassed and invited to participate in the study. All service providers (six) that consented and were able to participate in the time frame requested were included. The service providers encompassed both commercial and non-commercial laboratories and groups to generate BGA and EVC data. Known volunteer donors with BGAs and EVCs representative of the diverse Australian population were selected for this study. The assessment criteria for determining suitability of a service provider for operational application must be strict and aligned with accreditation, legal, ethical and moral expectations of both the NSW Police Force and the community at large. This study used the following three categories for assessment:

1. Prediction accuracy
2. Clarity of reporting
3. Ability to generate results from all samples

For the purpose of this study, focus is placed on available service providers who could potentially contribute to operational environment needs, and assessment was based on the number of accurate predictions as compared to self-declared traits, with consideration given to the limitations described above. All service providers used in this study have been de-identified.

To assess clarity of reporting, service providers were asked to provide their report results as required for standard casework, if applicable. The service providers were based in several different countries and therefore report for consumers with varying legislative requirements. The reports were assessed for both accuracy of scientific content by a subject matter expert (SME) and for comprehension by a lay audience, such as law enforcement personnel (e.g., detectives and investigators), in the context of the adversarial legal system in place within NSW. The SME was a forensic biologist employed within an operational policing and forensic agency, with extensive academic and research expertise in DNA intelligence. Three factors were considered when reviewing reporting styles: consistent language, ease of interpretation and overall clarity.

Finally, cost and turnaround time are an important consideration in the operational application of any specialist service; therefore, these parameters were also considered. Whilst these points were not specifically requested of service providers, the service request was made for “...within usual work timeframes...” and quotes for service provision were provided based on the pre-determined sample numbers that would be submitted.

MATERIALS AND METHODS

Donor Selection, Sampling, and Collection of Data

Ten known donors of varied BGA and EVCs relevant to the Australian population (**Table 1**) were sourced voluntarily from within the Forensic Evidence and Technical Services Command (FETSC) of the NSW Police Force. All donors provided informed consent and were de-identified, 1–10. The study was conducted in compliance with the *National Health and Medical Research Council (2015)*, consisting of a series of guidelines developed in accordance with the *National Health and Medical Research Council Act 1992* (NHMRC guidelines) (National Health and Medical Research Council, 2015).

Participants were asked to self-declare the following BGA and EVC information, which was confirmed by an independent evaluator at the time of collection to ensure consistency (**Table 1**). Photographs of the donors' face, eyes and hair were also obtained (not shown).

- BGA: self-declared over three generations (self, parents, maternal and paternal grandparents, as per general biological pedigree definition). The degree of admixture was determined by a SME based on the donor's self-declared BGA over the three generations;
- Eye colour: self-declared using categories blue, grey, green, hazel and brown;
- Skin colour: self-declared, based on an area of their body not exposed to light at age 20. Skin colour categories were fair/pale, medium, olive and dark.
- Hair colour: self-declared at age 0–4 years, 20 years old and current age. Natural hair colour categories were fair/blonde, light brown, light red/ginger, dark red/australian, dark brown and black.
- Hair greying: self-declared percentage of grey currently, and approximate age at which greying occurred.

All six service providers used in this study were de-identified, denoted A–F. Samples of saliva and blood were collected as instructed by the service provider. **Table 2** outlines the DNA sample type for each provider. DNA extracts were prepared from saliva stained Whatman FTATM MiniCard from each donor by the NSW Health Pathology Forensic and Analytical Science Services laboratory. DNA extraction was performed using PrepFiler[®] Automated Forensic DNA Extraction Kit [Thermo Fisher Scientific (TFS)] and quantification using the Quantifiler[®] Trio DNA Quantification Kit (TFS). A 5-mm hole punch of the saliva-stained Whatman FTATM MiniCard was sent to Providers A and D and DNA extract to Providers B and C for downstream analysis. The blood-stained Whatman FTATM MiniCard was provided to Provider A only. Neat saliva samples in proprietary collection tubes were sent to Providers E and F. Providers A, E and F did not disclose the minimum quantity of DNA required to perform testing.

Analysis by Service Providers

The testing undertaken by each provider is shown in **Table 2**, in addition to the marker panels, genotyping platforms and analysis methods used as indicated in the provider's results report or as declared by the provider. Providers B, C, and D all tested for eye colour, hair colour and BGA. Skin colour and age prediction were only tested by Providers C and A, respectively. Providers E and F only generated results for BGA and the marker panels, genotyping platforms and analysis methods used were not disclosed. The results from these providers were sent directly to the donors. The donors then chose to provide the results for this study. A SME and operational forensic scientists (biologists) assessed the results, analysis, prediction accuracy and reporting from the providers.

RESULTS

The summarised results generated by the service providers were assessed using three main criteria: prediction accuracy, clarity of reporting and ability to generate results from all samples. Additional criteria used to assess the providers included cost and turnaround time for analysis and reporting.

It was known prior to commencing this study that Providers E and F did not conduct analysis of EVCs and that their service does not include analysis of casework samples. However, these providers were included in the study for comparative purposes to assess variation in BGA prediction, accuracy and reporting styles between the service providers. A summary of the prediction performance for all service providers is shown in **Supplementary Table 1**.

Prediction Accuracy

The prediction accuracy of each service provider for eye colour, hair colour, skin colour, age and BGA, respectively, is shown in **Table 3**. The prediction accuracy (%) indicates the number of correct predictions of the total predictions made. As indicated, results were not obtained, or not available, for all samples and testing type.

Eye, Hair, and Skin Colour

Four of the six providers conducted eye colour analysis. Only two providers (Providers A and B) generated results for all 10 donors. Provider B achieved the highest prediction accuracy for eye colour (90%) followed by Provider C (89%). However, the eye colour of donor 3 was not predicted correctly by any of the providers. Donor 3's self-declared eye colour was hazel; however, it was predicted to be blue by all providers. Categorised as an intermediate eye colour, hazel eye colour has an expected prediction accuracy of 74% using Irisplex SNPs (Walsh et al., 2011a,b; Walsh et al., 2012, 2013). However, donor 10's hazel eye colour was correctly predicted by Provider B but not by Provider A. Providers C and D did not return eye colour results for this sample. Self-declared brown and blue eye colours were correctly predicted.

Three of the six providers tested for hair colour. Only two providers (Providers B and C) generated results for all 10 donors.

TABLE 1 | Donor's self-declared biogeographical ancestry (BGA) and externally visible characteristics (EVCs).

Donor	BGA	Eye colour	Hair colour	Skin colour	Age (years)
Donor 1	Non-admixed South Asian	Brown	Black	Olive	49
Donor 2	Non-admixed Pacific Islands	Brown	Black	Olive	39
Donor 3	Admixed European/Aboriginal	Hazel	Dark brown	Medium	25
Donor 4	Non-admixed Middle Eastern	Brown	Dark brown	Medium	47
Donor 5	Non-admixed East Asian	Brown	Black	Medium	44
Donor 6	Non-admixed Middle Eastern	Brown	Dark brown	Olive	49
Donor 7	Non-admixed European	Blue	Blonde	Fair/Pale	27
Donor 8	Non-admixed South East Asian	Brown	Black	Olive	59
Donor 9	Non-admixed East Asian	Brown	Black	Medium	35
Donor 10	Non-admixed European	Hazel	Red	Fair/Pale	52

The degree of admixture was determined using the BGA declared by the donor over three generations.

Provider D achieved the highest prediction accuracy for hair colour (86%). Providers B and C both achieved a hair colour prediction accuracy of 80%. The incorrect results generated by Provider B were for dark brown and blonde hair predicted to be black and brown, respectively. Provider C incorrectly predicted dark brown hair as red for donors 3 and black for donor 4. Provider D incorrectly predicted blonde hair as brown for donor 7.

Skin colour was only tested by Provider C and results were obtained for all 10 donors. However, the skin colour prediction accuracy was 50% with the incorrect predictions being for olive or medium skin colours predicted as white for donors 1, 3, 4, 6, and 8.

Age

Figure 1 displays the age of the eight donors tested, against the predicted age range given by Provider A. **Table 4** provides an example of provider A's reporting style of predicted age ranges presented as a "mean age" and age range with a 95% confidence interval (as reported by the provider). The overall age prediction accuracy was 25%. **Figure 1** demonstrates the variation in the age ranges provided across the donors. Although a small dataset, no trends in relation to correct predictions of older or younger donors were observed, nor were the predictions consistently below or above the correct age. A decrease in predicted age range did not correlate with an increase in successful age prediction.

BGA

All six providers tested for BGA. Providers A, B, C, D, and F returned results for the 10 donors. Providers B, E and F achieved the highest prediction accuracy for BGA (100%) followed by Provider A (90%). Results from Providers E and F were disseminated directly to the donors. Donor 2 did not return their BGA results from Provider E for inclusion in this study. Provider E did not generate a result for Donor 10.

The assays and genotyping platforms used by the providers varied greatly and included SNaPshot assays, massively parallel sequencing assays and high-density SNP arrays (**Table 2**). Therefore, the markers analysed, reference sets and prediction algorithms also varied, not allowing a direct comparison of prediction accuracies between providers. Additionally, the

reporting styles of the providers ranged from referring only to geographic ancestry to including statements about ethnicity. However, of the traits predicted in this study, the highest prediction accuracies (100%) were generated for BGA prediction from three of the six providers. BGA prediction was offered by all service providers.

Reporting Clarity

The service providers were instructed to provide a report that did not require additional interpretation and could be disseminated directly to investigators. Therefore, ease of comprehension by a lay person was a primary consideration in our assessment. Examples of reported results provided below have been selected to best represent the variation in reporting from service providers and to reflect the challenges associated with interpretation of these results. It was known prior to commencing this study that due to service capabilities, Provider D would only provide the genotyping platform's onboard analysis software output without interpretation of the results. Comparative analysis of reporting was categorised into three key components: consistent language, ease of interpretation by a lay audience and overall clarity.

All service providers reported EVC results using consistent language within their reports. However, the reporting style for EVC results varied greatly between providers. The reported predicted phenotype is indicated as correct or incorrect based on comparison to the donor's self-declared EVC.

To compare the donors' self-declaration to the categories reported by service providers, the following process was applied when assessing the accuracy of hair and eye colour predictions:

- (i) Self-declared brown (light or dark) hair colour: Any service provider predictions of "brown," "dark brown" or "light brown" were recorded as correct.
- (ii) Self-declared "light red or ginger" and "dark red or auburn" hair colour: Service provider predictions of "red" were recorded as correct.
- (iii) Service provider predictions indicating a range of hair colours for a donor (e.g., Brown/Black) were recorded as correct if *any* of the hair colours predicted matched the donor's self-declaration.

TABLE 2 | Sample types, assays, platforms and analysis methods for BGA and EVC prediction as disclosed by the six service providers.

Provider	A	B	C	D	E	F
Sample type	Whatman FTA TM MiniCard—Blood Whatman FTA TM MiniCard—Saliva (5-mm hole punch)	DNA extracts from saliva on Whatman FTA TM MiniCard	DNA extracts from saliva on Whatman FTA TM MiniCard	Whatman FTA TM MiniCard—Saliva (5-mm hole punch)	Provider E [®] and F [®] DNA Collection Kit	
BGA	Marker panel and genotyping platform <ul style="list-style-type: none"> Custom 41-plex SNP panel on MiSeq FGx (Illumina[®]) mtDNA control region (Sanger) PowerPlex[®] Y23 (Promega) Analysis <ul style="list-style-type: none"> SNIPPER, PCoA, STRUCTURE EMPOP database Y-HRD database 	Marker panel and genotyping platform <ul style="list-style-type: none"> Precision ID Ancestry Panel (TFS) (165 autosomal SNPs) Ion PGMTM System (TFS) Ion ChefTM (TFS) Ion 316TM v2 BC chips (TFS) Ion PGMTM Hi-QTM View Sequencing (TFS) Analysis <ul style="list-style-type: none"> HiD SNP Genotyper Plugin (TFS) STRUCTURE PCoA (Microsoft R) 	Marker panel and genotyping platform <ul style="list-style-type: none"> Custom marker panel (144 autosomal SNPs, panel in development) AmpF/STRTM Y FilerTM PCR amplification kit (TFS) Ion GeneStudio S5 System (TFS) Analysis <ul style="list-style-type: none"> Genotyper software (TFS) Modified Genotyper software (TFS) NEVGEN Y-DNA Haplogroup Predictor Haplogrep and EMPOP Emma. Phylotree v.17 mtDNA analysis SNIPPER, PCoA, STRUCTURE 	Marker panel and genotyping platform <ul style="list-style-type: none"> ForenSeq DNA Signature Prep Kit (Verogen) (231 STRs and SNPs) MiSeq FGx (Verogen) Analysis <ul style="list-style-type: none"> ForenSeq UAS (Verogen) 	Undisclosed	
EVCs	Marker panel and genotyping platform <ul style="list-style-type: none"> IrisPlex assay SNaPshot[®] Multiplex Kit (Applied BiosystemsTM) 3130xl Genetic Analyzer (Applied BiosystemsTM) Analysis <ul style="list-style-type: none"> IrisPlex Webtool (Erasmus) 	Marker panel and genotyping platform <ul style="list-style-type: none"> Ion AmpliseqTM DNA Phenotyping Panel—24 SNPs (TFS) Ion ChefTM System (TFS) Ion 314TM v2 BC chips (TFS) Ion PGMTM Hi-QTM View Sequencing (TFS) Analysis <ul style="list-style-type: none"> HlrisPlex Webtool (Erasmus) 	Marker panel and genotyping platform <ul style="list-style-type: none"> In-house multiplexes (incorporates HlrisPlex SNPs) Analysis <ul style="list-style-type: none"> SNIPPER IrisPlex Webtool (Erasmus) 	Marker panel and genotyping platform <ul style="list-style-type: none"> ForenSeq DNA Signature Prep Kit (Verogen) (231 STRs and SNPs) MiSeq FGx (Verogen) Analysis <ul style="list-style-type: none"> ForenSeq UAS (Verogen) 		
Age	Marker panel and genotyping platform <ul style="list-style-type: none"> Custom marker panel (two multiplexes; 7plex and 5plex) MiSeq FGx (Verogen) Analysis <ul style="list-style-type: none"> Statistical models SVMp, LASSO, ANN 					
Minimum Quantity	Undisclosed	1 ng total DNA	300 ng total DNA	4 ng total DNA (20 μ l of 0.2 ng/ μ l)	Undisclosed	

TABLE 3 | The prediction accuracy (%) of each service provider for eye, hair and skin colour, BGA, and age.

	Provider A	Provider B	Provider C	Provider D	Provider E	Provider F
Eye colour	80%	90%	89%*	86%*	–	–
Hair colour	–	80%	80%	86%*	–	–
Skin colour	–	–	50%	–	–	–
BGA	90%	100%	60%	50%	100%	100%
Age	25%*	–	–	–	–	–
Cost/sample (USD)	\$682	\$556	\$802	\$164	\$96	\$96
Turnaround time (days)	66	30	114	21	28	28

– Service not provided.

* Result not provided for all 10 donors.

Costs as at 2017/2018 pricing.

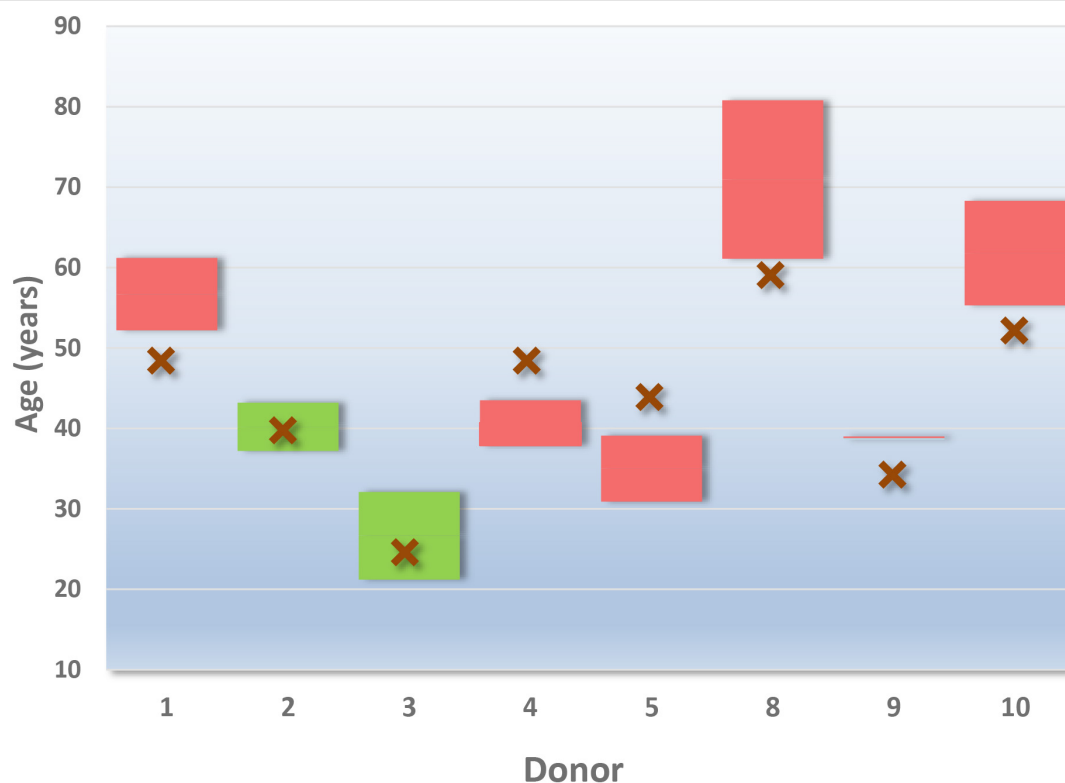


FIGURE 1 | Provider A age predictions for 8 of the 10 donors. The declared age of the donor is indicated by an “X” (red). The boxes outline the predicted age range given by the service provider (reported by the provider as a “95% confidence interval”). Green boxes encompass a correct prediction. Red boxes indicate an incorrect prediction for the donor.

- (iv) Self-declared “green,” “grey,” or “hazel” eye colours: Service provider predictions of “intermediate” eye colour were recorded as correct.

To compare the donor’s country specific self-declaration of BGA (e.g., Chinese, British, Turkish) the donor’s ancestry was reclassified to a sub-geographic region (e.g., East Asia, European, Middle East). In the case of the admixed donor (Table 1), the authors accepted a BGA prediction as correct based on the degree of admixture and the dominant ancestral geographic region declared, in deference to the lack of informative markers and individuals within the reference sets representative of the

Australian Indigenous population (at the time of the study). The providers did not supply a list of countries specified within each sub-geographic region/population groups tested; therefore, the United Nations statistics division’s definition of sub-geographic regions was used (United Nations, 1999).

Provider A

The language used by Provider A to report eye colour and age for all donors was consistent throughout the report (Table 4). Provider A’s reporting style of eye colour was interpretable by a lay person. However, an explanation of the calculation for the prediction error rate was not provided. Provider A

TABLE 4 | Provider A reporting style for eye colour, age, and BGA prediction.

Provider A	Eye colour	Age	BGA
Reported result	Predicted phenotype: Brown Prediction error: < 1%	Mean age: 56.7 95% Confidence Interval: 52.2–61.2	This sample is most likely from a South Asian population, such as Pakistan, but the origin could also be within an area that includes Pakistan and Iran, or less likely be within an area extending into Iraq or India. This prediction is not excluded by the Y-chromosome analysis and the STRUCTURE plot that shows an admixed population typical of some individuals in our reference pakistani population

used the Erasmus IrisPlex/HIrisPlex predictor for eye colour phenotyping. At the time of reporting, the *IrisPlex & HIrisPlex DNA Phenotyping Webtool User Manual Version 1.0* (Forensic Molecular Biology Department of Erasmus MC, n.d.) states that this tool received “overall prediction accuracies” of 94% for blue eyes, 74% for intermediate eye colour and 95% for brown eye colour. Provider A commonly reported prediction errors of < 1%. Whilst this may be correct from the service provider’s perspective, without explanation of how this error rate was determined, there is potential for an investigator to incorrectly assume that the eye colour predictions from Provider A have a > 99% accuracy.

A discrepancy in eye colour prediction accuracy was observed between Providers A and B for donor 10. Although the two providers used the same eye colour markers and webtool, different genotyping methods were used [Provider A used an IrisPlex SNaPshot assay and manual interpretation whilst Provider B used a HIrisPlex-based MPS assay and automated interpretation (Table 2)]. These differences may explain Provider A’s incorrect prediction. However, as Provider A did not provide the genotype data, it was not possible to determine whether the incorrect prediction of Provider A was a result of a genotyping error or differences in interpretation thresholds and reporting criteria applied by each provider.

Although the reported result for age did not require further interpretation, the age ranges varied considerably from ~5 months (Donor 9) to ~19 years and 8 months (Donor 8). Without additional explanation, this reporting style may result in law enforcement personnel associating a degree of confidence with the relative size of the age range reported. For example, the larger the age range reported (i.e., 19 years and 8 months), the less confident an investigator might be in the prediction of age for that donor, and vice versa.

Provider A’s BGA results used varying language throughout their report. The use of terminology (i.e., “very confident”) was inconsistent within and between the donor results. Several of the summaries of donor results reported by Provider A were unclear and, at times, contradictory. For example, “*This sample is likely from an Asian population, but it is not typical of East Asian or South Asian populations. An East Asian and South-East Asian origin is suggested by mtDNA [...] STRUCTURE reveals an admixture with a major East Asian and a minor South Asian contribution.*”

Provider A also used terminology in their report that infers race and skin colour rather than just geographic origin of the donor; e.g., “*this is a Caucasian individual with white European*

ancestry.” This terminology was not used consistently in the report, as two additional donors with similar ancestry to donor 3 were not described in this manner. The term “Caucasian” is widely misunderstood, used as a synonym for “white” and holds no contemporary geographic links (Freedman, 1984; Bhupal and Donaldson, 1998). This combined with Provider A’s “. . . white European” classification highlights the need for standardised terminology when reporting BGA. It is also noted here that there is use of language that infers skin colour, when no such test had been performed.

Provider B

Provider B used consistent language throughout the reporting of all EVCs and BGA (Table 5). Their results were readily understandable by a lay audience. Provider B used terms common in verbal scales (“likely,” “very likely”) to report the EVC and BGA predictions. Provider B’s report did not indicate whether specific criteria were applied or predictive values were used to support this terminology. However, consistent use of these terms allowed comprehension of, and increased confidence in, the results by a lay audience.

Provider C

Eye colour was determined via the output from two different tests Snipper eye and Erasmus eye (as reported by the provider), resulting in likelihood ratio (LR) and *p*-value statements with a predicted colour listed in bold text (Table 6). As Provider C did not combine the results into a single prediction, it was determined that the emphasised eye colour would be understood as the predicted phenotype from Provider C. In every case, the predictions from the Snipper eye and Erasmus eye tests concurred. It is not known how discordant results would be reported should they occur. Similarly, for hair and skin colour prediction, a LR was provided with a colour emphasised. Again, it was assumed that the emphasised colour was the prediction given by Provider C.

Provider C used consistent language throughout the reporting of all EVCs and BGA. However, it was observed that Provider C reported the same BGA result “*Eastern European or Middle Eastern ancestry*” for half of the donors in the study, with two correct predictions.

Provider D

As indicated previously, the reports from Provider D were solely based on instrumentation output of BGA and eye and hair colour analysis without further interpretation or reporting (Table 7).

TABLE 5 | Provider B reporting style for eye colour, hair colour, and BGA prediction.

Provider B	Eye colour	Hair colour	BGA
Reported result	The donor of this DNA is most likely to have brown eyes	The donor of this DNA is most likely to have dark brown hair	The donor of this DNA has a majority ancestral genetic contribution from Europe. Examples include Hungary, Greece and Denmark. They are more likely to have European ancestry than any other continental BGA*. They are likely to have a majority of ancestors (e.g., parents, grandparents) from Europe.

*BGAs include African, Middle Eastern, European, South Asian, East Asian, Oceanian, indigenous American.

TABLE 6 | Provider C reporting style for eye colour, hair colour, skin colour, and BGA prediction.

Provider C	Eye colour	Hair colour	Skin colour	BGA
Reported result	<i>Snipper</i> eye: 9,000 times more likely Brown than Intermediate <i>Erasmus</i> eye: <i>p</i> -value of 0.977 for Brown eye colour	<i>Snipper</i> hair: 1,130,436 times more likely Dark than Fair; insufficient predictive value for Brown vs. Black differentiation	<i>Snipper</i> skin: 649,715 times more likely White than Intermediate	Eastern european or Middle eastern ancestry

TABLE 7 | Provider D reporting style for eye colour, hair colour, and BGA prediction.

Provider D	Eye colour	Hair colour	BGA
Reported result	Intermediate: 0.01 Brown: 0.99 Blue: 0.00	Brown: 0.70 Red: 0.02 Black: 0.12 Blond: 0.17	 <p>Biogeographical ancestry results Distance to nearest centroid 4.49 1,000 genomes populations with samples in centroid with sample</p>
Population			
Population	Abbreviation	Count	In Training Data
British in England and Scotland	GBR	69	70
Finnish in Finland	FIN	71	75
Puerto Ricans from Puerto Rico	PUR	6	52
Colombians from Medellin, Colombia	CLM	8	50
Iberian population in Spain	IBS	6	6
Utah Residents (CEPH) with Northern and Western European ancestry	CEU	71	72
Mexican Ancestry from Los Angeles, United States	MXL	2	54
Toscans in Italy	TSI	86	98

Therefore, the prediction results were determined by a SME and forensic biologists within the project team.

Provider D did not indicate a singular predicted eye or hair colour. A range of prediction values assigned to the categories of “Intermediate,” “Brown” and “Blue” for eye colour, and “Brown,” “Red,” “Black” and “Blond” for hair colour was

provided. The investigator would be required to assume that the highest reported predictive value indicated would be the predicted phenotype.

As expected, the instrumentation output ensured that Provider D’s results were reported consistently, with the same language and presentation of result throughout the

report. The BGA predictions were interpreted by the project team by selecting the population cluster where the donor sample was indicated on the graph. For example, in the case of the donor shown in **Table 7**, due to the overlapping population clusters, it was determined that the predicted BGA was European/Admixed American. The potential difficulties for law enforcement personnel to accurately grasp the instrumentation output highlight the need for expert interpretation and reporting.

Provider E

The results from Provider E were presented in a tabular style with percentages given for continental groups (e.g., European—94.5% and South Asian—1.4%). Sub-population groups were also listed under each continental group (e.g., British and Irish—84.5%, West African—1.2%). Due to copyright requirements from the provider, an example result is not shown here. Provider E used consistent language and reporting style across all 10 donors. The results reported were easily interpretable for a lay audience.

Provider F

Similar to Provider E, Provider F listed percentages against population groups but focused on the sub-populations rather than continental groups (e.g., Great Britain—34%, Europe West—17%, and so on). “Low confidence regions” were also listed. Additionally, Provider F supplied a global map with the “ethnicity estimate” (their terminology) presented as shaded circles over the relevant areas. Likewise, due to copyright requirements, the reporting style of Provider F is not shown here.

Provider F used consistent language and reporting style across all 10 donors. The results reported did not require additional interpretation. The use of a map as a visual aid as seen in Provider F’s reporting assisted in understanding the results. This reporting approach for law enforcement agencies may avoid misinterpretation of geographical regions and terminology. Any “low confidence regions” stated by Provider F were excluded from the assessment of Provider F’s predictive ability for BGA.

Ability to Generate Results From All Samples, Costs, and Turnaround Time

Provider C returned eye colour prediction results for 90% of donors. Provider D produced eye and hair colour results for 70% of donors and Provider E was unable to return a BGA result for Donor 10. All other providers were able to return results for 100% of samples tested. All service providers were administered the quantity and quality of the DNA samples requested (**Table 2**). No explanation was provided by those service providers unable to return results for all samples.

Table 3 lists the costs and turnaround times for each provider. However, a direct comparison of the cost and turnaround time was not possible as the types of services, assays and technology used by each provider were often different.

DISCUSSION

The criteria applied to assess the six service providers in this study were selected to determine suitability for the application of FDP to casework samples. The service providers analysed samples from 10 donors (where possible) that self-declared their BGA, eye, hair and skin colour and age. This study indicated that the prediction accuracies and validated methodologies for BGA, hair and eye colour were appropriate for application to casework. Additional EVCs tested (skin colour and age) required more extensive research and development to increase the prediction accuracies. However, the authors note that considerable progress has been made on age and skin colour prediction since this study was conducted.

The donor samples received by the service providers were pristine, high source saliva or blood samples, unlike samples routinely encountered in casework. Casework samples often collected from trace evidence are of compromised quality (degraded). It is unlikely that casework samples will exceed the DNA quantity or quality of a sample retrieved directly from the donor source (e.g., a pristine or reference sample). Therefore, the inability to generate results from the donor samples was a point of consideration.

Of the four providers (A, B, C, and D) that could service law enforcement, based on the findings from the trial of 10 known donors using the criteria outlined in this study, Provider B was deemed suitable for use in NSW Police Force casework. A high prediction accuracy was observed for eye colour (90%), hair colour (80%), and BGA (100%). Provider B’s reporting style also satisfied the clarity assessment with clear, concise and effectively communicated results. Generating results occurred within a suitable time frame (30 days) and average cost/sample. Results were provided for all donor samples.

As a result of this study, FDP was incorporated into routine NSW Police casework. Future considerations for full operationalisation include assimilation into a quality framework with regular proficiency testing as per routine forensic analyses and accreditation requirements. Utilising the lessons learnt, the SME was engaged to interpret Service Provider B’s data analysis output and report the results using a reporting style template developed for dissemination directly to investigators.

Based on interaction and feedback from investigators, the ideal reporting template would include clear and concise language comprehensible by a non-expert/lay audience. The performance of the assays, etc. would be assessed by the scientific expert; therefore, the test characteristics mentioned previously do not require interpretation by the lay audience. Although indicating the accuracy of the prediction should be communicated to the investigator, language used to communicate this would not use scientific or specialist terms such as LR. Exclusions may be reported where possible and the limitations of the tests should be clearly indicated. FDP reporting style and dissemination of the results are important considerations for law enforcement agencies that will be addressed in a subsequent manuscript.

Observations made throughout this study have highlighted the need for caution and further discussion surrounding FDP, specifically the interpretation and reporting of results

for law enforcement consumption. In general, greater contextual understanding of outcomes could be achieved through standardised reporting terminology. The key observed points from this study relate to (i) definition of sub-geographic regions for BGA predictions, (ii) avoidance of association of BGA prediction with an individual's physical appearance, and (iii) standardisation of nomenclature for broader comprehension of results.

Regarding the issue of definition of sub-geographic regions, the reported BGA results are provided on a continent or sub-continent level. Most individuals are expected to refer their ancestry as a country-specific declaration (i.e., Chinese) as opposed to a continental or continental sub-regional scale (i.e., East Asian). This presents a challenge regarding how best to correlate the two in order to (1) define which countries lay within the reported sub-geographic region and (2) reach a consensus between the different service providers of how this should be defined and reported to achieve consistency.

A lay person's interpretation of countries that may be included in a sub-geographic region, such as "East Asia" and "Middle East," may be influenced by a number of factors such as the individual's conscious and unconscious biases (life experience, education, social, and political context) (Samuel and Prainsack, 2018). To exacerbate the issue, definitions of countries included in sub-geographic regions are subject to change with shifting political and social influences/circumstances. It is recommended that each service provider provides a comprehensive list of countries within a region that align with their reporting of BGA prediction, or a map that would include their definitions of sub-geographic regions relevant to their reference populations.

EVC prediction accuracy assessment required alignment of the self-declared and reported eye and hair colour categories. In such a comparison, the differences in categories used are a potential source of error and highlight the need for standardised collection and reporting of EVCs to remove subjectivity. Future assessments may benefit from provision of defined self-declaration and reporting categories to both participants and providers to increase consistency.

A separate issue relates to the association of BGA assessment with an individual's physical appearance. Prediction of the BGA of a donor is not the prediction of race, ethnicity or cultural background. It provides a prediction of the ancestral geographic or sub-geographic region of that donor. It is important to convey to law enforcement that although the affiliation between BGA prediction and assumption of physical appearance may align in some cases, BGA prediction does not imply the physical appearance of the DNA donor (Kayser and Schneider, 2012; Samuel and Prainsack, 2018).

The requirement for standardisation of nomenclature was the third issue highlighted from this study. Inferences of an individual's BGA or EVCs can be made using FDP. However, this may be a probabilistic prediction depending on the trait of interest. Therefore, it is possible for the nature of the information to be misunderstood (Enserink, 2011; Cino, 2016; Samuel and Prainsack, 2018). The reporting of EVC results from providers in this study highlighted the need for standardised language to indicate the test performance (positive

predictive value, negative predictive value, sensitivity, specificity) to assist with scientific interpretation. In addition, the service providers should provide the genotype data generated for every test undertaken to allow for independent verification of the results by the SME. Lay interpretation could be assisted by the provider indicating the % correct predictions per phenotype as a means of indicating potential error. Regardless of accurate predictions of BGA and EVCs by the service providers, equally as important is the delivery of the information.

The issues identified in this study support the involvement of a SME as a critical aspect of the interpretation and dissemination of FDP results to investigators. It was evident that providing an external report directly to investigators without SME review of the service provider's analysis and interpretation increases the potential for misinterpretation. Given that this is a form of intelligence used to generate investigative leads, there is also potential for inadequate or expert review of the results to misdirect an investigation.

It is clear from these findings that there is merit in developing standardised nomenclature and reporting of DNA intelligence. Benefits of this approach would ensure that DNA intelligence can be more extensively integrated within law enforcement investigations, effectively communicated to investigators and to minimise the potential for misinterpretation.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because approval for access to datasets must be granted. Requests to access the datasets should be directed to LA, atwo1lau@police.nsw.gov.au.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. 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

LA: data collation and analysis, logistics, background research, and writing—original draft. JR: conceptualisation, proposal, methodology, and writing—reviewing and editing. AS: methodology, logistics, and writing—reviewing and editing. MB: conceptualisation and supervision. RD: data interpretation and reporting, expert knowledge, methodology, and writing—reviewing and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the New South Wales Police Force.

ACKNOWLEDGMENTS

We would like to thank and acknowledge the NSW Health Pathology Forensic and Analytical Science Services

Laboratory and the de-identified service providers for their participation in the project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.568701/full#supplementary-material>

REFERENCES

- Al-Asfi, M., McNevein, D., Mehta, B., Power, D., Gahan, M. E., and Daniel, R. (2017). Assessment of the precision ID ancestry panel. *Int. J. Legal Med.* 132, 1581–1594. doi: 10.1007/s00414-018-1785-9
- Bhopal, R., and Donaldson, L. (1998). White, European, Western, Caucasian, or what? Inappropriate labeling in research on race, ethnicity, and health. *Am. J. Public Health* 88, 1303–1307. doi: 10.2105/ajph.88.9.1303
- Cheung, E. Y., Elizabeth, G. M., and McNevein, D. (2018). Predictive DNA analysis for biogeographical ancestry. *Austr. J. Forensic Sci.* 50, 651–658.
- Cino, J. G. (2016). Tackling technical debt: managing advances in DNA technology that outpace the evolution of the law. *J. Civil Legal Sci.* 5, 1–15.
- Daniel, R. (2016). "Personal communication," in *Human Identification and Solutions Conference*, Barcelona.
- Enserink, M. (2011). Can this DNA sleuth help catch criminals? *Science* 331, 838–840. doi: 10.1126/science.331.6019.838
- Forensic Molecular Biology Department of Erasmus MC (n.d). *IrisPlex, and HirisPlex DNA Phenotyping Webtool User Manual Version 1.0. Developmental Validation of the HirisPlex System: DNA-Based eye, and Hair Colour Prediction for Forensic, and Anthropological Usage*. Available at: <https://hirisplex.erasmusmc.nl/> (accessed November 14 2017).
- Freedman, B. J. (1984). Caucasian. *Br. Med. J.* 288, 696–698.
- Jin, S., Maretta, C., Margaret, H., Gerry, A., James, M. M., Sobia, M., et al. (2018). Implementing a biogeographic ancestry inference service for forensic casework. *Electrophoresis* 39, 2757–2765. doi: 10.1002/elps.201800171
- Kayser, M. (2015). Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes. *Forensic Sci. Int. Genet.* 18, 33–48. doi: 10.1016/j.fsigen.2015.02.003
- Kayser, M., and Schneider, P. M. (2012). Reply to bracking off population does not advance ethical reflection on EVCs: a reply to kayser and schneider. *Forensic Sci. Int. Genet.* 6, e18–e19.
- National Health and Medical Research Council (2015). *National Statement on Ethical Conduct in Human Research (2007) - Updated 2015. May*. Available at: <https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research#block-views-block-file-attachments-content-block-1> (accessed August 8th, 2019).
- Phillips, C., McNevein, 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.
- Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gómez-Tato, A., Álvarez-Dios, J., et al. (2009). Ancestry analysis in the 11-M madrid bomb attack investigation. *PLoS One* 4:e6583. doi: 10.1371/journal.pone.0006583
- Raymond, J., Atwood, L., and Sears, A. (2017). *DNA Phenotyping: Service Provider Trial. Project Report*. Internal Report. Sydney: Science and Research Unit, New South Wales Police Force.
- Samuel, G., and Prainsack, B. (2018). Forensic DNA phenotyping in Europe: views on the ground from those who have a professional stake in the technology. *New Genet. Soc.* 38, 119–141. doi: 10.1080/14636778.2018.1549984
- Schneider, P. M., Barbara, P., and Manfred, K. (2019). The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry. *Deutsches Arzteblatt International* 51–52, 873–880.
- Scudder, N., Kelty, S. F., Grant, J. B., Montgomerie, C., Walsh, S. J., Robertson, J., et al. (2020). *Differing Perception of DNA Evidence and Intelligence Capabilities in Criminal Investigations*. Available at: <https://www.preprints.org/manuscript/202002.0004/v1> (accessed April 7, 2020).
- Touchette, N. (2003). *Genome News Network. 13 June*. Available online at: http://www.genomenewsnetwork.org/articles/06_03/serial.shtml (accessed April 15, 2020).
- United Nations (1999). *Standard Country or Area Codes for Statistical use (M49)*. New York, NY: Statistics Division of the United Nations Secretariat.
- Walsh, S., Liu, F., Ballantyne, K. N., Van Oven, M., Lao, O., and Kayser, M. (2011a). IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci. Int. Genet.* 5, 170–180. doi: 10.1016/j.fsigen.2010.02.004
- Walsh, S., Lindenbergh, A., Zuniga, S. B., Sijen, T., De Knijff, P., Kayser, M., et al. (2011b). Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. *Forensic Sci. Int. Genet.* 5, 464–471. doi: 10.1016/j.fsigen.2010.09.008
- Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., et al. (2013). The HIRISplex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci. Int. Genet.* 7, 98–115. doi: 10.1016/j.fsigen.2012.07.005
- Walsh, S., Wollstein, A., Liu, F., Chakravarthy, U., Rahu, M., Seland, J. H., et al. (2012). DNA-based eye colour prediction across Europe with the IrisPlex system. *Forensic Sci. Int. Genet.* 6, 330–340. doi: 10.1016/j.fsigen.2011.07.009

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

Copyright © 2021 Atwood, Raymond, Sears, Bell and Daniel. 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.



Combined Low-/High-Density Modern and Ancient Genome-Wide Data Document Genomic Admixture History of High-Altitude East Asians

Yan Liu^{1†}, Mengge Wang^{2*†}, Pengyu Chen^{3,4†}, Zheng Wang², Jing Liu², Lilan Yao^{3,4}, Fei Wang², Renkuan Tang⁵, Xing Zou^{2*} and Guanglin He^{2,6*}

OPEN ACCESS

Edited by:

Cemal Gurkan,
Turkish Cypriot DNA Laboratory
(TCDL), Cyprus

Reviewed by:

Rita Rasteiro,
University of Bristol, United Kingdom
Kenneth K. Kidd,
School of Medicine Yale University,
United States

*Correspondence:

Mengge Wang
menggewang2021@163.com
Xing Zou
forensiczx@163.com
Guanglin He
guanglinhesu@163.com

[†]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: 11 July 2020

Accepted: 05 January 2021

Published: 11 February 2021

Citation:

Liu Y, Wang M, Chen P, Wang Z,
Liu J, Yao L, Wang F, Tang R,
Zou X and He G (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

¹School of Basic Medical Sciences, North Sichuan Medical College, Nanchong, China, ²Institute of Forensic Medicine, West China School of Basic Science and Forensic Medicine, Sichuan University, Chengdu, China, ³Key Laboratory of Cell Engineering in Guizhou Province, Affiliated Hospital of Zunyi Medical University, Zunyi, China, ⁴Center of Forensic Expertise, Affiliated Hospital of Zunyi Medical University, Zunyi, China, ⁵Department of Forensic Medicine, College of Basic Medicine, Chongqing Medical University, Chongqing, China, ⁶Department of Anthropology and Ethnology, Institute of Anthropology, National Institute for Data Science in Health and Medicine, and School of Life Sciences, Xiamen University, Xiamen, China

The Tibetan Plateau (TP) is considered to be one of the last terrestrial environments conquered by the anatomically modern human. Understanding of the genetic background of highland Tibetans plays a pivotal role in archeology, anthropology, genetics, and forensic investigations. Here, we genotyped 22 forensic genetic markers in 1,089 Tibetans residing in Nagqu Prefecture and collected 1,233,013 single nucleotide polymorphisms (SNPs) in the highland East Asians (Sherpa and Tibetan) from the Simons Genome Diversity Project and ancient Tibetans from Nepal and Neolithic farmers from northeastern Qinghai-Tibetan Plateau from public databases. We subsequently merged our two datasets with other worldwide reference populations or eastern ancient Eurasians to gain new insights into the genetic diversity, population movements, and admixtures of high-altitude East Asians via comprehensive population genetic statistical tools [principal component analysis (PCA), multidimensional scaling plot (MDS), STRUCTURE/ADMIXTURE, *f*₃, *f*₄, *qpWave*/*qpAdm*, and *qpGraph*]. Besides, we also explored their forensic characteristics and extended the Chinese National Database based on STR data. We identified 231 alleles with the corresponding allele frequencies spanning from 0.0005 to 0.5624 in the forensic low-density dataset, in which the combined powers of discrimination and the probability of exclusion were 1–1.22E-24 and 0.999999998, respectively. Additionally, comprehensive population comparisons in our low-density data among 57 worldwide populations via the Nei's genetic distance, PCA, MDS, NJ tree, and STRUCTURE analysis indicated that the highland Tibeto-Burman speakers kept the close genetic relationship with ethnically close populations. Findings from the 1240K high-density dataset not only confirmed the close genetic connection between modern Highlanders, Nepal ancients (Samdzong, Mebrak, and Chokhopani), and the upper Yellow River Qijia people, suggesting the northeastern edge of the TP served as a geographical corridor for ancient population migrations and interactions between highland and lowland regions, but also evidenced that late Neolithic

farmers permanently colonized into the TP by adopting cold-tolerant barley agriculture that was mediated *via* the acculturation of idea *via* the millet farmer and not *via* the movement of barley agriculturalist as no obvious western Eurasian admixture signals were identified in our analyzed modern and ancient populations. Besides, results from the *qpAdm*-based admixture proportion estimation and *qpGraph*-based phylogenetic relationship reconstruction consistently demonstrated that all ancient and modern highland East Asians harbored and shared the deeply diverged Onge/Hoabinhian-related eastern Eurasian lineage, suggesting a common Paleolithic genetic legacy existed in high-altitude East Asians as the first layer of their gene pool.

Keywords: 1240K dataset, ancient genomes, population history, forensic genetics, short tandem repeats, genetic polymorphism, East Asian highlander

INTRODUCTION

East Asia, one of the oldest centers of plant and animal domestication, is home to almost one-quarter of the world's population and encompasses substantial genetic, cultural, linguistic, and physical diversity. Understanding the peopling processes of East Asia or some unique harsh environment area is therefore of interest for elucidating how these extensive diversities arose and evolved. However, the comprehensive genetic history of East Asia is poorly understood due to the lack of ancient DNA from a denser genetic sampling or sparse sampling of modern East Asians and combined analyses of spatiotemporally diverse East Asian populations (Lu et al., 2016; Yao et al., 2017; Bai et al., 2018; He et al., 2020). Generally, patterns of genetic relatedness among present-day East Asians, especially for Han Chinese, run along a north-south cline (Qin et al., 2014; Chiang et al., 2018; Chen et al., 2019b; Gao et al., 2020b). Recent ancient genome-wide data of 26 ancient northern and southern East Asians (including Shandong Houli and Fujian Tanshishan cultural backgrounds) spanning 9,500–300 years ago indicated human population shifts and admixture in northern and southern China and confirmed the genetic division between northern and southern East Asians since early Neolithic (Yang et al., 2020). Wang et al. also reported genome-wide data from 383 modern and 191 ancient East Asians dating to around 6,000 BCE–1,000 CE and illuminated the dispersal models of the ancestors of Mongolic, Tungusic, Sino-Tibetan, Austronesian, Tai-Kadai, and Austroasiatic languages and showed the complex population interactions among different ancient East Asians (Wang et al., 2020). Additionally, Ning et al. reported 55 ancient genomes dating to 7,500–1,700 years ago from the Yellow River (Henan Yangshao, Longshan, and Shangzhou cultures and Qinghai Qijia culture), West Liao River (Hongshan and Xiajiadian cultures), and Amur River (Haminmangha culture) basins and illustrated a link between changes in subsistence strategy and human activities (migration and admixture; Ning et al., 2020). However, these ancient genomes from the lowland East Asians showed a finer-scale landscape of population origin, diversification, and admixture in the lowland regions, and the population genetic admixture

history of the highland region kept underrepresented and unclear due to the sparse genetic sampling of modern and ancient populations from the Qinghai-Tibet Plateau, which impedes our ability to connect temporally and geographically dispersed ancient East Asians and modern Tibetans.

The Qinghai-Tibet Plateau, also called the Tibetan Plateau (TP), a high-altitude arid steppe bounded by the world's tallest mountains, represents one of the most challenging environments with low temperature and hypobaric hypoxia for human settlement. As one of the last populated areas occupied by modern humans, the exact timing of the peopling of the TP and the migration trajectories of Tibetans have appealed to growing academic interests. The recovered paleoproteomic results of a Xiahe Denisovan mandible from the TP indicated that archaic hominins occupied the TP in the Middle Pleistocene epoch and successfully adapted to the high-altitude environments with the accumulation of *Endothelial PAS domain protein 1* (*EPAS1*) adaptive alleles (Chen et al., 2019a). Archeological investigations documented that the earliest modern human foraging of the TP may have begun at least ~40 to 30 thousand years ago (kya; Zhang et al., 2018). Considerable progress on the anthropological, archeological, and genetic perspectives of archaic and modern humans provided the conclusive evidence in support of the Paleolithic initial peopling of the TP and indicated that the permanent human occupation had taken place around 3.6 kya, which was most likely facilitated by the spread of barley/wheat-based agriculture (Qi et al., 2013; Chen et al., 2015; Lu et al., 2016; Meyer et al., 2017; Li et al., 2019b; Gao et al., 2020a; Ren et al., 2020). The matrilineal evidence revealed Tibetan-prevailing lineages of A11a1a and M9a1a1c1b1a and demonstrated that the ancestry of Tibetans could largely be traced back to the Neolithic millet farmers from northern China (Zhao et al., 2009; Qin et al., 2010; Qi et al., 2013; Li et al., 2019b; Wang et al., 2020). The coalescence ages of Tibetan-specific Y-chromosomal lineages served as another strong evidence that the earlier settlers on the TP could have survived in the Last Glacial Maximum (LGM) and contributed to the gene pool of present-day Tibetan populations (Qi et al., 2013). It also revealed that Neolithic expansions of low-altitude agriculturalists had a prominent impact on the genomic makeup of modern Tibetans (Qi et al., 2013). Besides, genome-wide

data revealed a relatively closer genetic affinity between Tibetans and Han Chinese and indicated that Tibetans arose from a mixture of multiple ancestral gene pools and most of the Tibetan gene pool could be attributed to the post-LGM arrivals of Neolithic ancestry (Qi et al., 2013; Lu et al., 2016; Yao et al., 2017). Linguistic study from the Bayesian phylogenetic analysis of the Sino-Tibetan language family has suggested that the Tibeto-Burman-speaking populations diverged from Han Chinese with an average coalescence age of approximately 5.9 kya (Zhang et al., 2019). Furthermore, genetic observations based on forensically related markers also revealed the consistent phylogenetic relationships between Tibetan and other geographically or ethnically different groups (He et al., 2018a,b; Wang et al., 2018b, 2020; Zou et al., 2018; Li et al., 2019a).

Taken together, current archeological, anthropological, genetic, and linguistic findings suggested that the initial Paleolithic occupation of the TP combined with later multiple migrations at different times and from different regions may have created the complicated and mosaic demographic history of Tibetans. However, available genetic data are insufficient to address the discrepancy between demographic history constructed by different regional studies and hamper the exploration of genetic variations of Tibetans based on the forensically related markers. Hence, extending the existing forensic reference database and dissecting the genetic differentiation among different Tibetan groups or between Tibetans and other reference populations based on the combined resolution of modern and ancient genomes is indispensable. Here, we mainly aimed to focus on the following topics: (I) explore the pattern of genetic diversity of highland East Asian based on short tandem repeat (STR) and single nucleotide polymorphism (SNP) data; (II) dissect the potential gene flow events between highland Tibetan-Burman speakers and close lowland East Asian populations; (III) explore whether there is a genetic continuity between modern Highlanders and ancient populations who were linked *via* the archeologically attested similarities of cultures from Nepal and upper Yellow River (Qijia people) and further explore the extent to which it was mediated *via* the population movement through a northeast geographical corridor; and (IV) evaluate to what extent of the barley/wheat agriculture spread in the Ganqing region was mediated *via* cultural diffusion or demic diffusion from the Fertile Crescent.

MATERIALS AND METHODS

Sample Collection

Here, we carried out the present study in 1,089 unrelated Tibetan individuals (593 males and 496 females) residing in Nagqu – the northeastern prefecture-level city of Tibet Autonomous Region (Figure 1A). All participants enrolled in the present study have signed the written informed consent form and are required to be the indigenous Tibetan people. Bloodstains were collected from people with no mixed marriage with people of other ethnic groups. This project was performed in accordance with the recommendations of the Declaration of Helsinki (Nicogossian et al., 2014) and approved by the

Ethics Committees of North Sichuan Medical College and Zunyi Medical University.

DNA Extraction, Quantification, and Genotyping

Human genomic DNA was extracted using the PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific) and quantified by employing the NanoDrop-2000 on the basis of the manufacturer's instructions. Twenty autosomal short tandem repeats (A-STRs) recommended by the Chinese National Database (CND) as well as two gender-determining genes (Amelogenin and Y-indel) were amplified simultaneously using the STRtyper-21G PCR assay on a ProFlex PCR System (Thermo Fisher Scientific) following the manufacturer's recommendation. ABI 3130 Genetic Analyzer (Thermo Fisher Scientific) was utilized to separate the PCR products, and the GeneMapper ID-X v.1.4 software was used to visualize the electrophoresis results.

Data Analysis

Analysis of Genetic Variations Based on Low-Density STRs

The online software of the STR Analysis for Forensics (STRAF; Gouy and Zieger, 2017) was adopted to evaluate the allelic frequencies and forensic statistical parameters of 20 A-STRs. The exact tests of linkage disequilibrium (LD) and Hardy-Weinberg equilibrium (HWE), as well as evaluation of the heterozygosity indexes (observed heterozygosity: H_o ; and expected heterozygosity: H_e), were conducted using the Arlequin v.3.5.2.2 (Excoffier and Lischer, 2010). Nei's pairwise genetic distances between Nagqu Tibetan and 56 worldwide reference populations were estimated *via* the Gendist package implemented in the PHYLIP v.3.695 (Retief, 2000) and imported into R software¹ for heatmap plotting. Frequency-based principal component analysis (PCA) of the 17 A-STRs among 57 worldwide populations (the detailed codes of population information is listed in **Supplementary Table S1** and **Figure 1A**) was carried out using the Multivariate Statistical Package (MVSP) software v.3.22 (Kovach, 2013). The Nei's distance matrix was then applied to perform the multidimensional scaling (MDS) analysis using the IBM SPSS v.21.0 and reconstruct a neighbor-joining (NJ) tree *via* the Molecular Evolutionary Genetics Analysis v.7.0 (Mega 7.0; Kumar et al., 2016). Furthermore, we employed the STRUCTURE v.2.3.4.21 (Evanno et al., 2005) to dissect the genetic similarity among 3,287 individuals from 11 Chinese populations with K values ranging from 2 to ~6 under the “correlated allele frequencies” and “LOCPRIOR” models.

High-Density Genome-Wide Data Analysis

We retrieved 1,233,013 SNPs of Tibetan and Sherpa from the Simons Genome Diversity Project (Mallick et al., 2016); ancient Tibetan genome-wide SNP data from eight Nepal individuals (Jeong et al., 2016) with cultural backgrounds of Chokhopani, Samdzong, and Mebrak; and 11 late Neolithic to Iron Age

¹<https://www.r-project.org/>

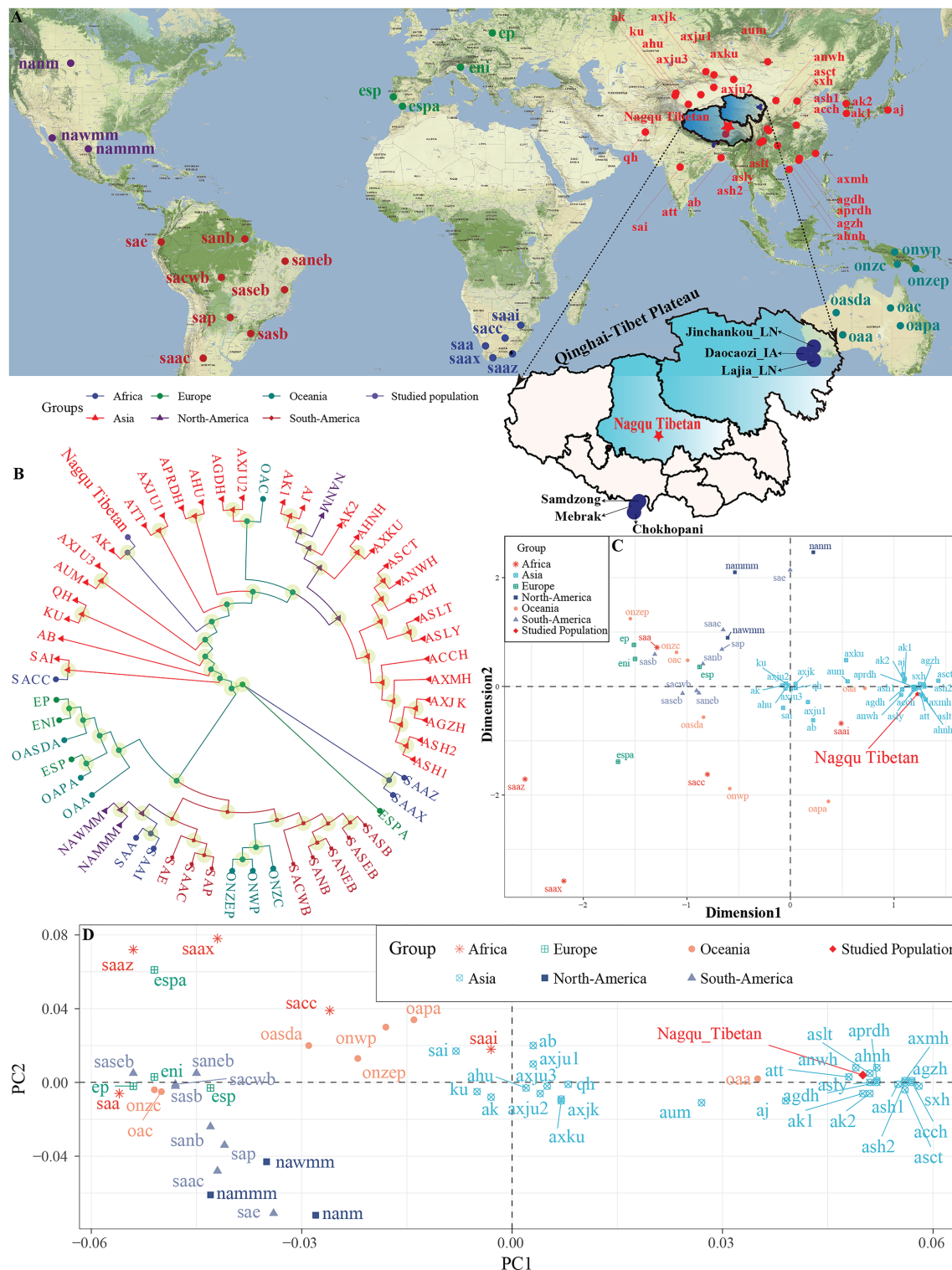


FIGURE 1 | Sampling geographical region and patterns of the genetic relationship between Tibetan and worldwide reference populations based on the STR low-density dataset. **(A)** Geographical position of Nagqu City and Qinghai-Tibet Plateau and other included reference modern and ancient populations. **(B)** Phylogenetic relationship between Nagqu Tibetan and other 56 worldwide reference populations based on the pairwise genetic distance. **(C)** Genetic relationship between Nagqu Tibetan and other 56 worldwide reference populations revealed by the multidimensional scaling plots. **(D)** Two-dimensional scaling plots of the top two components in PCA analysis. The full population names (codes) are submitted in **Supplementary Table S1**.

from the northeastern edge of the TP (Ning et al., 2020). We then merged the aforesaid data with other publicly available data of modern and ancient East Asians (Patterson et al., 2012; Yang et al., 2017, 2020; Lipson et al., 2018; Jeong et al., 2019; Liu et al., 2020; Ning et al., 2020). The geographical position and corresponding archeological periods are provided in **Supplementary Figure S1**. We then pruned SNPs in strong linkage disequilibrium by applying PLINK v.1.9 (Chang et al., 2015) with parameters of `--indep-pairwise 200 25 0.4`. We performed model-based clustering analysis using the ADMIXTURE v.1.3.0 (Alexander et al., 2009) with the 10-fold cross-validation (`--cv = 10`), presupposing the number of K values ranging from 2 to 8 in 100 bootstraps with different random seeds. The caution is that the clusters obtained in this model-based analysis are only “similarity” measures based on complex algorithms, and individuals are assigned to a cluster in whole or in part, which can be used to explore the genetic similarities and differences based on the shared components among them. We computed F_{st} values using the EIGENSOFT with the default parameters of `inbreed: YES` and `fstonly: YES`.

We computed outgroup f_3 -statistics using the *qp3Pop* program of the ADMIXTOOLS package (Patterson et al., 2012) and looked for evidence of maximized shared genetic drift. We also conducted admixture f_3 - and f_4 -statistics using the *qp3Pop* and *qpDstat* packages from the same program with the default parameters to assess the potential admixture signals from different source populations into the targeted populations. We calculated standard errors using the weighted block jackknife approach.

Applying the covariance of the allele frequency profiles as input, we ran TreeMix v.1.13 (Pickrell and Pritchard, 2012) with migration events varying from 0 to 8 to generate the topology with the maximum likelihood. Based on the results of the f -statistics, admixture graph modeling was carried out using the *qpGraph* software as implemented in the ADMIXTOOLS using central African of Mbuti as an outgroup (Fu et al., 2015). We applied the programs *qpWave* and *qpAdm* from the ADMIXTOOLS to model the targets as a combination of putatively selected source populations and estimate the ancestry proportions by solving a matrix of f_4 -statistics (Haak et al., 2015). We used a batch of outgroups and basic phylogenetic relationships followed Wang's model (Wang et al., 2020), which represented modern and ancient global genomic variations and provided a good resolution for distinguishing Tibetan Highlanders.

RESULTS

Genetic Diversity, Forensic Features, and STR-Based Population Comparisons

We genotyped 20 autosomal STRs and two Y-linked genetic loci for sex determination in 1,089 unrelated Nagqu highland Tibetans using the new generation of STRtyper-21G PCR amplification system. As displayed in **Supplementary Table S2**, one out of 20 STR loci (D3S1358) was deviated from the

HWE after applying the Bonferroni correction ($0.05/20 = 0.0025$), and LD was observed in the locus pair of TPOX-Penta E (**Supplementary Table S3**, 0.00020) after conducting the multiple tests of Bonferroni correction ($0.05/190 = 0.00026$). A total of 231 alleles were identified with the corresponding allelic frequencies spanning from 0.0005 to 0.5624 (**Supplementary Table S4**). The values of H_o and H_e , as well as forensic parameters, including discrimination power (DP), probability of exclusion (PE), and typical paternity index (TPI) are presented in **Supplementary Table S2**. The H_o varied from 0.6217 to 0.9183, and the H_e spanned from 0.6038 to 0.9182. The measured values of DP and PE were in the range of 0.7854–0.9865 and 0.3177–0.8329, respectively. The value TPI varied from 1.3216 to 6.1180. Additionally, the combined power of discrimination (CPD) value reached $1-1.22E-24$ in Nagqu Tibetan, and the value of the combined probability of exclusion (CPE) was 0.999999998.

We explored the genetic relationships between Nagqu Tibetan and other 56 reference populations *via* the pairwise genetic distances, PCA, MDS, and NJ tree. The pairwise genetic distances among 57 populations are listed in **Supplementary Table S5** and **Supplementary Figure S2**. The Chengdu Tibetan (ASCT) was identified as the genetically closest population to Nagqu Tibetan (0.012), followed by Liangshan Tibetan (ASLT, 0.0134) and Liangshan Yi (ASLY, 0.0146). The African AmaXhosa (SAAX) shows the largest genetic differences with Nagqu Tibetan (0.2097). Subsequently, MDS and NJ tree (**Figures 1B,C**) were depicted based on the pairwise genetic distance matrix. On the NJ tree (**Figure 1B**), all 57 worldwide populations were roughly grouped into two clades: Asian groups and other continental groups. It is interesting to find that the Nagqu Tibetan first clustered with Akto Kyrgyz (AK) and then clustered with Tibet Tibetan (ATT). There needed to be more caution that the NJ-based bifurcating tree just provided the basic framework of population relationship not only due to an NJ tree is an approximation to a fully additive tree but also the fitting process ignored the potential exited admixture events. Thus, TreeMix and *qpGraph*-based phylogenetic relationship reconstruction needed to be conducted and will be discussed in detail in the following contents. As displayed in **Figure 1C**, the Asian populations clustered close to each other, which can be further grouped into Sino-Tibetan (ST) cluster and Altai-Turkic (AT) cluster, and the North/South American populations formed a relatively looser cluster. Conversely, other continental populations were scattered in the left and lower right quadrants. Nagqu Tibetan was located close to Tibet Tibetan (ATT) and Liangshan Tibetan (ASLT), which was also surrounded by Han Chinese populations. PCA based on the top six components could explain 74.52% variance (PC1 to PC6: 34.18, 14.49, 11.70, 6.16, 5.00, and 2.99%). PC1 (**Figure 1D** and **Supplementary Figure S2A**) could distinguish the Asian populations from the others; besides, the Asian groups could be divided into two main clusters by the PC1: one contained Xinjiang and South Asian populations, and the other comprised Han Chinese, Hui, Yi, and Tibetan populations. The other five components could not separate any continental groups from the others (**Supplementary Figures S3B–D**).

Generally, the patterns revealed by MDS and NJ tree were in accordance with those observed in the PCA and heatmap. To directly dissect the Nagqu Tibetan ancestry component and explore the genetic similarity based on the shared ancestral components with different predefined K values, we conducted the STRUCTURE analysis assuming 2–6 predefined clusters (**Supplementary Figure S4**). We found that the fitted model with three clusters had the optimal K value. At $K = 2$, we identified two distinct components maximized, respectively, in ST and AT populations. At $K = 3$, population substructures of Han Chinese and Tibeto-Burman (TB) populations were observed within ST populations. Geographically, different components within the same language family gradually appeared with the increase of K values and the proportions of shared components were variable within ethnically different groups. Nagqu Tibetan consistently harbored a unique component and showed a closer genetic affinity with Chengdu Tibetan and Lhasa Tibetan.

ADMIXTURE, f_3 -Statistics, and Phylogeny Reconstruction Among Highlanders, Eurasian Modern/Ancient References Based on the 1240K SNPs

To study the demographic history and deep population history of East Asian Highlanders of Tibetan and Sherpa, we used the Tibetan and Sherpa individuals included in the Simons Genome Diversity Project (SGDP) as the new studied subject. We merged them with other publicly available modern and ancient Eurasian genomes based on the 1240K overlapping SNPs. The final dataset included 44 populations: 356 modern individuals from 21 East Asian groups and 1 central African Mbuti and 112 Chinese ancients from 22 spatiotemporally diverse archeological sites (**Supplementary Figure S1**). We pruned 335,589 linked SNPs from 1,233,013 SNPs and remained 897,424 markers for model-based ancestry sources modeling. Model-based ADMIXTURE results also showed the population similarities with different predefined genetic clusters. Individual and average population cluster-specific compositions are presented in **Figure 2A**; **Supplementary Figures S5, S6**; **Supplementary Table S7**. The fitted model with two predefined clusters separated Mbuti from other East Asians. The optimal cluster sources could be modeled and obtained when the three predefined genetic clusters ($K = 3$) were assumed (cross-validation error = 0.8832). Also, this three-population model showed that yellow ancestry was enriched in Taiwan Iron Age population (average proportion is 0.954 as blue component in **Supplementary Figure S7**), which also existed with a higher proportion (larger than 0.772) in the coastal late Neolithic southern East Asians (Tanshishan_LN and Xitoucun_LN) and modern southern Chinese Austronesian (Ami) and Tai-Kadai speakers (Xishuangbanna Dai). The other East Asian-dominant component (orange in **Figure 2A**, $K = 3$) maximized in the inland middle Neolithic northern East Asian Miaozigou individuals associated with Miaozigou culture (0.983), followed by the late Neolithic Shima and Neolithic Wuzhuangguoliang people in Shaanxi and other ancient Tibetan

and northern Chinese ancients (larger than 0.869). Modern Tibetan harbored 0.866 Miaozigou-related component and others from Hanben- or Mbuti-like component, and Sherpa derived 0.878 of their components related to this group. We identified two southern East Asian components when the model of four predefined cluster sources was used: island/coastal southern East Asian components maximized in Taiwan Iron Age Hanben people (0.962) and late Neolithic Xitoucun and Tanshishan (0.741 and 0.713, respectively) and inland southern East Asian component enriched in Tai-Kadai Dai which also existed with a high proportion in Chinese southern Tibeto-Burman Lahu, modern Austronesian Ami. and Hmong-Mien Miao and She. Similar to the patterns in $K = 3$, the third component was maximized in the inland northern Neolithic people. Based on the shared component in **Figure 2A**, modern Tibetan shared more components with Highland Sherpa.

We subsequently estimated the shared genetic drift between the highland East Asians (Tibetan and Sherpa) and other 350 lowland modern East Asian individuals from 20 populations, 118 lowland ancient East Asians from 33 populations, and 8 highland East Asian individuals from 3 Nepal populations (2,125-year-old Mebrak, 1,500-year-old Samdzong, and 2,700-year-old Chokhopani) via the outgroup f_3 -statistics in the form of $f_3(\text{Reference populations, Tibetan/Sherpa; Mbuti})$. Pairwise-shared genetic drift among 63 ancient and 22 modern East Asian populations were also calculated via $f_3(\text{Reference ancient/modern populations1, Reference ancient/modern populations2; Mbuti})$ and submitted in **Supplementary Table S7**. The observed larger f_3 values or green color in **Figure 2B** denoted more shared ancestry among two reference populations, and smaller f_3 values or red color meant less shared ancestry among them. The red color with Uyghur and the green color with the late Neolithic Wuzhuangguoliang people observed in the heatmap, respectively, showed their genomic differentiation and similarities with reference East Asians. The cluster patterns in the heatmap showed that Tibetan clustered with Nepal ancients and kept a close relationship with Sherpa. Focused on the genetic variations of Sherpa and Tibetan (**Figure 2C**), we found that the top shared ancestry with highland Tibetan and Sherpa was provided by Shaanxi Wuzhuangguoliang Neolithic people (0.3096 with Tibetan and 0.3121 with Sherpa). The indexes between Tibetan and four high-altitude populations (three Nepal ancients and one modern Sherpa) were larger than 0.3034, followed by late Neolithic Qijia people from the upper Yellow River basin (Jinchankou and Lajia) and modern lowland Tibeto-Burman-speaking Naxi and Yi and other northern modern and ancient populations. Consistent patterns of genetic affinity were observed in the relationship between Sherpa and other East Asian-associated reference populations.

We subsequently estimated admixture signals of Highland East Asians via admixture f_3 -statistics in the form of $f_3(\text{Source population1, Source population2; targeted populations of Tibetan/Sherpa})$. The observed statistically significant negative f_3 values with absolute Z scores larger than three indicated that the targeted investigated population was a mixed population with the possible ancestral populations related to the two used sources. No negative f_3 values were identified in $f_3(\text{Source$

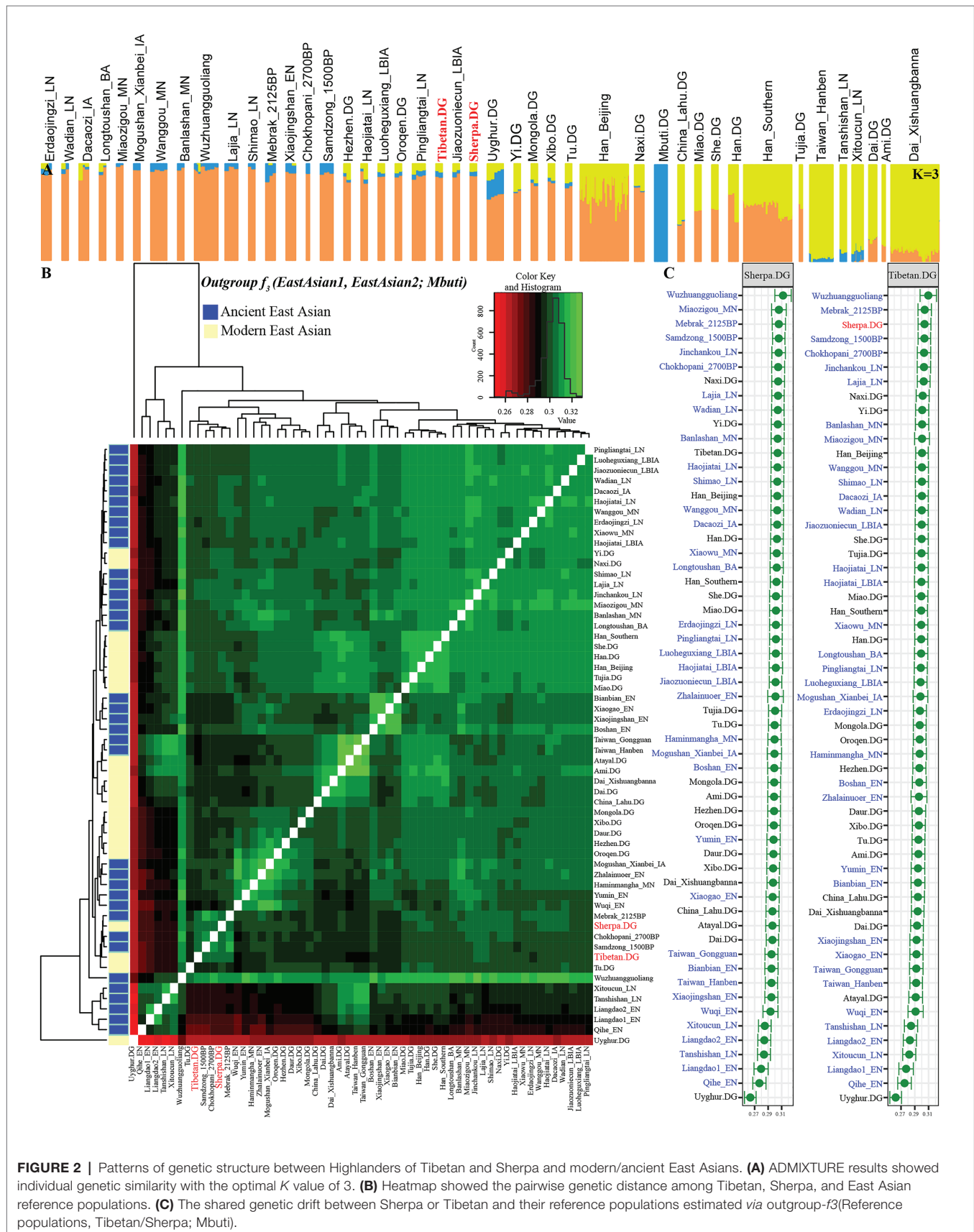


FIGURE 2 | Patterns of genetic structure between Highlanders of Tibetan and Sherpa and modern/ancient East Asians. **(A)** ADMIXTURE results showed individual genetic similarity with the optimal K value of 3. **(B)** Heatmap showed the pairwise genetic distance among Tibetan, Sherpa, and East Asian reference populations. **(C)** The shared genetic drift between Sherpa or Tibetan and their reference populations estimated via outgroup- f_3 (Reference populations, Tibetan/Sherpa; Mbuti).

population1, Source population2; Sherpa) among 1,653 pairs of modern and ancient East Asians, but eight population pairs with negative f_3 values were observed in f_3 (Source population1, Source population2; Tibetan) with one source from Nepal ancients and the other from modern/ancient northern East Asians (Supplementary Tables S8, S9). We should be cautious that the observed negative f_3 values with Z scores were larger than negative three. Thus, compared with obvious admixture signatures from northern and southern East Asians observed in the lowland East Asians, the highland Tibetan and Sherpa showed their unique genetic structure, which is different from other lowland East Asians.

We also calculated Wright's fixation index F_{st} among 42 modern and ancient populations (Supplementary Table S10), with the exception of the unexpected F_{st} values caused by the unbalanced sample size; Tibetan possessed the smallest genetic distance with Sherpa (0.0173), followed by Tu (0.0195), late Neolithic Pingliangtai (0.0236), Yi (0.0272), and Naxi (0.0289). However, Tibetan showed a close genetic relationship first with Chinese lowland Tibeto-Burman-speaking populations and then with Sherpa, which showed the more genetic influence or closer links between Tibetan and Lowland East Asian populations. Our F_{st} -based heatmap in Figure 3A revealed that modern and ancient populations showed their close genetic relationship within themselves. To further explore the phylogenetic relationships between Highlanders and lowland East Asians, we reconstructed three different phylogeny trees (Figures 3B–D). The first tree shown in Figure 3B was constructed using the genetic matrix of one minus outgroup- f_3 values ($1 - f_3$) and NJ algorithm. Here, we could identify southern Neolithic to Iron Age populations grouped together and then grouped with southern Chinese modern Austronesian, Tai-Kadai, Hmong-Mien, and Sinitic language speakers, which formed the southern East Asian branch. Hanben and Gongguan people from Taiwan kept the closest relationship with modern Austronesian Ami and Atayal. Sherpa and Tibetan possessed a strong genetic affinity and grouped first with three Nepal ancients and then with lowland Tibeto-Burman-speaking Tu, Naxi, and Yi and formed the TP branch. The observed Tibeto-Burman branch showed a close genetic relationship between modern lowland/highland Tibeto-Burman language speakers and ancient highland Nepal ancients. The northern ancient branch comprised early Neolithic to Iron Age individuals from Shandong and Henan provinces in the middle and lower Yellow River basin and from Shaanxi and Qinghai provinces in the upper Yellow River basin and West Liao River. Amur River ancient clustered with modern Tungusic and Mongolic speakers formed an Amur branch. The overall patterns observed in the f_3 -based phylogenetic relationship showed the TP branch was placed in the intermediate position between the northern East Asian branch and the southern East Asian branch, but far away from the Amur branch. The second NJ tree based on the F_{st} genetic distance matrices clustered one modern population branch and one ancient population branch (Figure 3C). Although there was separation between modern and ancient populations in the clustered results, we could also identify that Hanben was grouped with modern Ami, late Neolithic Pingliangtai

clustered with Yi and She, and the studied Tibetan and Sherpa Highlanders grouped with 1,500-year-old Samdzong. In the third one, we considered the gene flow events among the patterns of population splits and admixture among East Asians and reconstructed one maximized likelihood tree (Figure 3D). We found highland Tibetan and Sherpa grouped with their geographically/linguistically close populations. Similar clustered patterns were identified among southern modern and ancient East Asians and northern modern and ancient East Asians. No obvious gene flow events into Tibetans or from Tibetans into other East Asians were identified.

Genomic Affinity and Differentiation Between Sherpa and Tibetan Revealed by f_4 -Statistics

To comprehensively evaluate the genetic relationships between highland Tibetan and Sherpa, we performed four-population comparisons (f_4 -statistics) to explore the differentiated shared drifts between Highlanders and lowland East Asian reference groups compared with other East Asian reference groups in the form of f_4 (Modern/Ancient Chinese population1, Modern/Ancient Chinese population2; Tibetan/Sherpa, Mbuti). The observed significant negative f_4 values with the absolute Z scores larger than three (green color in the heatmap) denoted that our studied Tibetan and Sherpa shared more genetic drifts with Modern/Ancient Chinese population2 relative to the Modern/Ancient Chinese population1; otherwise, significant positive f_4 values (red color in the heatmap) denoted more shared alleles between Highlanders and Modern/Ancient Chinese population1 rather than Modern/Ancient Chinese population2. No significant negative or positive f_4 values (Z scores ranging from -3 to 3, gray color) denoted two Chinese reference populations formed one clade relative to our studied Highlanders. As shown in Supplementary Table S11 and Figure 4, f_4 (Xinjiang ancient/modern populations, other modern/ancient East Asians; Tibetan, Mbuti) was conducted to explore the relationships between Highlanders and northwestern Chinese populations (modern Uyghur and Iron Age Shirenzigou people). The results of significant negative f_4 values showed that Tibetan shared more derived alleles with both northern and southern Neolithic to present-day East Asians than with Xinjiang Iron Age to modern populations, which suggested little genetic materials associated with western Eurasian in Tibetans (Ning et al., 2019). Compared with 40,000-year-old Tianyuan people, Tibetan shared more alleles with modern Uyghur [f_4 : 3.546*standard error (SE)], Shirenzigou_IA (4.624*SE), and Shirenzigou_IA_E (8.593*SE) via f_4 (Xinjiang populations, China_Tianyuan; Tibetan, Mbuti). Compared with Shirenzigou individuals with stronger western Eurasian affinity, we found that Tibetan shared more alleles with modern Uyghur, Shirenzigou_IA, and Shirenzigou_IA_E. Moreover, genetic similarities between Tibetan and Shirenzigou_IA_E were further confirmed via f_4 (Shirenzigou_IA_E, Shirenzigou_IA/Uyghur; Tibetan, Mbuti) = 5.827*SE/9.473*SE. To study the genetic links between Tibetans and early East Asians, we carried out f_4 (Early Asians, modern/ancient East Asians; Tibetan, Mbuti) and found that Tibetan shared more derived

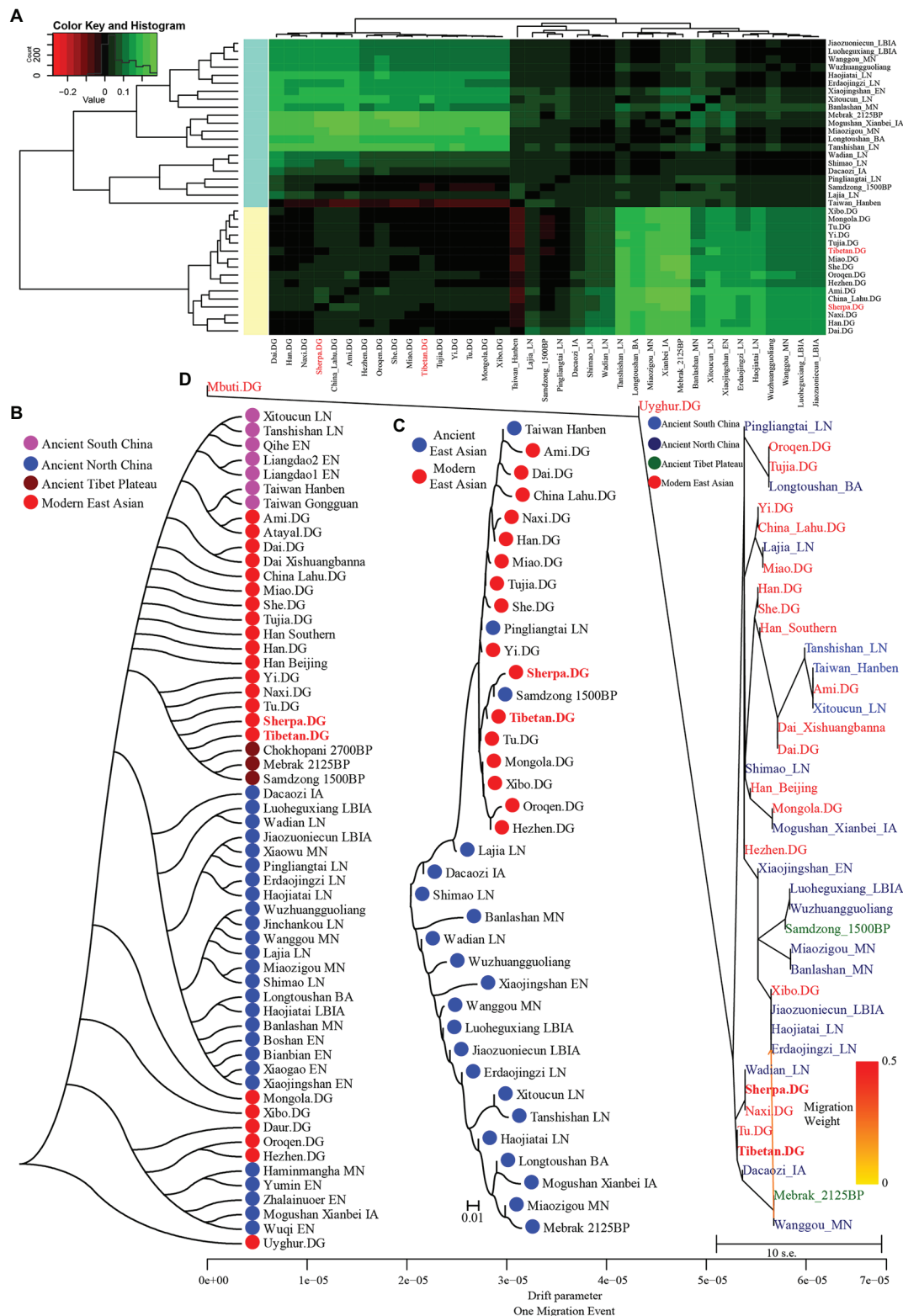
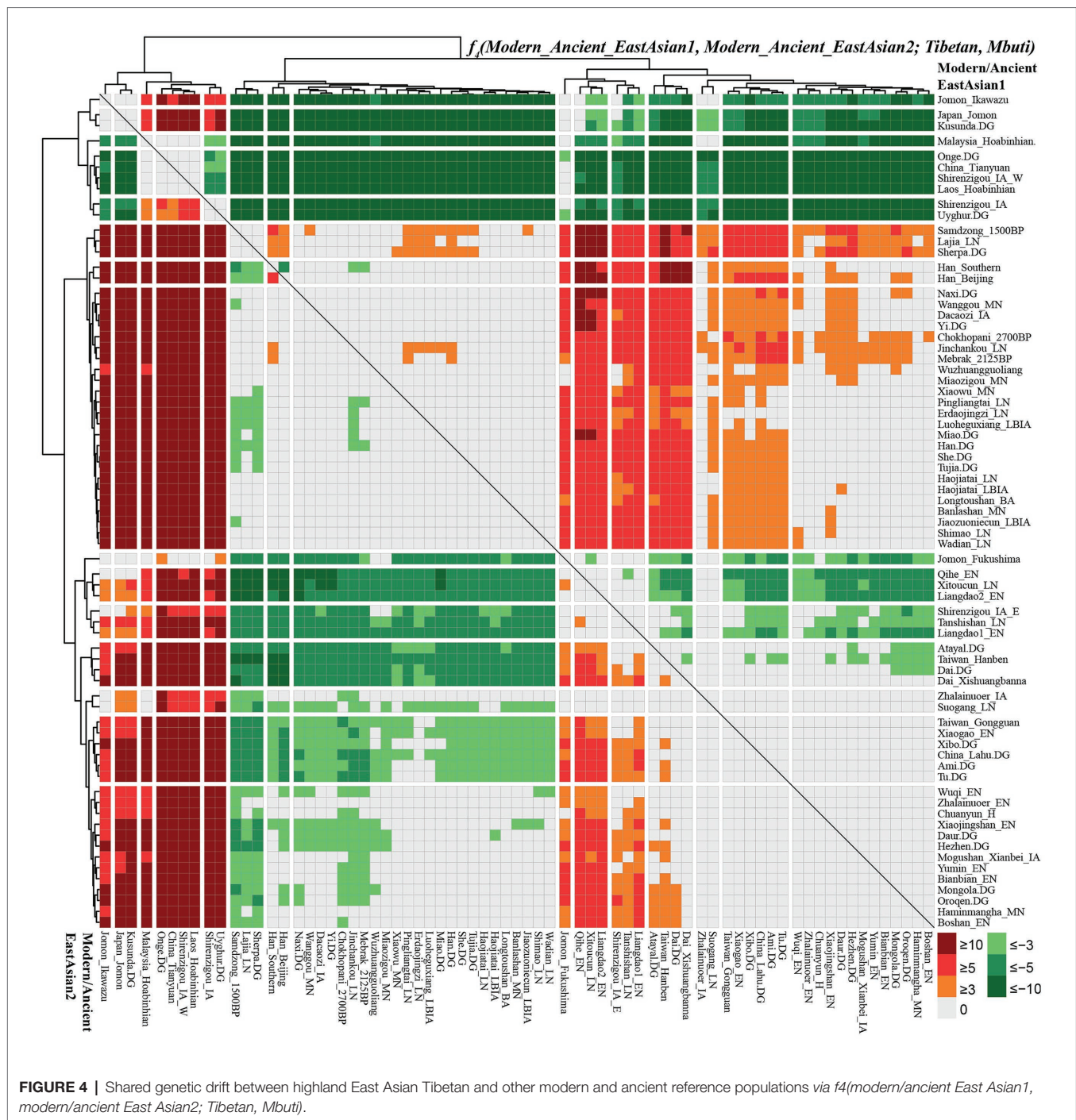


FIGURE 3 | Clustered patterns among East Asian Highlanders and other reference populations. **(A)** Heatmap of the pairwise F_{st} genetic distances among 42 included East Asian populations. **(B)** Neighbor-joining (NJ) tree was constructed using 1 – outgroup $f_3(\text{Source1 Source2}; \text{Mbuti})$. **(C)** The NJ tree was constructed via the F_{st} genetic distance matrixes. **(D)** The maximum likelihood tree showed the patterns of population splits and genetic admixture with one migration event.



alleles with Neolithic to present-day East Asians compared with deep East Asian lineages. Here, early East Asians were represented by Onge from South Asia, Hoabinhian people from Laos and Malaysia, and Tianyuan from Beijing. Compared with early East Asians, Tibetan shared more alleles with modern and ancient East Asians with negative f_4 values in the $f_4(\text{Early East Asians, modern/ancient Chinese populations; Tibetan, Mbuti})$. Some cases with the more shared genetic drifts between Tibetan and Jomon people were identified when we used Xinjiang Iron

Age to modern groups or 40,000-year-old Tianyuan people as the reference populations, such as $f_4(\text{Ikawazu Jomon, Shirenzigou_IA; Tibetan, Mbuti}) = 5.529 \times \text{SE}$. In summary, compared with northwestern Chinese populations with signatures of western Eurasian admixture and early East Asians, we found a strong genomic affinity between our studied Tibetan and northern/southern lowland East Asians. These observed genetic close relationships between Highlanders and East Asians showed that the gene pool of modern Tibeto-Burman speakers mainly

originated from East Asians, not from South Asia or Central Asia, although too many natural corridors and historic or prehistoric trade routes connected the TP and Central Asia or the Indian subcontinent (Jeong et al., 2016).

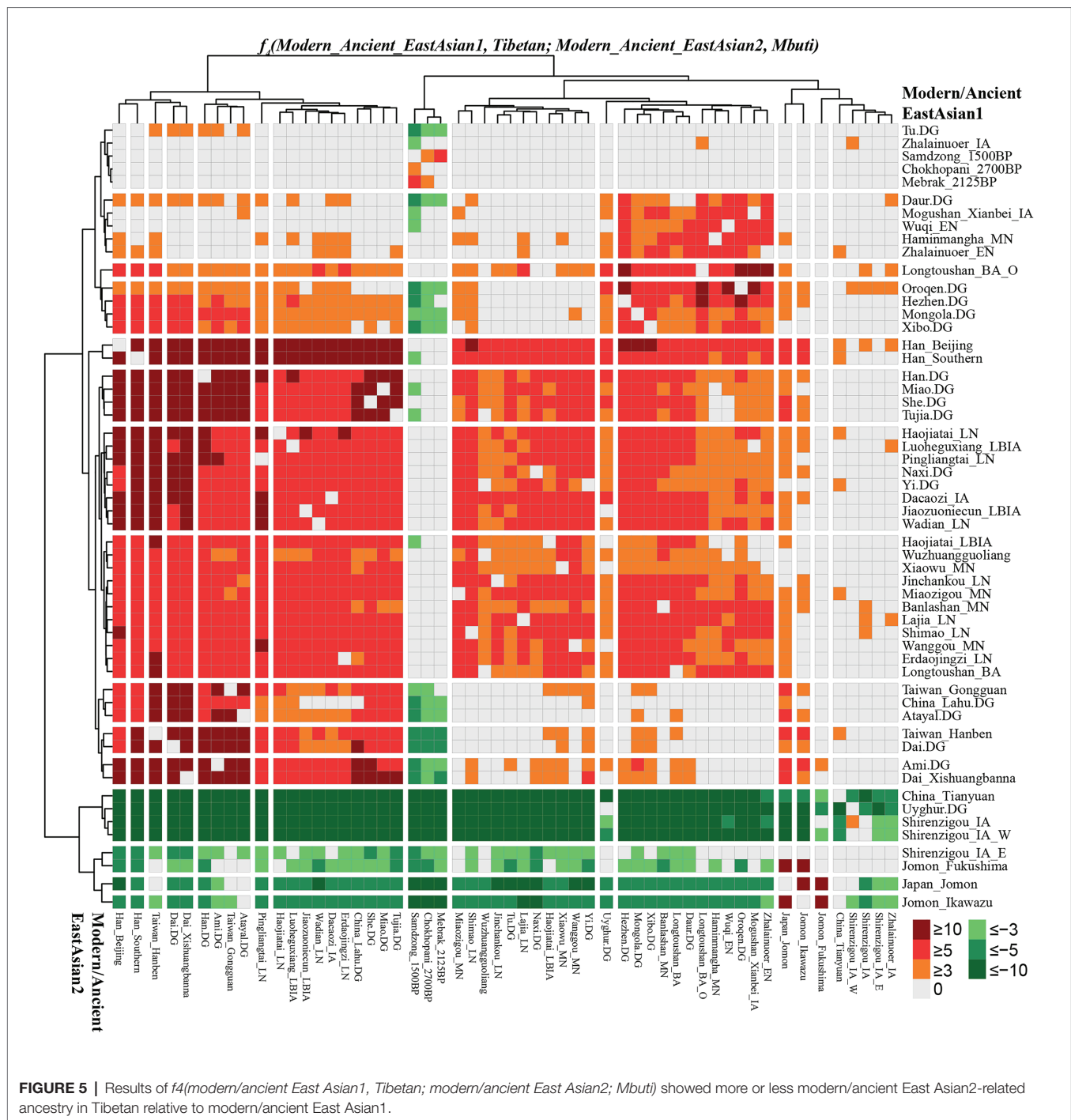
Focused on the population substructure within East Asians, the negative f_4 values in $f_4(\text{coastal Neolithic southern East Asians, modern/ancient northern East Asian; Tibetan, Mbuti})$ showed that Tibetan shared more alleles with northern East Asians. Positive values in $f_4(\text{Hanben/Atayal, coastal Neolithic southern East Asians; Tibetan, Mbuti})$ suggested Tibetan shared more alleles with Iron Age Hanben Taiwanese and their descendants than with their ancestors (Early Neolithic people), and positive f_4 values in $f_4(\text{inland modern southern East Asian Dai, coastal Neolithic to present-day southern East Asians; Tibetan, Mbuti})$ denoted Tibetan shared more alleles with inland southern East Asians than with island/coastal southern East Asians. Consistent positive f_4 values in $f_4(\text{Samdzong_1500BP/Lajia_LN/Sherpa, Lowland East Asians; Tibetan, Mbuti})$ showed that Tibetan had a strong genetic affinity with 1,500-year-old Samdzong people and Qijia people from Lajia, as well as the modern Sherpa. This obvious genetic affinity between modern Tibetans and ancients from Nepal and Qinghai showed the direct genetic contribution between Qijia culture-associated ancestral population and modern Tibetan or Nepal high-altitude people and modern Highlanders. Affinity between Tibetan and modern Tibeto-Burman speakers and other northern East Asians was further confirmed via positive f_4 values in $f_4(\text{northern East Asians, southern East Asians/Mongolic/Tungusic speakers; Tibetan, Mbuti})$. Focused on the Sherpa, as shown in **Supplementary Figure S8** and **Supplementary Table S12**, all green color denoted the significant negative f_4 values in $f_4(\text{modern/ancient East Asian1, modern/ancient East Asian2; Sherpa, Mbuti})$, which suggested Sherpa shared more derived alleles with lowland and highland northern East Asians compared with the early Asians, northwestern Chinese populations with western Eurasian admixture, ancients from coastal southeast China, islanders of Taiwan and Japanese Archipelago, and even some southern Chinese indigenous populations of Atayal and Dai. Red colors were observed when we used the following groups as the Modern/Ancient East Asian1: middle Neolithic populations (Miaozigou_MN, Wanggou_MN, Banlashan_MN, Wuzhuangguoliang), late Neolithic people (Wadian_LN, Haojiatai_LN, Shimao_LN), Qijia people (Jinchankou_LN, Lajia_LN, Dacaozi_IA), ancient Tibetans (Chokhopani_2700BP, Mebrak_2125BP, Samdzong_1500BP), and modern Sino-Tibetan (Naxi, Yi, Tibetan, Han, Han_Southern, Han_Beijing), which showed that Sherpa shared more alleles with them compared with southern East Asians or early Neolithic northern East Asians.

To further explore the genetic continuity and admixture of the Highlanders of Tibetan and Sherpa, we performed affinity f_4 statistics in the form of $f_4(\text{modern/ancient East Asian1, Tibetan/Sherpa; modern/ancient East Asian2, Mbuti})$. As shown in **Figure 5** and **Supplementary Table S13**, population lists of modern/ancient East Asian1 were presented in the right part and the other one was listed in the bottom part. Green colors showed the negative f_4 values, which suggested that Tibetan

harbored more ancestry derived from the groups related to the modern/ancient East Asian2. Here, we found that Tibetan possessed more ancestry from both northern and southern modern/ancient East Asians compared with northwestern Chinese populations and early Asians (Tianyuan, Hoabinhian, and Jomon). Most negative f_4 values were also observed in $f_4(\text{Qihе_EN/Liangdao2_EN, Tibetan; northern East Asians, Mbuti})$, suggesting that Tibetans harbored more northern East Asian ancestry. Red colors showed strong genetic affinity among lowland East Asians or more shared ancestry among them relative to Tibetan. We expected to observe the significant negative f_4 values if the included modern/ancient East Asian2 was the direct ancestor of Tibetan. Interestingly, $f_4(\text{modern/ancient East Asian1, Tibetan; Samdzong_1500BP/Sherpa/Mebrak_2125BP/Chokhopani_2700B, Mbuti})$ showed negative f_4 statistical values. However, no similar signals were identified in the late Neolithic Lajia or Jinchankou populations. Furthermore, no significant f_4 values should be observed when Nepal ancients were the unique ancestral source, or negative f_4 values could be obtained if there were some additional admixture gene flow into modern Tibetan in $f_4(\text{Samdzong_1500BP/Sherpa/Mebrak_2125BP/Chokhopani_2700B, Tibetan; modern/ancient East Asian2, Mbuti})$. No statistically significant f_4 values were observed here suggesting that ancestral populations related to the Nepal ancients were the direct ancestors of modern Tibetans. Although differentiated shared alleles were observed between Highlanders of Tibetan and Sherpa illustrated via $f_4(\text{Tibetan, Sherpa; Tu/Atayal/Niaozigou_MN/Wadian_LN/Erdaojingzi_LN, Mbuti})$, similar patterns of shared genetic drift were identified in Sherpa populations (**Supplementary Figure S9** and **Supplementary Table S14**).

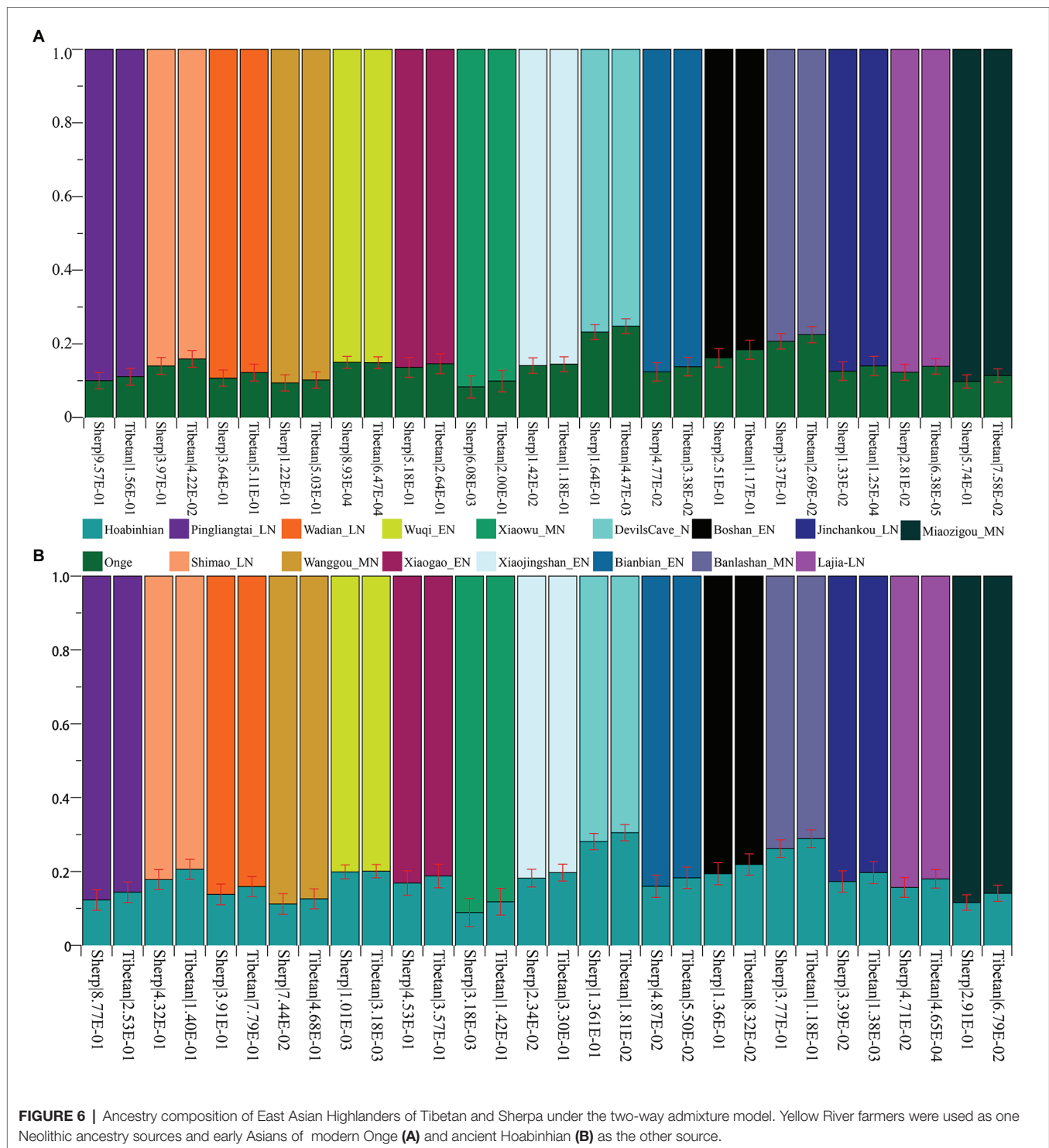
Genetic Admixture History Reconstruction of Highlanders of Tibetan and Sherpa via *qpWave/qpAdm* and *qpGraph*

Subsequently, to explore the plausible models of admixture fitted well of Highlanders and estimate the corresponding ancestry proportion, we used the statistical tool of *qpWave* to explore the minimum number of the possible ancestral populations and *qpAdm* to qualify ancestry proportion. Eight populations (Mbuti, Russia_Ust_Ishim, Russia_Kostenki14, Papuan, Australian, Mixe, Russia_MA1_HG, Mongolia_N_East) were used as the base outgroup set. Our *qpWave* results of $p_{\text{rank0}} < 0.05$ showed at least two ancestral populations could be used to model the ancestry composition of our included Tibetan and Sherpa. We first used the Ancient Ancestral South Indian of Onge as the southern source population represented as the deep diverged eastern Eurasian ancestry, which recently was hypothesized as the representation of indigenous South Asians in the study of the formation of human populations in South Asia (Narasimhan et al., 2019). Fifteen Neolithic northern East Asians from the Yellow River basin, West Liao River basin, Amur River basin, and other northern China and Russia were used as the other northern ancestral source. As shown in **Figure 6** and **Supplementary Table S15**, six models could be used to fit the observed genetic variations in both Sherpa and Tibetan with a large proportion of northern East Asian ancestry and small highly diverged eastern Eurasian ancestry: included two



coastal early Neolithic northern East Asian models (Boshan_EN and Xiaogao_EN), two inland middle Neolithic northern East Asian models (Miaozigou_MN and Wanggou_MN), and two inland late Neolithic northern East Asian models (Wadian_LN and Pingliangtai_LN). Besides, we also found Sherpa could be modeled as the admixture of 0.860 ± 0.023 Shimao-related ancestry and 0.140 Andamanese hunter-gatherer-related ancestry, 0.768 ± 0.02 DevilsCave_N-related ancestry and 0.232 Onge-related ancestry, and 0.793 ± 0.021 Banlashan Hongshan people-related ancestry

and 0.207 Onge-related ancestry. Similarly, Tibetan could be modeled as the mixing of 0.855 ± 0.020 Xiaojingshan_EN-related ancestry and 0.145 Onge-related ancestry, or 0.901 Xiaowu_MN-related ancestry and 0.099 Onge-related ancestry. Sherpa people could be modeled as approximately 0.870 of their ancestry derived from Qijia people associated with Lajia and Jinchankou populations with marginal nonsignificant p values ($1.33\text{E-}02$ and $2.81\text{E-}02$). When we substituted Hoabinhian from Laos with Onge as the southern ancestral source, the aforementioned six models (two



early Neolithic sources of Xiaogao and Boshan, two middle Neolithic sources of Wanggou and Miaozigou, and two late Neolithic sources of Pingliangtai and Wadian) could be fitted well of two included Highlanders with relative higher ancestry proportion from Hoabinhian-related ancestry. Another three models (Bianbian-EN-Hoabinhian: 0.840 for Sherpa and 0.817 for Tibetan; Banlashan_MN-Hoabinhian: 0.738 for Sherpa and

0.711 for Tibetan; and Shimao_LN-Hoabinhian: 0.877 for Sherpa and 0.856 for Tibetan). Middle Neolithic Xiaowu (0.882) and Xiaojingshan_EN (0.818) could be used as the northern East Asian sources for the model of the formation of modern Tibetan, and DevilsCave_N could be used as the source for modeling modern Sherpa with 0.719 derived from northern sources.

Finally, to reconstruct a deep population admixture history of the Highlanders of Tibetan and Sherpa based on the 1,233,013 SNPs, we used the basic phylogenetic framework from Wang et al. with the terminal modern populations of Mbuti, Onge, archaic population of Denisovan, and Paleolithic to Iron Age populations of Loschbour, Tianyuan, Liangdao2_EN, Lajia_LN, Chokhopani, and two eastern Mongolia Neolithic people (Wang et al., 2020). After adding Tibetan and Sherpa populations from the Simons Genome Diversity Project, we found Tibetan could be modeled as mixing from three source populations (**Figure 7**): coastal early Neolithic northern East Asian Bianbian_EN-related (Houli people: 0.040), inland late Neolithic northern East Asian Lajia_LN-related (Qijia people: 0.787), and deeply diverged East Eurasian-related (first layer of indigenous people, 0.173). For Sherpa, we used the middle Neolithic Yangshao people (Xiaowu_MN-related) as one of the northern East Asian sources, which could be modeled as the admixture of 0.73 ancestry directly derived from the northern main ancestral lineage and obtained additional 0.27 ancestry from southern East Asian lineage. In this situation, Sherpa was modeled as 0.09 ancestry from a group related to the middle Neolithic Yangshao people, 0.7644 from the ancestral population related to Lajia_LN, and 0.1456 from deeply diverged eastern Eurasian.

DISCUSSION

The Qinghai-Tibet Plateau and the surrounding great mountain ranges are home to cultural, genetic, and linguistic diversity since prehistoric or historic times, although nature environments, such as high-altitude hypoxia, resource scarcity, cold stress, and rough terrain, to some extent hindered the process, scale, and speed of the population's settlement in this world's high plateau. Archeological documents from Xiahe Denisovan mandible in northeastern TP (3,280 m above sea level) and abundant blade tool assemblage in the Nwya Devu site (4,600 m above sea level) successively demonstrated that humans colonized this high-altitude area from late Middle Pleistocene (160,000 years ago) to late Paleolithic stage (40,000–30,000 years ago; Zhang et al., 2018; Chen et al., 2019a). Genetic evidence for the high-altitude adaptative Denisovan-derived EPAS1 haplotype observed in modern Tibetan further showed a partial genetic continuity or archaic introgression between Denisovan and modern East Asian Highlanders. However, the demographic history and fine-scale genetic structure of modern and ancient Highlanders kept unclear and needed to be comprehensively explored. In the present study, we first used forensic short tandem repeat markers with high polymorphic and informative features to explore the genetic relationships between highland Tibetan and worldwide reference populations based on the allele frequency spectrum and found that East Asian Highlanders had a close genetic relationship with modern Tibeto-Burman-speaking populations and northern Han Chinese. This observed pattern of population relationship based on low-density genetic markers was consistent with recent linguistic evidence for the North China origin of modern Sino-Tibetan language (Sagart et al., 2019; Zhang et al., 2019).

To further clarify the population relationship and potential gene flow events, we subsequently used one high-density dataset comprised of the 1240K SNP genetic markers focused on the Highlanders of Tibetan and Sherpa and compared them with all available Chinese ancient and modern reference populations (Patterson et al., 2012; Yang et al., 2017, 2020; Lipson et al., 2018; Jeong et al., 2019; Liu et al., 2020; Ning et al., 2020; Wang et al., 2020) to carry out another comprehensive population genetic relationship analysis. Ancestry composition *via* the ADMIXTURE model-based cluster result showed a genetic affinity between Tibetan and Sherpa and their close genetic relationship with eight Nepal ancient individuals from a cultural background associated with Chokhopani, Mebrak, and Samdzong. This observed genetic similarity and continuity based on the 1240K dataset were consistent with Jeong's original finding of long-term genetic stability (Jeong et al., 2016). Genetic affinity and continuity among ancient Nepal populations and modern Tibetan and Sherpa were further evidenced *via* the more shared genetic drift in *f*-statistics and close phylogenetic relationships in the NJ tree and *qpGraph*-based phylogeny framework. Besides, we also identified a close genetic relationship between modern Sherpa/Tibetan and ancient Qijia people from the upper Yellow River basin (Lajia and Jinchankou), suggesting Qijia people as the representative of Neolithic millet farmers played an important role in the formation of modern Tibetans although they shared more alleles with Neolithic Yangshao, Longshan people from Central Plain in Henan Province, and Houli people from Shandong Province. Our autosome-based genetic links between ancient populations from northeast TP were consistent with recent archeological, Y-chromosomal, and mitochondrial evidence for the colonization and peopling of the Qinghai-Tibet Plateau (Chen et al., 2015; Wang et al., 2018a; Zhang et al., 2018; Li et al., 2019b; Ding et al., 2020). Archeologically attested charred grains and the corresponding carbonization dating data provided by Chen et al. suggested that a novel agropastoral economy facilitated Neolithic millet farmers to enjoy year-round living and to successfully occupy the northeastern TP around 3,600 years ago (Chen et al., 2015). Mitochondrial DNA (mtDNA) variations of modern Tibetan also provided clues that the upper Yellow River millet farmers first adopted cold-tolerant barley agriculture and then permanently inhibited it in the TP (Li et al., 2019b). Ancient mitogenomes of 5,200- to 300-year-old humans from Tibet, Gansu, Qinghai and Sichuan provinces also revealed that the D4j1b-represented ancestral population expanded from the low-altitude area to the core region of the TP around 4,750 to 2,775 years ago (Ding et al., 2020). Uniparental genetic evidence from Y-chromosome phylogeny also showed that the Yellow River farmers with the paternal founding lineage of Oa1c1b-CTS5308 dispersed to the TP had triggered the formation and expansion of modern high-altitude Tibeto-Burman speakers (Wang et al., 2018a). Thus, our findings combined with evidence from the aforementioned archeological or uniparental contents consistently supported that the northeastern edge of the TP is an important geographical corridor for ancient human movements and admixtures between low altitude and high altitude.

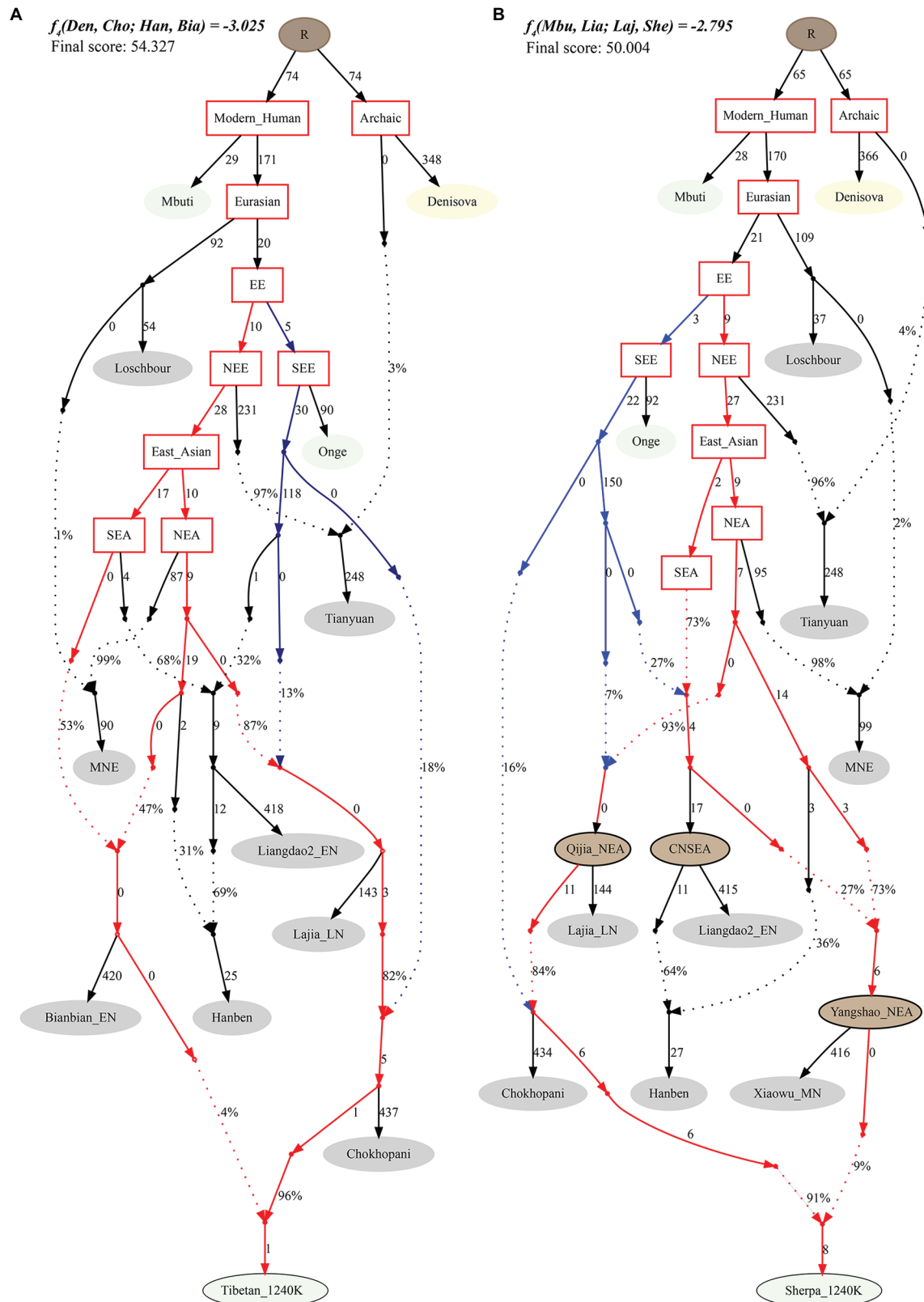


FIGURE 7 | Phylogenetic framework of Tibetan (A) and Sherpa (B) based on the 1240K high-density datasets.

Both southwestern agricultural charred cereal grains (barley and wheat) and northern China Yellow River dryland millet charred cereal grains (foxtail millet and broomcorn millet) were identified in the Neolithic archeological sites in the northeastern TP (Chen et al., 2015), suggesting that the communication of the adaptation of agriculture techniques existed there. A close genetic connection combined with these archeological records evidenced that the northeastern TP is the main geographical corridors of the peopling of TP. However, whether this mixed agriculture system was caused by human population movements and admixture or only acculturation of skills is unclear. We performed a series of population genetic analyses to clarify the admixture sources and progress. First, f_3 and f_4 -statistics did not identify more shared genetic drift with western Eurasian populations. Second, the observed genetic variations observed in Highlanders of Sherpa and Tibetan could be competently explained *via* a two-way admixture model with one deeply diverged Onge/Hoabinhian-related eastern Eurasian lineage and one northern East Asian lineage. Third, our *qpGraph*-based admixture graph model fitted well without a gene flow from western Eurasian populations. Thus, our genetic phylogenetic evidence supported that the upper Yellow River millet farmers adopted western barley and wheat agriculture techniques *via* an adaptation of the idea and not the direct movement of people. The cultural diffusion model was recently also evidenced *via* mitochondrial haplotype and haplogroup data (Li et al., 2019b). Li et al. recently discorded the founder maternal lineages (M9a1a1c1b1a and A11a1a) of Neolithic millet farmers based on the combined analyses of radiocarbon dating of cereal remains and mtDNA-based haplogroup geographical distribution among 8,277 Tibetans and 58,514 individuals from surrounding populations. Their founding supported that Yellow River millet farmers adopting barley agriculture successfully colonized the East Asian high-altitude region. In summary, our admixture- f_3 results, symmetrical- f_4 analyses, and *qpGraph*-based phylogeny did not identify obvious western Eurasian-related gene flow events in Qijia people and modern Highland East Asians, which suggested that cultural communication did not involve large-scale population movements and admixtures from Central Asia or western Eurasia.

Different from subpopulation structures observed in Highland East Asians (Zhang et al., 2017), our present study identified a genetic similarity between Sherpa and Tibetan, which may be caused by the small sample size and low density of genetic sampling. Thus, denser sampling of geographically/ethnically/linguistically diverse modern highland East Asians and ancient populations should be done to clarify the population substructure and demographic history of modern and ancient highland/lowland East Asians. Regardless of the fact that the limitations of sample size and population numbers existed here, our ancestry composition estimation and phylogeny reconstruction revealed multiple stages of genetic admixtures of both Tibetan and Sherpa. Paleolithic ancestry was estimated to over 10% when we used the South Asian Onge (shared deeply diverged haplogroup D) and early Asians of Laos Hoabinhian as the deep ancestral source. This deeply diverged eastern Eurasian

identified in modern East Asian Highlanders and 2,750-year-old Nepal ancient was consistent with Paleolithic sublineages of haplogroup D-M174 (D1a1-M15 and D1a2-p99), which was the representative lineage or genetic legacy of Paleolithic TP local residing hunter-gatherers (Wang et al., 2018a). This finding combined with Paleolithic archeological documents (Zhang et al., 2018), genetically attested Denisovan EPAS1 haplotype (Huerta-Sanchez et al., 2014), and Paleolithic paternal/material founding lineages (Qi et al., 2013) supported that both Paleolithic and Neolithic genetic legacies co-existed in Iron Age to modern highland East Asians.

CONCLUSION

Our population genetic or genomic analyses showed that both high-density and low-density datasets in the present study revealed the close genetic relationship between Highlanders and lowland Tibeto-Burman-speaking populations, forensic-related STR-based analysis showed limitations for finer-scale genetic structure dissection due to its relatively lower resolution with the forensically developed systems. We used STR-based datasets to evaluate the genetic diversity and forensic characteristics as well as to uncover the genetic similarities and differentiation between the studied Tibetan group and 56 reference populations and found that the STR amplification system was informative and discriminative in Nagqu Tibetan and could be applied in the construction of the Chinese national STR datasets. Comprehensive worldwide or nationwide population comparisons demonstrated that Nagqu Tibetan keeps the genetic affinity with ethnically close Chengdu Tibetan, Liangshan Tibetan, and Tibet Tibetan. Furthermore, population structure and demographic history reconstruction based on the high-density 1240K dataset showed that Highlanders of Tibetan and Sherpa possessed a close genetic relationship with Qijia culture-related people (Lajia and Jinchankou), suggesting that the northeastern edge of the Tibetan Plateau is an important geographical corridor for population movements and admixtures in the progress of permanent human settlement of the TP. No western Eurasian admixture signatures were identified in modern and ancient populations of the core region and northeastern edge of the TP, suggesting that the late Neolithic upper Yellow River millet farmers' adoption of barley and wheat agriculture from the Fertile Crescent of southwestern Asia was mediated *via* the cultural diffusion model and not *via* the demic diffusion model. Finally, the observed shared deeply diverged Onge/Hoabinhian-related eastern Eurasian lineage into modern Tibetan, Sherpa, and 2,700-year-old Chokhopani demonstrated that a common Paleolithic genetic legacy widely existed in all highland East Asians.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Zunyi Medical University. The patients/participants provided their written informed consent to participate in this study. Informed consent was obtained from all participants included in the study.

AUTHOR CONTRIBUTIONS

GH, MW, and XZ conceived the idea for the study. GH, PC, XZ, YL, and MW performed or supervised wet laboratory work. GH, MW, XZ, PC, ZW, JL, LY, FW, YL, and RT analyzed the data. GH, MW, YL, and XZ wrote and edited the manuscript.

REFERENCES

- Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109
- 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, 1696–1704. doi: 10.1038/s41588-018-0250-5
- 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. doi: 10.1186/s13742-015-0047-8
- Chen, F. H., Dong, G. H., Zhang, D. J., Liu, X. Y., Jia, X., An, C. B., et al. (2015). Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 B.P. *Science* 347, 248–250. doi: 10.1126/science.1259172
- Chen, F., Welker, F., Shen, C. C., Bailey, S. E., Bergmann, I., Davis, S., et al. (2019a). A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature* 569, 409–412. doi: 10.1038/s41586-019-1139-x
- Chen, P., Wu, J., Luo, L., Gao, H., Wang, M., Zou, X., et al. (2019b). Population genetic analysis of modern and ancient DNA variations yields new insights into the formation, genetic structure, and phylogenetic relationship of northern Han Chinese. *Front. Genet.* 10:1045. doi: 10.3389/fgene.2019.01045
- Chiang, C. W. K., Mangul, S., Robles, C., and Sankararaman, S. (2018). A comprehensive map of genetic variation in the World's largest ethnic group—Han Chinese. *Mol. Biol. Evol.* 35, 2736–2750. doi: 10.1093/molbev/msy170
- Ding, M., Wang, T., Ko, A. M., Chen, H., Wang, H., Dong, G., et al. (2020). Ancient mitogenomes show plateau populations from last 5200 years partially contributed to present-day Tibetans. *Proc. Biol. Sci.* 287:20192968. doi: 10.1098/rspb.2019.2968
- 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. (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
- Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., et al. (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature* 524, 216–219. doi: 10.1038/nature14558
- Gao, J. Y., Hou, G. L., Wei, H. C., Chen, Y. C., E, C. Y., Chen, X. L., et al. (2020a). Prehistoric human activity and its environmental background in Lake Donggi Cona basin, northeastern Tibetan Plateau. *The Holocene* 30, 657–671. doi: 10.1177/09596836198955
- Gao, Y., Zhang, C., Yuan, L., Ling, Y., Wang, X., Liu, C., et al. (2020b). PG. Han: the Han Chinese genome database and analysis platform. *Nucleic Acids Res.* 48, D971–D976. doi: 10.1093/nar/gkz829
- 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
- 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, 207–211. doi: 10.1038/nature14317
- 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., Wang, Z., Su, Y., Zou, X., Wang, M., Liu, J., et al. (2018a). Genetic variation and forensic characterization of highland Tibetan ethnicity revealed by autosomal STR markers. *Int. J. Legal Med.* 132, 1097–1102. doi: 10.1007/s00414-017-1765-5
- He, G., Wang, Z., Zou, X., Chen, X., Liu, J., Wang, M., et al. (2018b). Genetic diversity and phylogenetic characteristics of Chinese Tibetan and Yi minority ethnic groups revealed by non-CODIS STR markers. *Sci. Rep.* 8:5895. doi: 10.1038/s41598-018-24291-5
- Huerta-Sanchez, E., Jin, X., Asan, B., Bianba, Z., Peter, B. M., Vinckenbosch, N., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194–197. doi: 10.1038/nature13408
- Jeong, C., Balanovsky, O., Lukianova, E., Kahbatkyy, 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
- Jeong, C., Ozga, A. T., Witonsky, D. B., Malmstrom, H., Edlund, H., Hofman, C. A., et al. (2016). Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. *Proc. Natl. Acad. Sci. U. S. A.* 113, 7485–7490. doi: 10.1073/pnas.1520844113
- Kovach, W. L. (2013). MVSP: a multivariate statistical package for windows. version 3.22. Pentraeth: Kovach Computing Services.
- Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054
- Li, Y. -C., Tian, J. -Y., Liu, F. -W., Yang, B. -Y., Gu, K. -S. -Y., Rahman, Z. U., et al. (2019b). Neolithic millet farmers contributed to the permanent settlement of the Tibetan Plateau by adopting barley agriculture. *Natl. Sci. Rev.* 6, 1005–1013. doi: 10.1093/nsr/nwz080
- Li, L., Ye, Y., Song, F., Wang, Z., and Hou, Y. (2019a). 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
- Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietruszewsky, M., et al. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* 361, 92–95. doi: 10.1126/science.aat3188
- Liu, D., Duong, N. T., Ton, N. D., Van Phong, N., Pakendorf, B., Van Hai, N., et al. (2020). Extensive ethnolinguistic diversity in Vietnam reflects multiple sources of genetic diversity. *Mol. Biol. Evol.* 37, 2503–2519. doi: 10.1093/molbev/msaa099
- 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, 580–594. doi: 10.1016/j.ajhg.2016.07.002

All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by a grant from the Fundamental Research Funds for the Central Universities (YJ201651).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.582357/full#supplementary-material>

- Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., et al. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206. doi: 10.1038/nature18964
- Meyer, M. C., Aldenderfer, M. S., Wang, Z., Hoffmann, D. L., Dahl, J. A., Degering, D., et al. (2017). Permanent human occupation of the central Tibetan plateau in the early Holocene. *Science* 355, 64–67. doi: 10.1126/science.aag0357
- Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., et al. (2019). The formation of human populations in South and Central Asia. *Science* 365:eaat7487. doi: 10.1126/science.aat7487
- Nicogossian, A., Kloiber, O., and Stabile, B. (2014). The revised world medical Association's declaration of Helsinki 2013: enhancing the protection of human research subjects and empowering ethics review committees. *World Med. Health Policy* 6, 1–3. doi: 10.1002/wmh3.79
- 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
- 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, 2526.e2524–2532.e2524. doi: 10.1016/j.cub.2019.06.044
- 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
- Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8:e1002967. doi: 10.1371/journal.pgen.1002967
- Qi, X., Cui, C., Peng, Y., Zhang, X., Yang, Z., Zhong, H., et al. (2013). Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the tibetan plateau. *Mol. Biol. Evol.* 30, 1761–1778. doi: 10.1093/molbev/mst093
- Qin, P., Li, Z., Jin, W., Lu, D., Lou, H., Shen, J., et al. (2014). A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese. *Eur. J. Hum. Genet.* 22, 248–253. doi: 10.1038/ejhg.2013.111
- Qin, Z., Yang, Y., Kang, L., Yan, S., Cho, K., Cai, X., et al. (2010). A mitochondrial revelation of early human migrations to the Tibetan Plateau before and after the last glacial maximum. *Am. J. Phys. Anthropol.* 143, 555–569. doi: 10.1002/ajpa.21350
- Ren, L., Dong, G., Liu, F., D'alpoim-Guedes, J., Flad, R. K., Ma, M., et al. (2020). Foraging and farming: archaeobotanical and zooarchaeological evidence for Neolithic exchange on the Tibetan Plateau. *Antiquity* 94, 1–16. doi: 10.15184/aqy.2020.35
- Retief, J. D. (2000). Phylogenetic analysis using PHYLIP. *Mol. Biol.* 132, 243–258. doi: 10.1385/1-59259-192-2:243
- 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
- Wang, Z., He, G., Luo, T., Zhao, X., Liu, J., Wang, M., et al. (2018b). Massively parallel sequencing of 165 ancestry informative SNPs in two Chinese Tibetan-Burmese minority ethnicities. *Forensic Sci. Int. Genet.* 34, 141–147. doi: 10.1016/j.fsigen.2018.02.009
- Wang, L. X., Lu, Y., Zhang, C., Wei, L. H., Yan, S., Huang, Y. Z., et al. (2018a). Reconstruction of Y-chromosome phylogeny reveals two neolithic expansions of Tibeto-Burman populations. *Mol. Gen. Genomics.* 293, 1293–1300. doi: 10.1007/s00438-018-1461-2
- Wang, C. -C., Yeh, H. -Y., Popov, A. N., Zhang, H. -Q., Matsumura, H., Sirak, K., et al. (2020). The genomic formation of human populations in East Asia. *bioRxiv*. doi: 10.1101/2020.03.25.004606 [Preprint].
- 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
- Yang, M. A., Gao, X., Theunert, C., Tong, H., Aximu-Petri, A., Nickel, B., et al. (2017). 40,000-year-old individual from Asia provides insight into early population structure in Eurasia. *Curr. Biol.* 27, 3202.e3209–3208.e3209. doi: 10.1016/j.cub.2017.09.030
- Yao, H. B., Tang, S., Yao, X., Yeh, H. Y., Zhang, W., Xie, Z., et al. (2017). The genetic admixture in Tibetan-Yi Corridor. *Am. J. Phys. Anthropol.* 164, 522–532. doi: 10.1002/ajpa.23291
- Zhang, X. L., Ha, B. B., Wang, S. J., Chen, Z. J., Ge, J. Y., Long, H., et al. (2018). The earliest human occupation of the high-altitude Tibetan Plateau 40 thousand to 30 thousand years ago. *Science* 362, 1049–1051. doi: 10.1126/science.aat8824
- Zhang, C., Lu, Y., Feng, Q., Wang, X., Lou, H., Liu, J., et al. (2017). Differentiated demographic histories and local adaptations between Sherpas and Tibetans. *Genome Biol.* 18:115. doi: 10.1186/s13059-017-1242-y
- 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
- Zhao, M., Kong, Q. P., Wang, H. W., Peng, M. S., Xie, X. D., Wang, W. Z., et al. (2009). Mitochondrial genome evidence reveals successful late Paleolithic settlement on the Tibetan Plateau. *Proc. Natl. Acad. Sci. U. S. A.* 106, 21230–21235. doi: 10.1073/pnas.0907844106
- Zou, X., Wang, Z., He, G., Wang, M., Su, Y., Liu, J., et al. (2018). Population genetic diversity and phylogenetic characteristics for high-altitude adaptive Kham Tibetan revealed by DNATyper(TM) 19 amplification system. *Front. Genet.* 9:630. doi: 10.3389/fgene.2018.00630

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

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

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: frontiersin.org/about/contact



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

@frontiersin



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership