INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Jorge
Ruiz Ramírez

PhD Thesis

Development and validation of large-scale SNP sets for forensic purposes

Santiago de Compostela, 2024

Doctoral Programme in Advances and New Strategies in Forensic Science

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

DOCTORAL THESIS

DEVELOPMENT AND VALIDATION OF LARGE-SCALE SNP SETS FOR FORENSIC PURPOSES

Jorge Ruiz Ramírez

Directors: Mª Victoria Lareu Huidobro and Christopher Paul Phillips

ESCUELA DE DOCTORADO INTERNACIONAL DE LA UNIVERSIDAD DE SANTIAGO DE COMPOSTELA
PROGRAMA DE DOCTORADO EN AVANCES Y NUEVAS ESTRATEGIAS EN CIENCIAS FORENSES

SANTIAGO DE COMPOSTELA
2024

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Este trabajo ha sido parcialmente financiado por el *Programa de axudas á etapa predoutoral* de la Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria de la Xunta de Galicia, España (ED481A-2020-039); así como los proyectos de investigación BIO2016–78525-R, PID2022-141224OB-I00, PID2019–107876RB-I00 de la Agencia Estatal de Investigación, 2013-2016 y el Programa de Investigación e Innovación Horizonte 2020 de la Unión Europea (acuerdo nº740580)

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

A mis padres

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AGRADECIMIENTOS

Tras más de un lustro desde que, ignorante de mí me mudé a esa misteriosa y desconocida región del noroeste peninsular llamada Galicia, muy a mi pesar, acabaré echando en falta estas tierras frías y húmedas. Después de finalizar la tortura que supone enfrentarse a escribir una tesis doctoral, únicamente me queda intentar expresar todo mi agradecimiento, en la medida de lo sea capaz de expresar, a todas las personas que me han acompañado todos estos años a lo largo de esta etapa.

En primer lugar, quería dar las gracias a mis directores de tesis, Maviki y Chris. Maviki, me diste la oportunidad de entrar al laboratorio e iniciarme en el mundo de la genética forense. Gracias por tu cercanía, apoyo y por mantener el laboratorio a flote. Cuando entré, recuerdo que me dijiste que a Galicia se llega llorando y se va llorando. Lo de llegar llorando quedó patente, y me temo que de lo segundo tampoco me libro. Chris, I cannot express all the gratitude that I have for all the knowledge that you willingly (and patiently) shared with me over these years, although I fear I may have picked some bad habits as well (like *prochristinating* with writing a thesis). But, beyond all the academic support, I wanted to thank you for all the care and dedication that you have given me and all the enthusiasm and joy that you transmit.

Siguiendo por orden jerárquico, a las tres integrantes del despacho postdoc, también conocido como cuarto de los listos. A las tres, agradeceros el gran esfuerzo que hacéis por mantener en funcionamiento el faro y el nivel de profesionalidad que demostráis día a día. Ana Freire Aradas, que fue la primera persona conocí al entrar al laboratorio y que, aunque a veces el muro invisible pueda parecer infranqueable, has demostrado ser una persona llena de empatía y bondad. A la señora Ana Mosquera Miguel, una de las personas más altruistas que he conocido, que siempre está dispuesta a ayudar con una sonrisa (hacia arriba) y de buena gana. A Maripú, que con tu capacidad para el *multitasking* has estado siempre para ayudar, aunque estuvieras con otras mil brozas al mismo tiempo. Gracias por implicarte tanto, aunque no te tocara, y que por mucho que te guste refunfuñar lo haces encantada porque te preocupas por tus niños. Grazas tamén á señora María Amelia, fonte de sabedoría do noso laboratorio. Grazas pola túa paciencia e por todas as oportunidades desaproveitadas de darnos a Adri e a min una boa colleja.

A los dos compañeros con los que inicié la tesis, Adrián y Lorena. Señor Adrián, aunque mi agradecimiento carezca de tu pomposidad y no haga uso de figuras retóricas, espero conseguir expresar mi gratitud hacia su persona por acompañarme estos años de principio a fin y por tu predisposición para ayudar en lo que haga falta sin siquiera preguntar. Gracias por tu compañía, durante la tesis, actividades extraescolares y viajes (con suerte algún día explicarás nuestro misterioso viaje a *coney island*). Sin ti la mesa de Amaia no sería la más limpia. Lorena, gracias por enseñarnos tanto y dejar que estos dos críos que te hicieran compañía en tus últimos años de doctorado.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Como no, gracias al resto de integrantes del cuartucho que, aunque sean incapaces de juzgar un libro por su portada, trajisteis la alegría y el caos a este nuestro laboratorio. Señora Amaia, eskerrik asko por tu ingenuidad (en el buen sentido) y por los buenos ratos desde el Caña Hueca hasta el desván. Nunca creí que algún día normalizaría tanto hablar de los amigos del cementerio. A Lucia, máxima autoridad de la Entidad Nacional de Cartelería y mi libanesa preferida. En ti recae a partir de ahora el deber del chisme del laboratorio. A los dos señores de la parabólica que jamás vi funcionar tan mal. Javi, terrorista informático digno sucesor de Amelia y Miguel (versión no cayetana), testigo silencioso del cuartucho.

Vorrei anche ringraziare tutte le ragazze italiane che hanno visitato il laboratorio in tutti questi anni: Noemi, Giorgia, Flavia, Francesca (Paca), l'Emma e Desucha, la mia coinquilina preferita. Grazie a voi non solo ho imparato la vostra lingua, ma avete condiviso con me la vostra cultura e, soprattutto, la vostra gastronomia. Poverine voi, perché presto tornerò a trovarvi e, probabilmente, occuperò le vostre case.

A mis señoras amigas de la carrera, Claudia, Elizabeth, Tamara, Marisa, Chema y Azucena. Mariclaudia, también conocida como la de los stickers, compañera de desgracias y penurias. Quien nos iba a decir que después de tantos años parece que vemos la luz al final de la tesis. Elizabeth, que no sé si a lo próximo que acudiré será a tu boda, la defensa de tu tesis o tu funeral. Tamara, mi fiel compañera de interrail que nos hiciste ubicar a todos el Bierzo con ese acento norteño (que al final me he acabado quedando yo). Chema, enguerina con inclinación por temas orgánicos y *community manager* de La Casa de Papel. Marisilla, galega adoptiva que algún día me prestará su muestra de ADN para estudios poblacionales. Gracias también a Sergio por llevarme de excursión por media Galicia. Y a mis amigos de Elche, Jose, Ángela, Mari, Daniel, María Isabel, y a Helena, artista emergente y futura diseñadora de la portada alternativa de esta tesis. Gracias también al señor Daniel, que ha tenido que sufrir las consecuencias de lo que es escribir una tesis doctoral, aunque yo también.

Gracias toda mi familia, que ha estado detrás de mi apoyándome en todo momento. A mi abuela, financiadora junto con la Xunta de Galicia de esta tesis (aunque cuando acabe el doctorado seguirá pensando que su nieto el doctor es el licenciado en medicina). A mi tía Fini. A los hipoglúcidos de mis hermanos Pablo y Javier, a Marta, a mis sobrinos Siena y Boston, al aquelarre de mis tías y a mis primos. Y, en último lugar, querría darle las gracias a mis señores padres, que poco a poco han visto a su hijo crecer a distancia, a la par que adquirir un acento foráneo y un vocabulario un tanto peculiar. Más allá de toda la ayuda que me habéis dado y del flete de *tuppers* congelados, gracias por vuestro apoyo incondicional aún sin entender qué es lo que estaba haciendo y haber estado ahí siempre, sin tener muy claro qué hacía, para qué servía hacer una tesis doctoral o aun habiendo perdido la pista del país en el que me hallaba. Esta tesis va dedicada a vosotros.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

One of the advantages of being disorganized is that one is always having surprise discoveries

– A.A. Milne (*Winnie-the-Pooh*)

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CONFLICTOS DE INTERÉS

Jorge Ruiz Ramírez, como autor de esta tesis doctoral, declara que no hay conflictos de interés asociados a este trabajo.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

INDEX

INDEX i

ABBREVIATIONS iii

RESUMO 1

SUMMARY 9

1. INTRODUCTION 19

1.1 THE ADVENT OF FORENSIC SCIENCES 19

1.2 THE DISCOVERY OF THE DNA AND THE FIRST STEPS IN FORENSIC GENETICS 21

1.2.1 The haemogenetics and protein markers 21

1.2.2 RFLPs and DNA fingerprinting 21

1.2.3 The PCR and steps towards modern-day forensics 22

1.2.4 Prospects of shift from length-based to sequence-based analyses 23

1.3 DNA MARKERS OF FORENSIC INTEREST 24

1.3.1 DNA polymorphisms 25

1.3.1.1 STRs 25

1.3.1.2 SNPs 26

1.3.1.3 Indels 27

1.3.1.4 Microhaplotypes 28

1.3.2 Lineage markers 29

1.3.2.1 Mitochondrial DNA 29

1.3.2.2 The Y chromosome 31

1.3.3 The X chromosome 32

1.4 OTHER GENETIC MARKERS OF FORENSIC INTEREST 33

1.4.1 Epigenetic markers 33

1.4.2 RNA 33

1.5 ANALYSIS METHODS USED FOR FORENSIC TESTING 34

1.5.1 STR profiling 35

1.5.2 DNA Sequencing 36

1.5.2.1 First generation sequencing 36

1.5.2.2 Next Generation Sequencing 37

1.5.2.2.1 MPS sequencing workflow *39*

1.5.2.3 Third-generation sequencing 40

1.5.3 Forensic DNA phenotyping 41

1.5.4 Bio-geographic ancestry 43

1.5.4.1 The structure of human population 43

1.5.4.2 Inference of biogeographic ancestry using genetic markers 46

1.5.5 Chronological age prediction 47

1.5.6 Investigative Genetic Genealogy 48

2. OBJECTIVES 51

3. METHODOLOGY 55

3.1 RETRIEVING GENOTYPES FROM WGS PROJECTS 55

3.2 SELECTING MARKERS FOR MPS PANEL CONSTRUCTION 55

3.2.1 Selecting markers for ancestry inference 56

3.2.2 Selecting markers for Human Identification 57

3.2.3 Other considerations for marker selection 57

3.3 DNA TYPING 57

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.3.1 Sample selection **57**
3.3.2 Library preparation and genotyping **58**
3.3.3 Bioinformatic analysis **58**
3.4 POPULATION DATA ANALYSIS **58**
3.5 PEDIGREE ANALYSES **60**
3.6 FORENSIC VALIDATION OF AN MPS PANEL **60**

4. RESULTS **65**
Article 1. Development and Evaluation of the Ancestry Informative Marker Panel of the VISAGE Basic Tool 65
Article 2. Development and evaluations of the ancestry informative markers of the VISAGE Enhanced Tool for Appearance and Ancestry 86
Article 3. Broadening the Applicability of a Custom Multi-Platform Panel of Microhaplotypes: Bio-Geographical Ancestry Inference and Expanded Reference Data 111
Article 4. *Eurasiaplex*-2: Shifting the focus to SNPs with high population specificity increases the power of forensic ancestry marker sets 125
Article 5. A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel 138
Article 6. Inter-platform evaluation of the MPSplex large-scale tri-allelic SNP panel for forensic identification 152

5. DISCUSSION **173**
5.1 ASSESSING PANELS FOR BIOGEOGRAPHIC ANCESTRY INFERENCE **173**
5.1.1 Development of new panels for ancestry inference **174**
5.1.2 Exploring the applicability of an already existing panel for BGA prediction **178**
5.1.3 Towards the creation of a population-specific maker panel **181**
5.2 A PANEL FOR KINSHIP ANALYSIS: MPSPLEX **184**
5.2.1 Marker selection and panel development **184**
5.2.2 Assessment and validation of the MPSplex panel **186**

6. CONCLUSIONS **191**

7. BIBLIOGRAPHY **197**

ANNEXES **219**
ANNEX I. SUPPLEMENTARY TABLES FOR ARTICLE 6 ..... **219**
ANNEX II. SUPPLEMENTARY FIGURES FOR ARTICLE 6 ..... **227**
ANNEX III. SUPPLEMENTARY FILE 1 FOR ARTICLE 6 ..... **234**

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ABBREVIATIONS

AFR: sub-Saharan Africa
AIM: ancestry informative marker
aiSNP: ancestry informative SNPs
AME: America
BGA: bio-geographical ancestry
bp: base pair
CE: capillary electrophoresis
CEPH: Centre Etude Polymorphism Humain
chr: chromosome
cM: centimorgan
CODIS: Combined DNA Index System
ddNTP: dideoxynucleotide triphosphate
DNA: deoxyribonucleic acid
dNTP: deoxynucleotide triphosphate
DTC: direct-to-consumer
DVI: Disaster Victim Identification
EAS: East Asia
EDNAP: European DNA Profiling Group
e.g.: *exempli gratia*, for example
ENFSI: European Network of Forensic Science Institutes
ESS: European Standard Set
et al.: *et alia*, and others
EUR: Europe
EVC: externally visible characteristics
FDP: forensic DNA phenotyping
FFPE: formalin-fixed, paraffin-embedded
GDA: genetic distance algorithm
GWAS: genome-wide association studies
HGDP: Human Genome Diversity Project
IBD: identical by descent
i.e.; *id est*, that is
iiSNP: identity informative SNP
I_n: informativeness for assignment
indel: insertion/deletion polymorphism
IGG/FIGG: (Forensic) investigative genetic genealogy
ISFG: International Society for Forensic Genetics
MCMC: Markov Chain Monte Carlo
ME: Middle East
MH: microhaplotype
MHC: major histocompatibility complex
MPS: massively parallel sequencing
mtDNA: mitochondrial DNA
NAF: north Africa
NGS: next generation sequencing
NJT: neighbour-joining trees
OCE: Oceania

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ONT: Oxford Nanopore Technologies
PC: Principal Component
PCA: Principal Component Analysis
PCAST: President's Council of Advisors on Science and Technology
PCR: polymerase chain reaction
piSNP: phenotype informative SNP
PPi: pyrophosphate
PSD: Population Specific Divergence
RFLP: restriction fragment length polymorphism
RNA: ribonucleic acid
ROI: region of interest
RT-PCR: reverse transcription polymerase chain reaction
SAS: South Asia
SBE: single base extension
SNP: single nucleotide polymorphism
STR: short tandem repeat
SWGDAM: Scientific Working Group on DNA Analysis Methods
TGS: third-generation sequencing
TFS: Thermo Fisher Scientific
VCF: Variant Call Format
WGS: whole genome sequencing
X-SNP: X-chromosome SNP
X-STR: X-chromosome SNP
Y-SNP: Y-chromosome SNP
Y-STR: Y-chromosome SNP

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

RESUMO

Os STRs (*Short Tandem Repeats*), polimorfismos de ADN formados por secuencias curtas repetidas en tándem, son o tipo marcador por excelencia ao que se recorre na casuística forense para a resolución da gran maioría dos casos, xa que permiten unha análise rápida, son moi informativos e é unha ferramenta estandarizada, o que facilita á súa vez a comparación de resultados entre laboratorios a nivel internacional e o seu almacenamento en bases de datos de interese forense. As principais aplicacións deste tipo de marcadores baséanse na súa capacidade de individualización, é dicir, de diferenciar o perfil xenético de dúas persoas diferentes (fóra de xemelgos monocigotos). Así, isto permite, por exemplo, determinar a probabilidade dunha relación de parentesco entre dous presuntos familiares, mediante a valoración estatística dos seus perfís xenéticos ou establecer a probabilidade de que un sospeitoso contribuíse ao perfil da mostra obtida nunha escena do crime.

Os principais desafíos aos que se enfronta a xenética forense na actualidade engloban aquelas situacións nas que o uso dos marcadores STRs non achega suficiente información para a resolución do caso, o que ilustra as limitacións destes marcadores. Exemplos destes casos poden darse debido á distancia da relación de parentesco presentada, onde o poder estatístico dos paneis de STRs actuais non é o suficientemente elevado como para proporcionar uns resultados concluíntes; ou debido a que non existe un sospeitoso co que se poidan realizar comparacións e o cotexo nas bases de datos non dá lugar a ningunha coincidencia. A comunidade forense desenvolveu multitude de ferramentas creadas especificamente para abordar este tipo de casos. Por exemplo, en casos de parentesco afastado, pódese suplementar a información dos STRs con paneis de marcadores alternativos, incrementando así o número total de marcadores xenotipados e, en consecuencia, o poder estatístico da análise. No caso de carecer dun sospeitoso, existen as denominadas ferramentas de intelixencia de ADN ou *DNA intelligence tools*, que permiten inferir determinadas características do doante dunha mostra de ADN a partir da propia mostra, como a súa aparencia física (cor de ollos, pelo ou pel) ou a súa orixe bioxeográfica. Existen técnicas que mesmo, mediante o estudo de marcadores epixenéticos, permiten a predición da idade cronolóxica dun individuo. A información achegada por estas ferramentas, aínda que non sirva para identificar a un individuo, permite reducir o universo de sospeitosos e orientar desta forma a investigación policial para centrarse nun conxunto máis reducido de individuos.

Fóra da predición da idade, para a que se estudan cambios na metilación do ADN, estas ferramentas adoitan empregar marcadores alternativos como os SNPs (*Single Nucleotide Polymorphisms*), que son polimorfismos do ADN nos que hai unha variación nunha única posición da secuencia nucleotídica, ou microhaplotipos (MHs), conxuntos de SNPs agrupados

USC DE COMPOSTELA

en rexións de menos de 300 pb. A adopción de novas tecnoloxías de secuenciación, baseadas no emprego de instrumentos de secuenciación masiva en paralelo (*Massively Parallel Sequencing*, MPS), permiten o xenotipado simultáneo dunha gran cantidade de marcadores cunha moi alta sensibilidade (é dicir, partindo de cantidades moi reducidas de material xenético), o que as converte nunha ferramenta de gran utilidade para os estudos forenses. O interese dos investigadores nestas ferramentas, así como a súa aplicabilidade en análise forense constatouse nos últimos anos co desenvolvemento e implementación dunha gran variedade de paneis para o seu uso neste tipo de plataformas. Con todo, gran parte destas ferramentas limítanse á análise dun número relativamente reducido de marcadores, lonxe da gran capacidade que albergan os secuenciadores de MPS. Por iso, o desenvolvemento de paneis de maior escala, baseados no emprego de marcadores seleccionados coidadosamente para proporcionar a maior informatividade, é clave para a creación dunha nova xeración de paneis que permitan expandir os límites das ferramentas actuais.

Nos traballos comprendidos na presente Tese Doutoral descríbese a avaliación e desenvolvemento dun total de cinco paneis de MPS distintos para o seu uso forense: os paneis *VISAGE Basic Tool* (BT) e *VISAGE Enhanced Tool* (ET) *for appearance and ancestry*; un panel composto por 113 microhaplotipos; o panel *Eurasiaplex*-2 e o panel MPSplex. Os estudos das catro primeiras ferramentas céntranse na súa aplicabilidade para a inferencia da orixe bioxeográfica, mentres que os estudos do panel MPSplex están enfocados na súa implementación para casos de parentesco e identificación.

Paneis enfocados á realización predición de ancestralidad

Os paneis VISAGE BT e ET son paneis que foron desenvoltos dentro do marco de traballo do proxecto europeo VISAGE (*Visible Attributes Through GEnomics*) perseguindo un obxectivo común: a combinación nun único ensaio marcadores para a inferencia da orixe bioxeográfica e características físicas. A diferenza entre estes dous paneis radica no nivel de detalle e o número de características que analiza cada un e, polo tanto, no número de marcadores que engloban. O **panel VISAGE BT** representa a primeira etapa do proxecto e foi ideada como unha ferramenta de menor alcance, que incluía SNPs de predición de aparencia física de cor de pelo, pel e ollos; así como marcadores de predición de orixe bioxeográfica compilados para a diferenciación de seis grupos poboacionais – África subsahariana, Europa, sur de Asia, leste de Asia, nativo América e Oceanía. Os marcadores de orixe bioxeográfica foron seleccionados dunha colección de SNPs recolleitos de distintos paneis forenses xa establecidos e foron elixidos balanceando a diverxencia específica poboacional. Para comprobar a informatividade dos marcadores seleccionados para a predición da orixe bioxeográfica, recompiláronse xenotipos de distintos proxectos de secuenciación, como o *1000 Genomes Project* ou o *Simons Genome Diversity Project*, e avaliáronse mediante diversos métodos de análises: STRUCTURE, MDS (*Multi Dimensional Scaling*) e NJT (*Neighbour Joining Tree*); así como unha validación cruzada empregando o portal *Snipper* e usando un algoritmo de distancia xenética (*Genetic Distance*

Algorithm, GDA) dispoñible tamén en *Snipper*. Neles observouse, en xeral, unha boa separación dos grupos poboacionais que se tiñan como obxectivo, así como unha clasificación correcta no 100% dos casos na validación cruzada. Os individuos con orixe bioxeográfica mixta, mostraron unha alta correlación nas porcentaxes de coancestralidade entre as análises realizadas con STRUCTURE e as estimacións cos marcadores incluídos no panel *Human Origins*, un panel con máis de 500.000 SNPs. Esta correlación era especialmente alta en individuos con mestura de dúas orixe bioxeográfica, mentres que as predicións en poboacións onde os existía unha mestura de tres grupos poboacionais distintos mostraron un maior grao de imprecisión.

O **panel VISAGE ET**, unha versión mellorada e máis precisa do panel VISAGE BT, ademais de ter como obxectivo predicir un maior número de características físicas (cor de cellas, presenza de pencas, morfoloxía do cabelo e patróns de calvicie masculina), foi desenvolvido para lograr a diferenciación dun maior número de grupos poboacionais, incluíndo así mesmo unha maior variedade no tipo de marcadores de orixe bioxeográfiaca. Ademais de SNPs autosómicos, este panel inclúe SNPs de cromosoma X (X-SNPs), SNPs de cromosoma Y (Y-SNPs) e MHs. Os SNPs autosómicos foron elixidos non só para manter a diferenciación do mesmo número de poboacións que a BT a pesar de comprender un número máis limitado, senón para que tamén fosen capaces de distinguir poboacións de Oriente Medio. A estratexia que se seguiu para iso consistiu na diminución do número de marcadores para poboacións como África ou Oceanía (poboacións para as que, xeralmente, marcadores seleccionados para distinguir outras poboacións tamén son informativos); aumentando á súa vez o número de SNPs para distinguir as poboacións de Oriente Medio; e ampliando a selección de SNPs trialélicos informativos para a distinción de poboacións do sur de Asia e Europa. Seguindo unha estratexia de análise en dúas fases empregando o software STRUCTURE (realizando primeiro unha análise continental, seguido dunha segunda análise dos grupos poboacionais de Eurasia), conseguiuse diferenciar con éxito as sete poboacións, ademais de observar unha separación entre poboacións de Oriente Medio e norte de África. A poboación de referencia empregada para o norte de África consistía unicamente na poboación alxerina do CEPH e, por tanto, os resultados poderían limitarse a unha variación exclusiva para esta poboación e non ser representativos para o resto da rexión. A pesar diso, a detección de coancestralidade do norte de África en individuos de Marrocos, incluída como poboación de estudo, podería indicar que os resultados obtidos débense á diverxencia de ambos grupos poboacionais, aínda que a dispoñibilidade dun maior número de poboacións sería necesaria para poder confirmar estes resultados. Os MHs, pola súa banda, a pesar de que mostraron unha menor eficiencia na separación de grupos poboacionais (xa que soamente constituían unha pequena fracción dos marcadores incluídos no panel), mostraron moi bos resultados na predición de ancestralidade a partir de mesturas de ADN.

Respecto a os marcadores de cromosoma X e Y incluídos no panel, a pesar de que presentan unha capacidade máis limitada con respecto aos marcadores autosómicos, poden achegar información sobre as liñaxes materna e paterna en individuos con orixe bioxeográfico mixto.

USC DE SANTIAGO DE COMPOSTELA

Para a selección dos 85 Y-SNPs buscouse un equilibrio entre o espazo que ocupaban no *multiplex* e a diferenciación dos haplogrupos. Algúns marcadores, por exemplo, a pesar de ser informativos para o mesmo haplogrupo foron elixidos xa que presentaban unha distribución xeográfica distinta. Os marcadores do cromosoma X foron incluídos como un substituto de marcadores de ADN mitocondrial xa que, aínda que non supoñían un substituto completo do mesmo debido ao seu patrón de herdanza máis complexo (pódense herdar de ambos os proxenitores e presentan recombinación), os X-SNPs elixidos exhibían contrastes nas frecuencias alélicas para os cinco principais grupos de poboacións continentais (África subsahariana, América, Oceanía, Europa e leste de Asia). Ademais, ao tratarse de ADN nuclear, era posible a sua coamplificación co resto de marcadores nun mesmo ensaio. Cinco dos dezaseis SNPs do cromosoma X atópanse na rexión centromérica do cromosoma e os seus valores de recombinación estimados eran moi baixos. Por iso, explorouse a análise destes marcadores mediante dúas estratexias distintas: (i) empregando os 5 X-SNPs do centrómero como un bloque haplotípico e observando a súa distribución nos distintos grupos poboacionais; e (ii) empregando o conxunto de 16 marcadores nunha única análise de compoñentes principais (PCA) tendo en consideración o ligamento entre eles. A pesar de que o uso das cinco poboacións de referencia nas análises de PCA non deu lugar a unha separación efectiva, obtivéronse bos resultados coa redución a tres poboacións de referencia (América, África e Europa), e o estudo dos haplotipos do centrómero revelou a existencia de haplotipos específicos de poboación, que ademais representaban unha gran porcentaxe dos haplotipos presentes nas súas respectivas poboacións. O estudo da ancestralidade do cromosoma X, en conxunto cos haplogrupos do cromosoma Y, ademais, revelaron patróns de nesgo de sexo no entrecruzamento de distintas poboacións mixtas, como na poboación Calunga (unha poboación rural brasileira descendentes de escravos de orixe africana) na que se detectou nunha alta porcentaxe cromosoma Y de orixe europea e unha maioría de haplogrupos de cromosoma X de orixe africana.

Por outra banda, o **panel de 113 MHs** (102 marcadores autosómicos e 11 de cromosoma X) foi un panel desenvolto por de la Puente et al. (2020) para o seu uso como panel de identificación, seleccionando marcadores altamente polimórficos e, á súa vez, cun reducido tamaño de amplificación para poder empregarse en mostras degradadas. Neste estudo posterior, con todo, explorouse a utilidade deste panel máis aló de fins identificativos, comprobando a súa informatividade na realización de predicións de orixe bioxeográfica. A estratexia empregada para estas análises seguiu unha aproximación similar á empregada co panel VISAGE ET. En lugar de realizar unha única análise para todos os grupos poboacionais de forma conxunta, optouse de novo por dividilo en dúas fases – unha primeira análise incluíndo cinco grupos poboacionais continentais (África, Europa, leste de Asia, Oceanía e América) e unha análise subseguinte, dependente dos resultados desta primeira exploración, para separar seis grupos poboacionais de Eurasia (norte de África, Oriente Medio, Europa, sur de Asia e leste de Asia). A nivel de poboacións globais, logrouse realizar a clara separación dos cinco grupos, utilizando STRUCTURE, NJT e MDS. A diferenciación das poboacións de Eurasia, por outra banda, amosou ser máis limitada, especialmente para as poboacións do norte de África e Oriente Medio onde existía un maior solapamento. Ademais destes estudos, tamén se explorou o uso do panel

USC DE COMPOSTELA

no eido onde os MHs amosan un maior desempeño: as mesturas de ADN. Empregando mesturas simuladas de dous individuos, conseguiuse inferir con éxito a orixe bioxeográfica de ambos contribuíntes nas tres probas realizadas, mesmo con perfís parciais e en casos onde a orixe bioxeográfica dos dous contribuíntes é a mesma ou con individuos de orixe bioxeográfica mixta.

En canto ao **panel *Eurasiaplex*-2**, o desenvolvemento desta ferramenta enfocouse na diferenciación das poboacións do sur de Asia (*South Asia*, SAS) do mesmo xeito que a primeira versión do mesmo. Con todo, a diferenza do seu predecesor, o deseño do panel realizouse empregando un novo tipo de marcador de orixe bioxeográfica: SNPs cun alelo con especificidade absoluta para estas poboacións. Debido á novidade deste tipo de marcador, realizouse unha nova procura partindo dos datos do *1000 Genomes Project*, filtrando aqueles SNPs bialélicos cunha frecuencia do alelo alternativo maior do 10% para SAS e menor do 1% para América, leste de Asia, Oceanía e Europa. Os SNPs candidatos resultantes deste cribado foron filtrados en base a parámetros de calidade (como consistencia das frecuencias observadas en distintos estudos) e procurando seleccionar marcadores distanciados entre si para evitar seleccionar SNPs en desequilibrio de ligamento, dando lugar a un total de 36 SNPs que conformarían a versión final do panel. Seis destes SNPs seleccionados, situados no cromosoma 16, foron avaliados para ver a asociación entre estes marcadores. Os valores de recombinación suxeriron un ligamento forte entre tres deles debido á súa localización próxima ao centrómero do cromosoma, a pesar de estar relativamente separados en distancia física. Con todo, os datos poboacionais analizados indicaban que entre estes marcadores existía unha recombinación maior á esperada xa que se observaba unha disrupción do haplotipo específico de SAS, suxerindo que a asociación entre estes tres marcadores é máis baixa da estimada.

Tras analizar a presenza destes alelos en distintas poboacións, viuse que os individuos das poboacións do subcontinente indio presentaban un alto número de alelos, fronte a unha presenza case nula nas poboacións fóra do sur de Asia – unicamente nas poboacións paquistanís observouse unha distribución intermedia do número de alelos específicos. Por mor destas observacións expúxose a posibilidade da implementación de análises estatísticas baseadas no cálculo de alelos, que poderían explorarse como unha alternativa a análises máis complexas como STRUCTURE ou PCA. A capacidade deste panel para distinguir o resto de grupos poboacionais, como cabería esperar, foi practicamente nula, debido a que os SNPs seleccionados unicamente mostran variabilidade nas poboacións do sur de Asia. Este estudo conceptual, con todo, supón un éxito na proba de concepto do potencial que ten deste tipo de marcadores con alelos específicos, e constitúe o primeiro paso do desenvolvemento dun panel de maior escala conformado por alelos específicos para cada un dos grupos poboacionais.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Desenvolvemento dun panel de gran escala con fins identificativos

O **panel MPSplex**, pola súa banda, ideouse para a identificación de persoas desaparecidas e para a realización de estudos de parentesco afastado, cando o grao de relación entre os individuos vai máis aló de primeiro ou segundo grao, que constitúe o límite máximo que se consegue resolver mediante os *kits* de STRs convencionais. Co obxectivo de maximizar a informatividade do panel, optouse non só por seleccionar SNPs atendendo a súa heterocigosidade, senón que a procura centrouse na selección de SNPs que presentasen tres alelos distintos (SNPs trialélicos), xa que grazas á presenza deste terceiro alelo estes marcadores poden presentar un poder de discriminación maior ao dos SNPs bialélicos. Para o deseño do panel recorreuse, novamente, aos datos da Fase 3 do *1000 Genomes Project*, onde se identificaron un total de 271.934 marcadores candidatos. Realizouse un filtrado en base á heterocigosidade dos marcadores para seleccionar aqueles con unha heterocigosidade maior de 0,5 (límite máximo ao que chega un SNP bialélico) e un posterior cribado para diminuír o posible ligamento entre os marcadores, elixindo SNPs separados tanto fisicamente como en distancia de mapa (mínimo 1 centimorgan). Tamén foron eliminados SNPs que se sospeitaba que puideran presentar problemas ao ser incorporados nun *multiplex*, como elementos repetitivos nas rexións adxacentes ou indicios de atoparse en rexións duplicadas. Os 1.411 marcadores restantes tras aplicar os filtros foron incluídos nun prototipo de panel QIAseq (Qiagen) para comprobar o seu funcionamento, e novamente excluíronse diversos SNPs por presentar baixa cobertura, discordancias con ADNs control ou máis de dous alelos neste tipo de mostras. Os 1.270 SNPs trialélicos restantes (1.241 autosómicos e 29 de cromosoma X) foron suplementados con 46 MHs, conformando así a selección final de marcadores.

Para a avaliación dos SNPs seleccionados, así como para comprobar como puideran afectar outros factores como a adopción de SNPs trialélicos e a selección de marcadores con frecuencias alélicas balanceadas para maximizar a súa heterocigosidade, realizáronse distintos tipos de simulacións. Grazas a elas, observouse que os SNPs trialélicos que formaban parte do panel permitían resolver casos de parentesco de ata terceiro grao (curmáns), así como algunhas relacións de quinto grao (curmáns segundos). Así mesmo, constatouse o maior poder de discriminación dos SNPs seleccionados fronte a un panel equivalente de SNPs bialélicos e que, en xeral, os SNPs presentan unha maior informatividade cando as frecuencias alélicas estaban balanceadas – aínda que isto último non era tan notorio en paneis de gran tamaño debido á maior probabilidade de compartir alelos pouco frecuentes. Por outra banda, aínda que non fose o obxectivo do panel, avaliouse a capacidade destes marcadores para a realización de análise de orixe bioxeográfica. Tal e como se observou no estudo do panel de 113 MHs, a pesares de estar deseñado para non incluír SNPs con grandes contrastes de frecuencias alélicas entre distintos grupos poboacionais, conseguiuse realizar unha separación de seis grupos poboacionais (África subsahariana, Europa, sur de Asia, leste de Asia, Oceanía e América).

Nun traballo posterior, realizouse a validación forense do panel en tres plataformas de MPS distintas: Ion S5, MiSeq e GeneReader. A versión do panel que se empregou nesta avaliación fora modificada lixeiramente, xa que incluía 55 marcadores adicionais para a predición da orixe

bioxeográfica. Ademais, ao estar implementado nun panel QIAseq, o panel albergaba a peculiaridade de utilizar unhas secuencias de ADN denominadas UMIs (*Unique Molecular Indices*). Estas moléculas consisten en secuencias de ADN aleatorias que se engaden aos fragmentos de ADN durante a preparación de librerías previo á amplificación por PCR. Tras a secuenciación, esta marcaxe permite recoñecer todas aquelas lecturas que se orixinaron a partir do mesmo fragmento orixinal de ADN, e durante o procesamento bioinformático únense para xerar unha lectura consenso (denominada lectura UMI), que permite eliminar potenciais erros introducidos durante o proceso de amplificación ou de secuenciación.

A validación realizouse utilizando un conxunto de mostras común que foi secuenciado nas tres plataformas: mostras de referencia, tanto para avaliar o funcionamento do panel como para utilizar en estudos de concordancia, dilucións seriadas para comprobar a sensibilidade do panel e un estudo de mesturas. Ademais, tamén se incluíu, para ilustrar a aplicabilidade do panel, un caso de parentesco real, no que os marcadores xenotipados mediante electroforeses capilar non proporcionaron un resultado concluínte. Os resultados deste estudo demostraron un moi bo funcionamento xeral do panel, onde a gran maioría dos marcadores presentaban unha cobertura por encima do límite de detección. Así mesmo, observouse que presentaba unha alta sensibilidade nas tres plataformas analizadas, obtendo perfís case completos con 62 pg de ADN inicial; e unha alta capacidade para detectar mesturas, non só debido ao desbalance no número lecturas, senón a que se detectaron ata tres alelos distintos en varios loci. Grazas ao estudo de concordancia, así como ao transcurso da avaliación, puidéronse identificar varios marcadores potencialmente problemáticos como, por exemplo, marcadores discordantes con resultados de proxectos de secuenciación de xenomas completos debido a problemas no aliñamento ou SNPs que podían atoparse en rexións duplicadas, xa que se observaban tres alelos distintos en mostras provenientes dun único individuo. Finalmente, no estudo de parentesco púidose constatar o gran poder de discriminación do panel, logrando un valor de LR (*likelihood ratio*) de 42 ordes de magnitude superior ao obtido mediante a combinación de dous *kits* distintos de STRs, a pesar de obter xenotipos parciais e subliñando o gran poder de discriminación do panel e a súa aplicabilidade en casos de parentesco afastado.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

SUMMARY

STRs (Short Tandem Repeats), DNA polymorphisms formed by a core of 2-7 base pairs (bp) that is repeated in tandem, represent the gold standard in forensic genetics and are used for the resolution of the vast majority of cases. They can be used to perform a rapid analysis, are very informative and are a standardized tool, which in turn facilitates the comparison of results between laboratories and their storage in databases of forensic interest. The main application of this type of marker is based on its capacity for individualization, i.e. to differentiate the genetic profile of two different people (except for monozygotic twins). This allows, for example, to determine the probability of a kinship relationship between two suspected relatives through the statistical evaluation of their genetic profiles, or to establish the probability that a suspect has contributed to the profile of the sample obtained at a crime scene.

The main challenges facing forensic genetics today include situations where the use of STRs does not provide sufficient information for the resolution of certain scenarios, illustrating the limitations of these markers. Examples of these limitations are distant kinship cases, where the statistical power of current STR panels is not high enough to provide conclusive results; or when there is no suspect with whom comparisons can be made and there are no matches with databases. The forensic community has developed a myriad of tools created specifically to address these types of situations. For example, for distant kinship, the STR information can be supplemented with panels of alternative markers, consequently increasing the statistical power of the analysis. In the absence of a suspect, there is a set of tools known as “DNA intelligence tools” that allow certain characteristics of an unknown DNA donor to be inferred, such as their physical traits (eye, hair or skin color) or biogeographical ancestry. There are even techniques that, through the study of epigenetic markers, allow the prediction of an individual's chronological age. The information provided by these tools, even if it does not serve to identify an individual, makes it possible to reduce the pool of suspects and thus serve as a lead for the police investigation to focus on a smaller set of individuals.

Except for age prediction, for which changes in DNA methylation are studied, these tools usually employ markers such as Single Nucleotide Polymorphisms (SNPs), which are DNA polymorphisms in which there is a variation in a single position of the nucleotide sequence, or microhaplotypes (MHs), sets of SNPs grouped in regions shorter than 300 bp. The adoption of new sequencing technologies, based on the use of Massively Parallel Sequencing (MPS) methodologies, allows the simultaneous genotyping of a large number of markers with very high sensitivity (i.e. starting from very small amounts of genetic material), which makes them a very useful tool for forensic studies. The interest of researchers in these tools, as well as their applicability in forensic analysis, has been corroborated in recent years with the development

USC DE COMPOSTELA

and implementation of a wide variety of panels for use in this type of platforms. However, most of these tools are limited to the analysis of a relatively small number of markers, not fully utilizing the great capacity of MPS sequencers. Therefore, the development of larger scale panels, based on the use of carefully selected markers to provide the highest informativity, is key to the creation of a new generation of panels that will expand the limits of current tools.

The works included in this PhD Thesis describe the evaluation and development of a total of five different MPS panels for forensic use: the VISAGE Basic Tool (BT) and VISAGE Enhanced Tool (ET) for appearance and ancestry; a panel composed of 113 microhaplotypes (MHs); the *Eurasiaplex*-2 panel; and the MPSplex panel. Studies of the first four tools focus on their applicability for ancestry inference, while studies of the MPSplex panel focus on their implementation for kinship and identification applications.

Panels evaluated for ancestry inference applications

The VISAGE BT and ET panels are panels that were developed within the European VISAGE (*VISible Attributes Through GEnomics*) project framework, pursuing a common objective: the combination in a single assay of markers for the inference of ancestry and physical characteristics. The difference between these two panels lies in the level of detail and the number of features analyzed by each panel, and thus in the number of markers that they encompass. The **VISAGE BT panel** represents the first stage of the project and was devised as a smaller scope tool, which included physical appearance prediction SNPs for hair, skin and eye color; as well as biogeographic origin prediction markers compiled for the differentiation of six population groups - sub-Saharan Africa, Europe, South Asia, East Asia, Native America and Oceania. Ancestry markers were selected from a collection of SNPs collected from different established forensic panels and were chosen by balancing population-specific divergence. To test the informativeness of the markers selected for ancestry prediction, genotypes were collected from different sequencing projects, such as the 1000 Genomes Project or the Simons Genome Diversity Project, and evaluated by different analysis methods: STRUCTURE, MDS (Multidimensional Scaling) and NJT (Neighbour Joining Tree); as well as a cross-validation using the Snipper portal and using a Genetic Distance Algorithm (GDA), also available in Snipper. In general, a good separation of the targeted population groups was observed, as well as a correct classification in 100% of the cases in the cross-validation. Admixed individuals showed a high correlation in the co-ancestry percentages between the analyses performed with STRUCTURE and the estimates with the markers included in the Human Origins panel, a panel with more than 500,000 SNPs. This correlation was especially high in admixed individuals with two co-ancestries, whereas predictions in populations with a three-way co-ancestry pattern from showed a higher degree of imprecision.

The **VISAGE ET panel**, an improved and more refined version of the VISAGE BT panel, in addition to aiming to predict a larger number of physical characteristics (eyebrow color,

USC DE COMPOSTELA

freckling, hair shape and male pattern baldness), was developed to achieve differentiation of a larger number of population groups while also including a greater variety of ancestry markers. In addition to autosomal SNPs, this panel includes X-chromosomal SNPs (X-SNPs), Y-chromosomal SNPs (Y-SNPs) and MHs. The autosomal SNPs were chosen not only to maintain differentiation of the same number of populations as BT despite comprising a more limited number of markers, but also to be able to distinguish Middle Eastern populations. The strategy followed for this consisted of decreasing the number of markers for populations such as Africa or Oceania (populations for which, generally, markers selected to distinguish other populations are also informative); increasing in turn the number of SNPs to distinguish Middle Eastern populations; and expanding the selection of informative tri-allelic SNPs for distinguishing South Asian and European populations. Following a two-phase analysis strategy using STRUCTURE software (performing a continental analysis first, followed by a second analysis of the Eurasian population groups), we successfully differentiated the seven target population groups, observing also a distinction between Middle Eastern and North African populations. The reference population used for North Africa however consisted only of the Algerian CEPH population and, therefore, the results could be limited to variation unique to this population and not be representative for the rest of the region. Despite this, the detection of North African co-ancestry in individuals from Morocco, included as a study population, suggests that the results obtained may reflect the divergence of both population groups, although the availability of a larger number of populations would be necessary to be able to confirm these results. The MHs, on the other hand, although they showed a lower efficiency in the separation of population groups (since they only constituted a small fraction of the markers included in the panel), showed very good results in the prediction of ancestry from DNA mixtures.

Although the X and Y chromosome markers included in the panel have a more limited capacity than the autosomal markers, they can provide information on maternal and paternal lineages in admixed individuals. For the selection of the 85 Y-SNPs, a balance was sought between the space they occupied in the multiplex and the differentiation of different haplogroups. Some markers, although informative for the same haplogroup, were chosen because they had a different geographic distribution. X-chromosomal markers were included as a substitute for mitochondrial DNA markers because, although they do not represent complete alternative for mitochondrial DNA due to their more complex inheritance pattern (they can be inherited from both parents and show recombination), the X-SNPs chosen exhibited contrasting allele frequencies for the five major continental population groups (sub-Saharan Africa, America, Oceania, Europe and East Asia). Moreover, since they are located in a nuclear chromosome, it is possible to co-amplify them with the rest of autosomal and Y-markers in the same assay. Five of the sixteen SNPs on the X chromosome are found in the centromeric region of the chromosome and their estimated recombination values were very low. Therefore, the analysis of these markers was explored using two different strategies: (i) using the 5 centromeric X-SNPs as a haplotypic block and observing their distribution in the different population groups; and (ii) using the set of 16 markers in a single principal component analysis (PCA) taking into consideration the linkage between them. Although the use of the five

USC DE COMPOSTELA

reference populations in the PCA analyses did not result in an effective separation, good results were obtained with the reduction to three reference populations (America, Africa and Europe), and the study of centromere haplotypes revealed the existence of population-specific haplotypes, which also represented a large percentage of the haplotypes present in their respective populations. The study of X chromosome ancestry, in conjunction with Y chromosome haplogroups, further revealed patterns of sex bias in the interbreeding of different admixed populations, such as in the Kalunga population (a rural Brazilian population descended from slaves of African origin) in which a high percentage of X chromosome of African origin and a majority of Y chromosome haplogroups of European origin were detected.

On the other hand, the **panel of 113 MHs** (102 autosomal and 11 X chromosome markers) was conceived by de la Puente et al. (2020) as an identification panel, developed by selecting highly polymorphic markers with reduced amplification size suitable for the analysis of degraded samples. In the subsequent study, however, we explored the utility of this panel beyond identification purposes, testing its informativeness in making ancestry predictions. The strategy employed for these analyses followed a similar strategy to that employed with the VISAGE ET panel. Instead of performing a single analysis, for all population groups together, the evaluation was split it into two phases – a first analysis including five continental population groups (Africa, Europe, East Asia, Oceania and America) and a subsequent analysis, dependent on the results of this first STRUCTURE run, to separate six Eurasian population groups (North Africa, Middle East, Europe, South Asia and East Asia). At the continental level, a successful separation of the five groups was achieved using STRUCTURE, NJT and MDS. The differentiation of the Eurasian populations on the other hand was shown to be more limited, especially for the North African and Middle Eastern populations where there was more overlap. In addition to these studies, the use of the panel was also explored in the field where MHs are most proficient: DNA mixtures. Using simulated mixtures of two individuals, the ancestry of both contributors was successfully inferred in all three tests performed, even with partial profiles and in cases where the ancestry of the two contributors was the same or when an individual was admixed.

Regarding the ***Eurasiaplex*-2 panel**, the development of this tool focused on the differentiation of South Asian (SAS) populations, similarly to the first version of the panel. However, unlike its predecessor, the panel was designed using a new type of ancestry marker: SNPs with an allele with absolute specificity for these populations. Due to the novelty of this type of marker, a new search was performed using data from the 1000 Genomes Project, selecting biallelic SNPs with a frequency of the alternative allele greater than 10% for SAS and lower than 1% for America, East Asia, Oceania and Europe population groups. The candidate SNPs resulting from this screening were filtered based on quality parameters (such as consistency of the frequencies observed in different studies) and aiming to select markers distant from each other so as to avoid selecting SNPs in linkage disequilibrium, resulting in a total of 36 SNPs that would constitute the final selection of markers. Six of these selected SNPs, located on chromosome 16, were evaluated for linkage between them. Recombination values

USC DE SANTIAGO DE COMPOSTELA

suggested strong linkage between three of them due to their location close to the centromere of the chromosome, despite being relatively separated in physical distance. However, the population data analyzed indicated that the recombination between these markers was higher than expected since a frequent disruption of the specific SAS haplotype was observed, suggesting that the association between these three markers is lower than estimated.

After examining the allele distribution of these SNPs in different populations, it was observed that individuals from Indian subcontinent populations had a high number of SAS-specific alleles, whereas to almost no presence in populations outside South Asia - only Pakistani populations presented an intermediate distribution of specific alleles. Based on these findings, the possibility of implementing statistical analyses based on allele counts was considered as a potential alternative to more complex analyses such as STRUCTURE or PCA. The separation of other population groups however was, as expected, not possible, as the selected SNPs showed variability only within South Asian populations. Nonetheless, this pilot study served as proof of concept of the potential of population-specific markers, and is the first step towards the development of a large-scale ancestry panel comprising population-specific markers for each of the population groups.

Development of a large-scale panel for identification purposes

The **MPSplex panel**, on the other hand, was designed as a panel for missing persons identification and distant kinship studies, when the degree of relationship between individuals goes beyond the first or second degree, which is the maximum limit that can be resolved with routine STRs kits. In order to maximize the informativeness of the panel, markers were chosen not only according to their heterozygosity, but the search focused on the selection of SNPs that presented three different alleles (tri-allelic SNPs), given that these markers can present a greater power of discrimination than biallelic SNPs. For the design of the panel, the data from Phase 3 of the 1000 Genomes Project was used once again, where a total of 271,934 candidate markers were identified. A filtering was performed based on the heterozygosity of the markers to select those with a heterozygosity greater than 0.5 (maximum heterozygosity value that biallelic SNP can reach) and a subsequent screening to reduce the possible linkage between markers, choosing SNPs separated both physically and in map distance. SNPs that could present problems when incorporated into a multiplex, e.g., presenting repetitive elements in adjacent regions or indications of being in duplicated regions, were also removed. The 1,411 markers remaining were included in a prototype QIAseq panel (Qiagen) to test their performance, and several SNPs were removed due to low coverage, discordances with control DNAs or when more than two alleles were observed in single-source samples. The remaining 1,270 tri-allelic SNPs (1,241 autosomal and 29 X-chromosomal) were supplemented with 46 MHs, constituting the final selection of markers for this panel.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

For the evaluation of the selected SNPs, as well as to assess how other factors such as the adoption of tri-allelic SNPs, different types of simulations were carried out. It was observed that the tri-allelic SNPs that were part of the panel allowed the resolution of cases of kinship up to third degree (first cousins), as well as some instances of fifth degree (second cousins). It was also observed that the selected SNPs had greater discriminatory power than an equivalent panel of biallelic SNPs and that, in general, SNPs were more informative when allele frequencies were balanced – although the latter was not as noticeable in large panels due to the greater likelihood of observing rare alleles. On the other hand, although it was not the objective of the panel, the ability of these markers to perform biogeographic ancestry analyses was evaluated. As was observed in the study of the panel of 113 MHs, despite being designed not to include SNPs with large contrasts in allele frequencies between different population groups, a separation of six population groups (sub-Saharan Africa, Europe, South Asia, East Asia, Oceania and America) was achieved.

In subsequent work, the forensic validation of the panel was performed on three different MPS platforms: Ion S5, MiSeq and GeneReader. The version of the panel used in this evaluation had been slightly modified, as it included 55 additional markers for ancestry inference. In addition, since the markers had been implemented in a QIAseq panel, the panel also incorporated DNA sequences called UMIs (Unique Molecular Indices). These molecules consist of random DNA sequences that are added to DNA fragments during library preparation prior to PCR amplification. After sequencing, the UMI tagging makes it possible to recognize all the reads that have originated from the same original DNA fragment, and during bioinformatics processing they are merged to generate a consensus read (called UMI read), enabling the removal of potential errors introduced during the amplification or sequencing process.

The validation was performed using a common sample set that was sequenced on all three platforms: reference samples, used both to evaluate the performance of the panel and for concordance studies, a serial dilution to test the sensitivity of the panel, and a mixture study. In addition, a distant kinship case, in which markers typed by capillary electrophoresis did not provide a conclusive result, was also included to illustrate the applicability of the panel. The results of this study demonstrated an overall good performance of the panel, with the vast majority of markers showing coverage above the detection limit. Likewise, it was observed that the panel presented high sensitivity in the three platforms analyzed, obtaining almost complete profiles with 62 pg of initial DNA; and a high capacity to detect mixtures, not only due to the imbalance in the number of reads, but also because three different alleles were detected in several loci. Thanks to the concordance study, as well as throughout the evaluation, several potentially problematic markers could be identified. For example, some markers showed discordances with the results from whole genome sequencing projects caused by alignment artifacts, while some SNPs were suspected of being in duplicated regions, as three different alleles were observed in single-source samples. Finally, in the parentage study it was possible to confirm the great power of discrimination of the panel, achieving an LR (likelihood ratio)

USC DE SANTIAGO DE COMPOSTELA

value of 42 orders of magnitude higher than that obtained by combining two different STRs kits, despite having obtained partial genotypes.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

1. INTRODUCTION

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

1. INTRODUCTION

1.1 The advent of Forensic Sciences

Contrary to popular belief, in which forensic practices are associated with cutting-edge crime-solving methodologies (often derived from TV shows of questionable scientific accuracy that show a distorted reality, known as 'the CSI effect'), the procedures followed by forensic practitioners and the real definition of the term itself are far from that. The word *forensics* is a borrowing from Latin that etymologically indicates "of or belonging to the Forum, of, or connected with, the law courts" (Oxford English Dictionary, 2023) and the Forensic Sciences include all the scientific methodologies that are used in legal proceedings, to provide evidence that aids the resolution of a legal conflict; as, for instance, aiming to help locate the perpetrator of a crime and providing evidence which proves their guilt.

Even though the exact moment when this term was first used is uncertain, the use of forensic methodologies and the regulation of medical procedures by the law date back to ancient civilizations. In Ancient Egypt, researchers found evidence of the earliest application of autopsies (Kharoshah et al., 2011). The Code of Hammurabi, written by the Mesopotamia's King Hammurabi c. 1750 BCE, was one of the first legal codes registered in human history and among the 300 laws included in the Code, for instance, we can find the widely known and famous phrase "an eye for an eye" and, more relevant for this context, what represents the first conjoined written evidence of medicine and law. It included several regulations for practitioners of medicine and the punishment of certain actions – for instance, the practitioner would lose a hand if their intervention resulted in the death of the patient (when they were a free citizen, otherwise the punishment would be different depending on the social class) (Pearn, 2016). It should also be mentioned the paradigm example of the autopsy conducted by the physician Antistius of Julius Caesar after his assassination on 44 BCE, where he noted the 23 wound stabs that were inflicted on the Caesar; or the manuscript *Hsi Yuan Lu* (that could be translated to English as "The Washing Away of Wrongs"), written by Chinese lawyer Sung Tz'u in the 13th century, which is considered to be the first systematic approach to forensic science; and the author is acknowledged as the first Forensic Entomologist (Benecke, 2001).

Over the centuries, and thanks to many scientific developments, multiple disciplines have been incorporated into the forensic sciences, including toxicology, ballistics, entomology and the branch in which this Thesis is comprised: genetics. All of these fields together and from different perspectives, help to put together the pieces of a jigsaw, attempting to provide unbiased, objective data to the judge or jury, who will be the ones that make the ultimate decision.

USC
DE SANTIAGO
DE COMPOSTELA

As Phillips relates from personal experience (Phillips, 2015), and multiple studies have confirmed so far, visual testimonies tend not to be the most reliable source of information, both the victim or any interlopers' glances that witnesses make can be biased and volatile. Multiple studies (Fitzgerald et al., 2018; Heaton-Armstrong, 1995; Spinney, 2008) describe how the certainty of memories can begin to fade just a fortnight after the events (Gustafsson et al., 2022). But how reliable can the information provided by forensic sciences be? A 2016 report by the President's Council of Advisors on Science and Technology (PCAST), titled *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, critically evaluated the state-of-the-art in the forensics field at the time. The report raised awareness of the issues that challenged the validity of some of the methodologies that were used as evidence in a court of law. In particular, PCAST questioned the subjectivity, high false-positive rates or the absence of sufficient studies to assess the reliability and reproducibility that were present on some methodologies for the comparison of features such as bitemarks, latent fingerprints, firearm marks or even hair morphology (President's Council of Advisors on Science and Technology, 2016).

Despite this, the present-day forensic field is still in constant evolution, trying to provide the most unbiased, objective results to aid in a court of law and prevent any miscarriage of justice that could lead to the wrongful conviction of an innocent person. It must be noted nonetheless, that it is not the objective to provide the truth, but they are just tools, and even objective data can be distorted – e.g., the widespread occurrence of the prosecutor's fallacy, a misinterpretation of the data caused by statistical statements that are inverted, causing the transposition of conditionals – and in order to guarantee the quality of the results, several quality standards must be met. Otherwise, it could compromise the quality of the results and, therefore, their veracity, which could in turn cause the evidence to be disregarded in court. A notable example of this is the O.J. Simpson trial, a controversial case involving the ex-NFL player, who had presumably assassinated his ex-wife and her partner. The case drew a lot of attention from the general public and had a large media coverage due to the high-profile nature of the case and all the surrounding events (including a 45-minute police chase on the I-5 freeway). During the trial, the detailed questioning made by the defence team about handling of evidence by the police, led to the discreditation of the DNA evidence and to the acquittal of the suspect– although after a civil suit two years later he was found guilty of those crimes (as denoted by a famous artist's lyric, '*nothing good starts in a getaway car'*). Since then, laboratories have increased the security measures and increased their standards to reduce contamination of evidential material and ensure the quality of the analyses made (Butler, 2005).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.2 THE DISCOVERY OF THE DNA AND THE FIRST STEPS IN FORENSIC GENETICS

1.2.1 The haemogenetics and protein markers

The application of genetic methodologies to assist in legal procedures marks one of the most recent additions to forensic sciences, whose journey began in the early 20th Century with the birth of haemogenetics. In the year 1900, Karl Landsteiner first described the results of experiments, in which he was able to categorize blood samples by mixing blood from two different individuals, which would cause the agglutination of the red blood cells (Landsteiner, 1900). In a publication one year later, he further elaborated on those results, classifying blood samples into three groups (Landsteiner, 1901), which would become the ABO blood system, one of the most important classifications of blood types. The hereditary basis of ABO blood groups was demonstrated in 1908 by Esptein and Ottenberg, and posterior experiments from von Dungern and Hirschfeld proved that these markers followed a mendelian inheritance pattern (von Dungern & Hirschfeld, 1911). To a certain extent, the ABO blood group allowed the exclusion of individuals, and the first example of its applicability was in 1915 by Leone Lattes in paternity testing. Despite that, its capacity to discriminate individuals is very limited: the system only presented 4 different groups and their frequency distribution are also uneven, with not all groups frequent.

Since the ABO system was elucidated, other blood group markers and blood serum proteins were discovered and implemented in the analyses to increase the capacity of individualization, including Rhesus (RH), the MNSs or the P systems. However, they still presented several limitations: they had low variability (the match probability of two unrelated individuals was ~0.01–0.001), the instability of the proteins that were used for the testing and the large amount of sample required, represented a severe limitation on their use in forensic casework and other fluids of forensic interest do not contain all these markers (Geserick & Wirth, 2012; Goodwin et al., 2010; Jobling & Gill, 2004).

1.2.2 RFLPs and DNA fingerprinting

In parallel to the evolution of hemogenetics, several breakthroughs in the genetics field helped the establishment of the bases of forensic genetics including the discovery of the structure of DNA (Franklin & Gosling, 1953; Watson & Crick, 1953), and advances in molecular biology that revolutionized DNA analysis, like Southern blotting (Southern, 1975) or Sanger sequencing (Sanger et al., 1977b) (the latter explained in more detail in Section 1.5.2.1). In 1984, British geneticist Alec Jeffreys and his collaborators described what they termed ‘DNA fingerprints’. They used a type of analysis known as RFLP (restriction fragment length polymorphism), based on the use of tandem-repetitive regions of DNA or 'minisatellites', a type of DNA polymorphism to obtain patterns that were ‘*completely specific to an individual (or to his or her identical twin)*’ and could be used for human identification (Jeffreys et al.,

USC DE SANTIAGO DE COMPOSTELA

1985). This technique consisted of the fragmentation of the DNA sample using restriction enzymes, a posterior separation using a gel electrophoresis, and then DNA fragments were transferred to a membrane for the detection of the loci using labelled probes. Later improvements of the technique led to the use of single-locus probes (SLP), each one hybridizing to a single locus to increase the reproducibility of the results and simplify the interpretation. Although this method constituted the first use of DNA in a criminal case (the Pitchfork case), similarly to the hemogenetics, its applicability in criminal cases was still limited, drawbacks that would later be solved by PCR-based methodologies.

Another well-known and controversial example of the application of this methodology was during the Clinton–Lewinsky scandal resulting in the impeachment of the US President, William Jefferson Clinton. RFLP analysis with *Hae*III restriction enzyme and the use of 7 SLPs was applied to the DNA retrieved from a semen stain on Lewinsky's dress, that helped prove sexual activity between the President and his secretary, following allegations of the illicit affair (Butler, 2005).

1.2.3 The PCR and steps towards modern-day forensics

DNA analysis constituted a significant advance from methodologies based on protein analysis, but the sensitivity of such analyses still presented a major handicap, as the biological material required hindered its widespread use, together with the difficulty to analyse degraded material and the lack of standardization among laboratories (Goodwin et al., 2010). Nonetheless, in 1986, Karry Mullis presented a new tool that would become the mainstream technique in forensic genetics and became, to this day, the basic tool in most forensic analyses: the polymerase chain reaction or PCR (Mullis et al., 1986). This method, whose biological functioning is explained in more detail in Section 1.5.1, created a system to produce multiple copies of a region of interest (ROI) from small amounts of DNA, reducing the amount from µg to less than 1 ng. However, the implementation of PCR in forensic genetics was not immediate and it co-existed with DNA fingerprinting techniques for several years.

Initially, sequence polymorphisms in the HLA DQα locus were analysed using PCR procedures (Helmuth et al., 1990), but they were used as a last resort when SLP techniques failed due to their lower discrimination power and the difficulty of mixture interpretation. But soon the forensic community started to focus on the analysis of other types of markers, which through the following years would become established as the gold standard in the field: the development of STRs (short tandem repeats) or microsatellites (section 1.3.1). Several factors contributed to this development in a short space of time, first, many STRs were characterized, including the STRs that are analysed in routine analyses nowadays, such as D12S391, vWA, D21S11, D1S1656 or D18S51 (Kimpton et al., 1992; Lareu et al., 1998; Lareu et al., 1996; Sharma & Litt, 1992; Straub et al., 1993). Second, STRs started to be implemented into PCR multiplexes (co-amplification of several ROI within the same PCR reaction) using automated

USC DE COMPOSTELA

sequencers, allowing for the automation of the analyses that would allow the simultaneous screening of several loci; and other markers such as amelogenin, that provided sex determination, began to be included in such multiplexes (Kimpton et al., 1993; Mannucci et al., 1994; Oldroyd et al., 1995).

The nature of the results, where each STR was assigned their corresponding alleles in an unambiguous manner, allowed profiles from individuals to be stored in databases and for these to be screened when searches and profile comparisons were made. In 1995, the UK was the first country to create a national database and three years later Germany, The Netherlands and Austria had already implemented their own national databases (Martin et al., 2001). On the 27th May 2005, seven countries in the European Union (Austria, Belgium, France, Germany, Luxembourg, the Netherlands and Spain) signed the treaty of Prüm, that facilitated the exchange of DNA data. With these developments, a new issue arose: the need to have a standardized set of core STRs to harmonize the analysis and facilitate the exchange of information among different laboratories, ensuring the overlap of STRs used in databases from different countries. As a result, in 1999 ENFSI (European Network of Forensic Science Institutes) in collaboration with the European DNA Profiling Group (EDNAP) selected 7 STRs (TH01, vWA, FGA, D21S11, D3S1358, D8S1179 and D18S51) to form the European Standard Set (ESS). In the same way, the FBI had established two years before this a core set of STRs for the US named Combined DNA Index System (CODIS) containing a total of 13 markers: CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 and D21S11.

Due to the relatively small number of loci included in the core set and given the number of comparisons there was a high probability of encountering a significant number of random matches. Thus, in 2009 the EES was expanded to include 5 additional loci – two of which, D12S391 and D1S1656, had been characterised by Lareu and collaborators at the University of Santiago de Compostela. The expansion on the ESS also sought to incorporate and adapt the existing markers to adaptation of mini-STRs (STRs that have reduced-size amplicons) to increase the sensitivity and robustness of the assays, which proved to be especially useful when dealing with degraded samples (Butler et al., 2003; Gill et al., 2006). In 2017, CODIS underwent an expansion as well, including seven additional markers (Hares, 2015). Although not in the core loci, several other highly polymorphic STRs (e.g., D2S1338 or SE33) are commonly found in STR kits. Normally, around 20-24 STRs can be found

1.2.4 Prospects of shift from length-based to sequence-based analyses

In the current state-of-the-art of the field, STR analyses are robust, fast and have a very high statistical power of discrimination. STR profiling is well-established and is the procedure that represents the mainstream approach for any forensic genetic test. In addition, the implementation of quality standards has been applied, including the accreditation of labs under

USC DE COMPOSTELA

ISO/IEC 17025 and the regular performance of inter-laboratory comparison exercises promoted by the International Society for Forensic Genetics (ISFG) and ISFG's working groups, which aim to assure the quality and reproducibility of profiling results.

Nonetheless, the rapid pace at which new technological advances in DNA sequencing are developed, the increase in accessibility of laboratories to new instruments may challenge the current methodologies and cause a paradigm shift in the field. There standardization of sequencing procedures for STRs, with the implementation of a minimal sequence range for the reporting of each STR's repeat region, and the creation of a standardized framework for naming for sequence-based alleles (Gettings et al., 2024; Hoogenboom et al., 2021) are examples of how this transition may be already in motion. In addition, the rise of new sequencing technologies offers a myriad of new applications that have broadened applications beyond STR analyses and enhanced the power of current STR sequencing methodologies (Ballard et al., 2020; Ge & Budowle, 2020). Lastly artificial intelligence is starting to be applied to complex profile analysis and is certain to play a crucial role in the future (Galante et al., 2023).

Despite this, in a field such as forensics, old habits die hard and the transition from the current typing methods to new technologies will probably be a slow process. If it is successful, both will probably co-exist for a long time, thus ensuring backwards compatibility with the data from current methodologies.

1.3 DNA MARKERS OF FORENSIC INTEREST

When we compare the DNA between to individual, up to 99.9% of DNA is estimated to be shared among them, as was estimated upon the completion of the Human Genome Project (International Human Genome Sequencing Consortium, 2004). In essence, from the forensic genetics perspective that DNA is rendered useless: we are not interested in what we have in common, but rather we want to find the parts of the DNA that are different and make each individual unique, allowing to, for instance, individualize persons, or attempt to predict phenotypical characteristics that can be used in an investigation. These differences in the DNA sequence are what are known as polymorphisms: the presence of different variants (alleles) in the DNA sequence on a specific region (locus). Additionally, the frequency of an alternative allele must be at least 1% in order to be considered a polymorphism.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Polymorphisms vary from changes in the DNA that solely affect a nucleotide to variations that span for many megabases, as is the case of Copy Number Variations. Essentially, we can classify polymorphisms into two different categories: length polymorphisms and sequence. Each type of polymorphism has its own distinctive characteristics and particularities and, consequently, they can be used in different scenarios depending on the information that we want to obtain or how much biological material we have at our disposal. Figure I1 displays four of the most common DNA polymorphisms in forensic genetics: STRs, indels, SNPs and MHs.

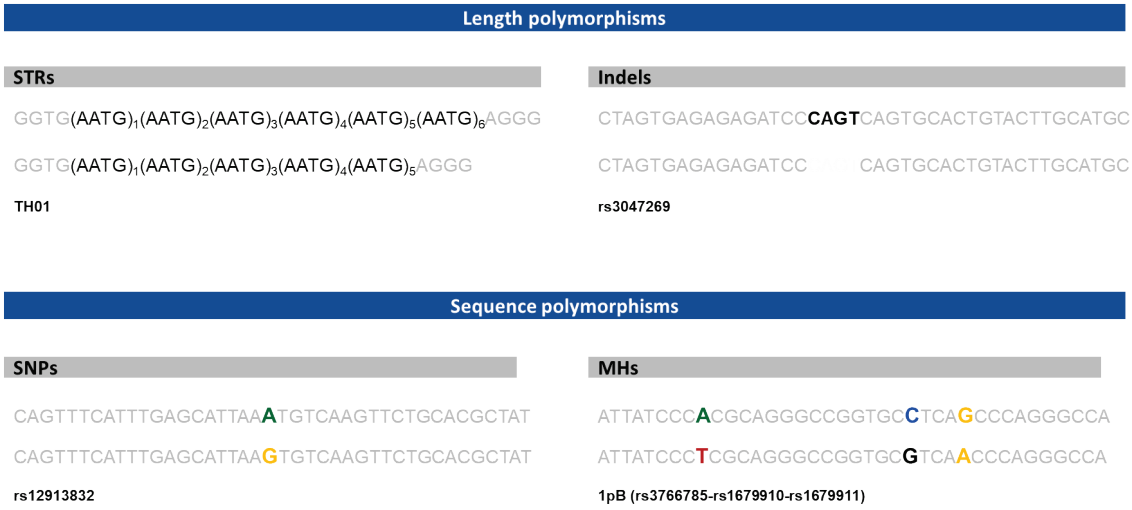


Figure I1. DNA polymorphisms of forensic interest

1.3.1 DNA polymorphisms

1.3.1.1 STRs

As previously mentioned, Short Tandem Repeats, also known as microsatellites, were one of the first markers that were implemented in routine forensic genetics analyses and they still represent, up to the present time, the gold standard in forensic genetic analyses. STRs are a type of length polymorphism which can be defined as a core of 2-7 bp that is repeated in tandem, to create the number of repeats that define the alleles (e.g., for the STR TH01, when the motif AATG is repeated seven times, the allele is referred to as a 7 allele). STRs are fairly abundant throughout the human genome and occur, on average, every 6-10 kb (Beckman & Weber, 1992). Their frequency is however inversely proportional to their complexity, and the longer the core sequence or the more varied are the structures of the repeat units that form the STR, the less abundant they are (Fan & Chu, 2007).

STRs are classified according to the length of the core sequence motif (most commonly: dinucleotide, trinucleotide, tetranucleotide, pentanucleotide repeat units) and depending on

whether the core repeat sequence is maintained intact or presents variations. Simple STRs comprise a unique core sequence that does not show variations, compound repeats are formed by two or more contiguous simple repeats, and complex STRs can have several core sequences of different length and different intervening sequences (Butler, 2005). The example mentioned earlier, TH01, constitutes a simple tetra-nucleotide STR, as its core is composed of four bases, and the core is invariable throughout the whole STR sequence. Sometimes STRs present incomplete allele units, producing what are termed non-consensus alleles (e.g., allele 9.3 in TH01, containing a TGA partial repeat). Mainly tetranucleotide STRs are used in forensic sciences given their level of polymorphism and stability, although some trinucleotide and pentanucleotide STRs are used in many forensic kits.

The key characteristic that has made STR markers ubiquitous in forensic analyses is their high level of polymorphism, as a single STR can present more than twenty different length alleles – a number which is further increased when including sequence variation within multiple same-size sets of repeats, especially in complex STRs such as D12S391 or D21S11. The combination of several STR loci markers creates "unique" profiles and can be a very powerful tool for identification or kinship analyses. The random match probability for just the 13 core CODIS loci (that is, the odds of two people sharing the same STR profile for those 13 markers) is lower than one in a trillion, when comparing unrelated individuals (Butler, 2005). The main reason behind this high level of polymorphism is their high mutation rate, which is around 10^{-3}-10^{-4}, several orders of magnitude higher than the average mutation rate in the human genome. The main mechanism for this is known to be the phenomenon of strand-slippage replication. Due to the repetitive nature of STR polymorphisms, sometimes during DNA replication a misalignment between the nascent and template strands occurs, and a loop of the mismatched repeat is created, giving rise to a copy with one more or one less repeat. If the mismatch repair system fails to correct this error, then the new strand will have a different allele length from the parental allele (Fan & Chu, 2007). The mutation rate however is not the same for all loci, but is dependent on several factors, including the length of the repeat unit, the sequence of the core or even the length of the allele. This phenomenon is also commonly observed *in vitro* when STRs are amplified in a PCR reaction, producing an artifact known as stutter (section 1.5.1).

1.3.1.2 SNPs

Single Nucleotide Polymorphisms (SNPs) represent the most simple type of sequence polymorphism, where only one base differs between individuals; yet they represent by far the most common type of variation, accounting for approximately ~85% of human genome variation (Goodwin et al., 2010). They can have up to four possible alleles, corresponding to the four different bases in DNA (A, C, T or G), but most SNPs are biallelic markers. Tri- and tetra-allelic SNPs are substantially less frequent, but they also have a significant presence in the genome, with more than 270,000 and 1,600 of them, respectively, in the 1000 Genomes Phase III data (Phillips et al., 2020). They have a homogeneous distribution in the genome and occur

every 300 bp. Their mutation rate is very low, especially when compared to STRs, and is estimated to be circa 2.3 x 10^{-8} (Nachman & Crowell, 2000). When the substitution does not change the class of nucleotide, it is called a transition; that is a pyrimidine (C or T) substitutes another pyrimidine or a purine (A or G) another purine, while the shift of a purine into a pyrimidine or vice versa is called a transversion. The mutation rate for transitions is however twice that for transversions and despite each base has one possible transition and two possible transversion (Zhang & Gerstein, 2003).

Their contrasting characteristics with STRs make them complimentary markers to this type of polymorphism, and depending on the application, their use can be beneficial and more suitable for some forensic analyses. Although STRs are not much more polymorphic, the discrimination power provided by limited numbers of STRs requires several more SNPs to be as informative, which also constitutes a limitation in mixtures analyses. Since STRs are a length polymorphism the analysis methods required for STR typing can be made in less time and effort (forensic marker genotyping methodologies are explained in more detail in section 1.5). On the other hand, SNPs are more abundant than STRs, they do not present stutter products, and they comprise a single nucleotide difference, so shorter amplicons are required for their amplification, which is especially advantageous for degraded DNA. The low mutation rate makes the nucleotide changes very stable throughout generations and it is more likely that two alleles come from the same mutation event than they have originated from two independent mutation events – which is an important characteristic to study complex kinship scenarios. It is also more likely that a SNP allele becomes fixed in a given population, which makes some of them useful for the study of an individual's ancestry.

The variety of SNPs with different characteristics in the genome makes them suitable for several different applications, depending on factors like their heterozygosity, their frequency distribution in different population groups, their correlation to externally visible characteristics (EVCs) or whether they are located in autosomal chromosomes or not. According to the type of application, SNPs are typically categorized into four main groups: identity informative (iiSNP), ancestry informative (aiSNP), phenotype informative (piSNP), and lineage informative SNPs (Budowle & van Daal, 2008).

1.3.1.3 Indels

Insertion-deletion polymorphisms, commonly referred to as indels, are, as the name suggests, length polymorphisms in which a sequence insertion or deletion is observed in a particular region of the genome. This variation in size can range from a single nucleotide to several hundreds, although the vast majority of them are short sequence differences of less than 16 bp (Weber et al., 2002). They represent the second most common type of variation in the genome after SNPs, occurring on average once every 1,500 bp and are estimated to represent ~15-18% of total genome variation (Mills et al., 2006; Mills et al., 2011).

USC DE COMPOSTELA

Indels represent an intermediate marker between STRs and SNPs as, despite sharing many characteristics with the SNPs, they also have some resemblance to STRs – even though they do not have the best of both worlds, unfortunately. They represent another type of length polymorphisms, so the typing technology used for them is almost identical to that used for STRs, with the advantage that their interpretation is simpler as they do not have stutter artifacts. The mutation rate for an indel is close to the values observed for SNPs, of around 10^{-9} (Nachman & Crowell, 2000), so they have two shared alleles which are more likely identical by descent (IBD) rather than identical by state . In the same way as SNPs, indels used in forensics generally benefit from short amplicon lengths, and are biallelic markers, so they have a low discrimination power when compared to STRs and are not very proficient in mixture analyses. Therefore, they tend to be employed as supplementary markers for kinship analyses (Fondevila et al., 2012), although they may be used for other applications, included for instance in panels for biogeographic ancestry (BGA) inference (Pereira et al., 2012) or the prediction of externally visual characteristics (EVCs) (Chaitanya et al., 2018) – applications further described in sections 1.5.3 and 1.5.4.

1.3.1.4 Microhaplotypes

Microhaplotypes (MHs) constitute a type of marker that have been recently adopted in forensic analyses. First proposed by Kidd et al in 2013 as a marker of potential forensic interest (Kidd et al., 2013), they are defined as a set of two or more SNPs closely sited in a region not larger than 200-300 bp that display, at least, three different haplotypes (allele combinations of the SNPs). Microhaplotypes benefit from the advantages of SNPs, such as short amplicons and no stutter products, while having a higher level of polymorphism given the different possible allele combinations of the composite SNPs.

The size limit established for MHs and their late implementation in forensic testing is not incidental, and both owe to the fact that they were conceived to be analysed through Massively Parallel Sequencing (MPS) technologies (see 1.5.2.2). The size limit designated for them corresponds to the size of reads that mainstream MPS sequencers used in forensic analysis can reliably sequence, allowing the whole marker to fit in a single read. This, in turn, allows the phase of the allele combinations on each sequence strand to be unambiguously phased, as phase cannot be determined when two or more SNPs are individually genotyped using, e.g., CE, and are heterozygous. If genotyped using Sanger sequencing or the SNPs are sequenced individually using technologies such as TaqMan or SNaPshot it is possible to determine the genotypes of the SNPs, but not how they are configured. Even though probabilistic phasing of the SNPs based on population data could be performed, rare haplotypes could potentially be lost.

Owing to their novelty and the increased accessibility to MPS sequencing, MHs represent a currently popular trend for marker set development in forensics, and more than a

USC DE COMPOSTELA

dozen panels have been presented in a very short time, generally focusing on the selection of highly polymorphic markers (Chen et al., 2018; de la Puente et al., 2020; Du et al., 2023; Gandotra et al., 2020; Kureshi et al., 2020; Oldoni et al., 2020; Pang et al., 2020; Qu et al., 2020; Staadig & Tillmar, 2021; Sun et al., 2020; Turchi et al., 2019; van der Gaag et al., 2018; Wang et al., 2023; Wu et al., 2021; Zhu et al., 2019; Zou et al., 2022).

1.3.2 Lineage markers

Lineage markers comprise the mitochondrial genome and the Y-chromosome, whose main difference from autosomal DNA relies on their transmission pattern, as they are inherited generally from just one copy of the mother's mitochondrial complement, or father's Y for male offspring with no alterations, so they are transmitted as a complete haplotype in each case. Their stability throughout generations has been subject to many studies to understand the history of human population migration and evolutionary (Cann et al., 1987; Jobling & Tyler-Smith, 2003; Vigilant et al., 1991) and today we have available phylogenetic trees and haplogroup classification for both markers (van Oven & Kayser, 2009; Y Chromosome Consortium, 2002). Their study in forensic genetics, despite not being markers that can be used to individualize, can also be very valuable in certain contexts.

1.3.2.1 Mitochondrial DNA

The human mitochondrial genome consist in a 16,569 bp circular DNA molecule that is located in the mitochondrial matrix in multiple copies, ranging from one to ten per mitochondrion (Satoh & Kuroiwa, 1991), and depending on the cell type the total number of mitochondria and therefore DNA copies in a single human cell can vary from 100 to more than 10,000 (Wai et al., 2010). Despite its relatively small size, the mitochondrial DNA (mtDNA) encodes for a total of 13 proteins, 22 transfer RNAs and two ribosomal RNAs (12S and 16S) (Goodwin et al., 2010), that are divided into the heavy (H) and light (L) chains.

The most accepted theory for the origin of mitochondria (and the reason why this organelle has its own genome) is the endosymbiotic theory postulated by Lynn Margulis. According to this theory, both the mitochondria and chloroplasts were prokaryotic organisms that were "captured" by the ancestors of eukaryotic cells and adapted to become the organelles that we know today (Sagan, 1967). The sequence of the complete mitochondrial genome was the first part of the human genome to be completed almost a decade prior to the beginning of the Human Genome Project (Anderson et al., 1981), and was revised a few years later to amend several sequence annotation errors (Andrews et al., 1999), giving rise to the revised Cambridge Reference Sequence (rCRS), the reference genome for mitochondrial analysis that is used currently.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

The mitochondrial genome mutation rate in comparison to the nuclear genome is much higher (around 100-fold) due to oxidative damage (Hashiguchi et al., 2004), most of the variation found concentrates in a small non-coding region termed the control region (Greenberg et al., 1983). The length of this region is around 1.1 kb and occupies the positions 16024-576 in the rCRS, in which two hypervariable sites HVS-I (16024-16365) and HVS-II (73-340) are located. A third hypervariable region, HVS-III (438 and 574) is also used. Forensic analysis focuses on the sequence analysis on these regions, as it is where most of the variability is found (which, again, is where the interest of forensic analyses lies) and this analysis establishes an individual's haplogroup. Traditionally, the study of mtDNA has been performed by Sanger sequencing, establishing the differences with the rCRS, but the technological developments of MPS has allowed the the exploration of the whole mitochondrial genome (see 1.5.2.2).

The study of mtDNA has provided valuable insights in the study human evolution (Cann et al., 1987) and for the study of mitochondrial diseases (Zardoya, 2020), but for forensic geneticists it provides a powerful tool for the analysis of maternal lineages. The main advantage of mtDNA analysis versus nuclear DNA resides in the large number of copies that are present in each cell, which can be used to analyse hair or very degraded samples. mtDNA does not undergo the traditional recombination mechanisms from nuclear DNA and it is transmitted from mother to children – excepting very rare occurrences that could be related to maternal diseases (McWilliams & Suomalainen, 2019). Sometimes, individuals may present more than one type of mitochondrial DNA, in the phenomenon known as heteroplasmy, which normally consist of a difference in just a single position (Goodwin et al., 2010). Outside human forensic genetics, the analysis of more conserved regions of mtDNA, like the cytochrome *b* and the 12S and 16S RNAs, can be used to determine the source of non-human DNA (Pereira et al., 2006).

EMPOP[1] is a curated database of mitochondrial variation that represents a very useful tool in the forensic context, hosting data for a total of 48,572 mitochondrial haplotypes (release 13, 2019). It allows the assignment of haplogroups based on sequencing data (Huber et al., 2018) and can provide the haplotype frequency estimates along with their global distribution. Thus, the study of mtDNA can also be used to establish an individual's maternal ancestry – although the study of this marker alone may not be representative of one's ancestry, for instance, when a distant atypical mtDNA is inherited (Phillips, 2015).

A well-known historical example of the application of mitochondrial DNA were the analyses performed to confirm the identity of what were thought to be the remains of the last Tsar and Tsarina of the Romanov Royal Family, found together with three of their children in a shallow grave near Yekaterinburg, Russia (Gill et al., 1994). Since no close relatives were alive, the comparison was done using mtDNA from a relative from the same maternal lineage: Prince Philip, the Duke of Edinburgh. The Duke of Edinburgh was asked for a DNA sample, as his maternal grandmother, Princess Victoria of Hesse and by Rhine, was the sister of Tsarina

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[1] https://empop.online/

Alexandra. The results indicated that the remains found in the Ipatiev House (now the location of the Church on Blood in Honour of All Saints Resplendent in the Russian Land) shared their mitochondrial lineage with Prince Philip, and STR analysis also established that the remains found belonged to the same family group. The mtDNA of Tsar Nicholas on the other hand, as it belonged to a different maternal lineage had to be compared with two living relatives, but due to scepticism caused by the presence of a heteroplasmic position, it was posteriorly confirmed with the analysis of the remains of his brother Georgij Romanov, who shared the same mitochondrial genotype (Ivanov et al., 1996). The same mitochondrial evidence was used to disprove Anna Anderson's allegations, who claimed to be Anastasia, one of the Tsar's daughters that had escaped her fate; and served to identify the two other missing children that had not been discovered, but were found in 2007 in a mass grave near the original grave (Coble et al., 2009).

1.3.2.2 The Y chromosome

The Y chromosome is one of the sex chromosome pair that determines the biological sex in humans and most mammals. Its size is much smaller than its X counterpart, roughly 1/3 of its size, and just ~10 Mb larger than chromosome 21 (the smallest autosomal chromosome). The pair of sex chromosomes present in males do not undergo recombination events except in the two Pseudo Autosomal Regions (PAR), so ~95% of the chromosome has no recombination and is transmitted as a haplotype block from fathers to sons (Goodwin et al., 2010) – except the occurrence of some ectopic crossing over between chromosomes, that generally leads to infertility. The non-recombining part, also referred to as the male-specific Y, only contains 78 protein-coding genes, much less than the 800 protein-coding genes on the X chromosome (Skaletsky et al., 2003). The two sexual chromosomes originated from the same pair of ancestral autosomes. The complex evolution process underwent by the proto-sex chromosomes, which occurred around 240 million years ago, and was catalysed by the appearance of the sex-determining region Y (*SRY*) gene. Both chromosomes progressively lost the capacity to recombine through several events of chromosomal inversions, eventually causing the degeneration of the Y chromosome, which suffered a decay in the number of functional genes and led to the loss of genetic material (Bachtrog, 2013).

Although, as previously stated, lineage markers cannot individualize, they can outperform autosomal markers for several applications. The most common application is the identification of a male contributor in mixtures with very skewed male to female ratio, circumstances that are unfortunately often observed in forensic laboratories when analysing sexual assault cases. The key characteristic of transmitting entirely to male descendants allows Y markers to be used for the study of patrilineal lineage in the same way as mitochondrial DNA for maternal lineage, which can be applied for familiar searching or distant kinship analyses. Continuing with monarchic examples, Y-STR data was recently used to identify the remains of King Richard III, the last English King of the House of York. The King met his fate in the Battle of Bosworth Field during the Wars of the Roses, and his remains were buried in the Greyfriars friary, in the middle of Leicester. The King's resting place was lost to time, until researchers from the

University of Leicester located the presumed burial location. The haplotypes from male-line descendants were used to identify the male lineage of the late King. Unfortunately, the remains did not match, possibly due to an extra-marital paternity event, so they were identified using the maternal lineage, by analysing mtDNA.

Similarly, to EMPOP, the forensic database YHRD[2] (Y Chromosome Haplotype Reference Database) serves as a repository of a total of 349,750 Y-STR haplotypes at release 69. It provides several useful tools for forensics, such as the possibility to obtain haplotype frequency estimates.

1.3.3 The X chromosome

The X chromosome represents the other sexual chromosome that forms part of the sex-determination system present in humans and most mammals. It has around 153 Mb of sequence and contains ~1100 genes, which makes it one of the chromosomes with the lowest gene density, while also the average sizes of its genes are relatively small, with 1.7% of the sequence representing exons of known genes (Ross et al., 2005). Unlike its counterpart, that is only present in a single copy in males and mainly specializes in sex-determination, the X-chromosome contains several other genes that are involved in female individuals who carry two copies of the chromosome. To compensate for the different gene dosage between the two sexes one of the two copies undergoes a silencing process regulated by a long non-coding RNA encoded in the *XIST* (X-inactive specific transcript) gene (Borsani et al., 1991) and its antagonist *TSIX (Lee et al., 1999)*, forming what are known as Barr bodies (Boumil & Lee, 2001). This inactivation occurs in the early stages of development, and is believed to occur randomly at one of the two chromosomes, and which remains inactivated.

These differences between males and females also play a crucial role in its transmission pattern: female individuals transmit one copy to their children, while male individuals are hemizygous and transmit the single X chromosome they carry to their daughters. (Fig. X) In addition, maternal X chromosomes can recombine during meiosis, while this is not the case between the X and Y chromosomes in males (excepting the PAR regions). Another peculiarity is that the recombination in the X chromosome is not even throughout the whole chromosome, and traditionally there are four linkage groups used for the X-STRs used in forensics, as regions in which the recombination rate is very low.

These unique characteristics make the data from X-chromosomal markers more challenging to interpret and the statistical analysis more complex, which usually leads to use of the autosomal information instead. However, the use of X markers such as X-STRs can provide strong statistical support in instances such as sisterhood tests, in which both individuals must

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[2] https://yhrd.org/

share a pair of identical by descent (IBD) X-alleles, or provide evidence in cases that would be undistinguishable if autosomal markers were used, which can occur in some incest scenarios (Gomes et al., 2020).

1.4 OTHER GENETIC MARKERS OF FORENSIC INTEREST

1.4.1 Epigenetic markers

DNA carries the information required for protein synthesis encoded in its nucleotide sequence. Yet, the regulation of gene expression cannot be explained by the variation of the sequence itself, as, for instance, different cell types with the same genetic information can present very diverse phenotypes. Epigenetic markers are defined as these underlying regulatory mechanisms that can modify gene expression without altering its nucleotide sequence, representing dynamic hereditable changes that can be influenced by genetic or environmental factors. There are different levels of epigenetic modifications, although they can be classified in three categories: histone modification (e.g., acetylation or phosphorylation), non-coding RNAs or DNA methylation (Bird, 2007).

The main epigenetic modification that is studied in this tool is DNA methylation, which consist on the addition of a methyl group ($-CH_3$) to the 5'C from C thar are followed by a G (known as CpG dinucleotides). Numerous and very diverse applications based on the use of epigenetic marks have been developed so far for their use as forensic tools, from mainstream techniques to estimate the chronological age (more detailed in section 1.5.5) to more niche applications such as the differentiation of *in vitro* generated DNA based on its epigenetic marks (Frumkin et al., 2010). In general, nonetheless, three main uses of epigenetics tools for their use in the forensic context can be distinguished: as previously mention, the prediction of an individual's age, to discern between two monozygotic twins (like the Lohan twins in *The parent trap*) that would otherwise not be possible to tell apart traditional autosomal markers and the determination of body fluids and tissues (Vidaki & Kayser, 2018), similarly to the approach using RNA (see next chapter).

1.4.2 RNA

Ribonucleic acid (RNA), despite its lower stability compared to DNA, has applicability as a forensic marker is a valuable tool for different applications. Unlike DNA, in which the main source of interest are the polymorphisms to distinguish, RNA analyses focus on the study of the expression patterns of RNA – mostly by converting RNA into DNA using RT-PCR (reverse transcription polymerase chain reaction).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Most of the variation found in gene expression is due to differences among tissues, while a smaller percentage is due to differences among individuals (Mele et al., 2015), so one of the main applications of RNA testing is identifying the biological source of a sample. Several RT-PCR assays have been developed for this purpose, comprising markers for the identification of body fluids and tissues of forensic relevance such as blood, semen, saliva, menstrual blood, skin or mucosa. (Fleming & Harbison, 2010; Lindenbergh et al., 2012). Other uses of RNA include the estimation of the post-mortem interval, determination of the time since deposition, wound age determination, age prediction, or the assignment of tissues to a specific individual in mixtures using coding region SNPs (Bauer, 2007; Haas et al., 2021). Some recent studies have also evaluated the applicability of RNA analysis to establish the time of deposition of a biological sample, that is, the moment of the day in which a sample has been produced, by mRNA analysis, studying the expression of genes implicated in the circadian rhythm or performing whole transcriptome analysis (Girón Santamaría, 2021; Gosch et al., 2023).

1.5 Analysis methods used for forensic testing

STRs analysis constitutes the main pilar in any forensic genetic investigation, provided it represents a powerful and sensitive tool that has enough statistical power for the resolution of most common situations found in routine analyses. However, some scenarios like the lack of a suspect profile or matches in police databases, the deconvolution of complex mixtures or statistical analyses of distant kinship cases are examples of recurring DNA analysis challenges and constitute some of the current challenges faced in forensic analyses (Fondevila et al., 2008; Gonzalez-Bao et al., 2025; Kayser, 2015).

In an effort to overcome these limitations, the forensic community has worked towards the construction of a wide variety of tools that go beyond the use of STRs. Large SNPs panels for instance can provide enough discriminatory power to provide conclusive results when distant relationships are being tested. When a dead end is reached in an investigation, there are tools that can provide information from a DNA donor which can help narrow down the pool of suspects. Among them, we find those termed as “DNA intelligence” tools, which can be used for the prediction of an individual’s ancestry or physical traits; or tools that can be used to infer the chronological age of the suspect. Examples of traits that can be inferred (or currently under development) are illustrated in Figure I2 and detailed in their respective sections.

The use of these tools, however, requires the understanding of their limitations and the acknowledgement of their prediction error (Boullon-Cassau et al., 2025). In addition, while the ethical concerns are beyond the scope of this Thesis, these tools must be used for their intended purpose, as they could be misused, for instance, as instruments of discrimination against certain population groups (as exemplified in *The Problem We All Live With* by Norman Rockwell).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

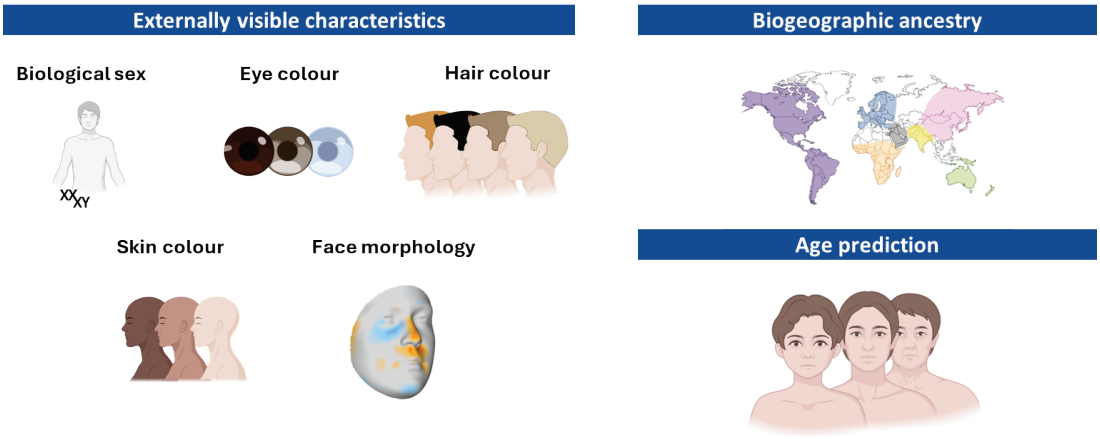


Figure I2. Illustration of different traits that could potentially be inferred from an unknown DNA donor. Created with Biorender.

1.5.1 STR profiling

The standard analyses routinely performed in forensic laboratories for STR profiling rely on the separation of the different alleles according to their size (which can be extended to indels as well). As previously mentioned, the availability of DNA sample is a key factor for most forensic procedures, which is why most of the standard procedures rely on the use of the PCR or polymerase chain reaction (Mullis et al., 1986), an enzymatic process that can be used to generate multiple copies from minimal amounts of DNA. The PCR is enzymatic process based on the use of a thermally stable polymerase (such as the polymerase I from *Thermus aquaticus* (Chien et al., 1976) (*Taq* polymerase)) which, throughout a cyclical thermal process produces the (almost) exponential amplification of DNA. In short, in each cycle DNA is denatured by heating the sample to ~96-98 ºC. Then, the temperature is lowered to ~55 ºC, to allow the annealing of the primers specifically designed to hybridise in the regions we want to amplify. Posteriorly, the temperature is set to the optimal temperature for the polymerase activity (~72 ºC for the *Taq* polymerase), initiating the synthesis of the complementary strand. After the PCR amplification, the resulting DNA amplicons are subjected to a capillary electrophoresis (CE) procedure, where DNA fragments are separated according to their size in base pairs (bp). DNA molecules, which are negatively charged, are exposed to an electric current and forced to migrate through small capillaries containing a polymeric gel. Depending on the size of the DNA the migration will vary, and larger molecules will progress at a slower pace through the pores.

When analysing STRs, several loci are simultaneously typed in a single multiplex reaction (that is, several primer pairs are used in the same reaction) and primers are typically labelled with different fluorophores, which can be used to separate not only according to their size, but also depending on the fluorescent signal emitted by the label introduced. However, several

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

factors may affect the success of the amplification, such as the degradation of the DNA or the presence of inhibitors in the DNA sample like hematin, carmine or indigo blue dye.

1.5.2 DNA Sequencing

Sequencing methodologies, which are the techniques used to establish the nucleotide sequence in DNA or RNA molecules, have evolved significantly over time. The strategies used for sequencing are very diverse, and over time have undergone substantial improvements from more rudimentary and low-throughput technologies to the high-throughput sophisticated techniques that are used nowadays.

1.5.2.1 First generation sequencing

Nearly half a century ago, the earliest approaches for DNA sequencing started to emerge. Some of the first methodologies were the 'plus and minus' system by Coulsoun and Sanger (Sanger & Coulson, 1975), used for sequencing of the bacteriophage phi X174 (the first DNA genome ever sequenced) (Sanger et al., 1977a); or Maxam-Guilbert sequencing, based on the inference of the nucleotide sequence by the use of different chemical treatments that produced the cleavage on certain key points (Maxam & Gilbert, 1977). However, the most influential sequencing method, still used up to date, was Sanger sequencing (Sanger et al., 1977b) which was the method chosen for the generation of the first draft of the human genome (Lander et al., 2001).

The method developed by Sanger is based on the use of labelled dideoxynucleotides triphosphate (ddNTPs), that are mixed (in a smaller ratio) with regular deoxynucleotides triphosphate (dNTPs). The process consists on an amplification similar to the PCR, in which dNTPs are added sequentially into the nascent DNA strand until a ddNTP is incorporated. Due to the lack of a hydroxyl end, the DNA polymerase cannot continue adding new nucleotides and the elongation process is interrupted. The products of the sequencing reaction are subjected to an electrophoresis to separate the fragments according to their size. The original method employed ^{32}P labelled ddNTPs, requiring four parallel sequencing reactions and electrophoresis. Modern day Sanger sequencing, however, makes use of fluorescently labelled ddNTPs, each of which incorporate a different fluorescent, reducing the number of reactions performed to one per sample; and automated sequencers, which perform the separation using CE and automatic detection of the labelled fragments with a laser.

To date, Sanger sequencing hasn't gone out of style and is still a widely used technique and far from being substituted by modern technologies, due to the low cost and rapid analysis. In forensic analyses, however, this technique is mainly reserved for mtDNA analysis. While SNP typing is also possible through Sanger sequencing, this analysis is highly inconvenient as it is restricted to a single locus per reaction. In such instances, Single Base Extension (SBE) sequencing methodologies, also known as mini-sequencing (Sokolov, 1990), constitute a

sensitive and affordable alternative, that allow the multiplexing of up to 30-40 SNPs in a single reaction. SBE methods (normally referred by the most common commercial kit for SBE analysis, SNaPshot™) consist in an approach using labelled ddNTPs, similarly Sanger sequencing. After a multiplex PCR amplification (and a subsequent purification to remove unused nucleotides and primers), the SBE reaction is carried out, in which the probes, one for each of the target locus, hybridize a nucleotide upstream of the location of the SNP. Then, a single labelled ddNTP is incorporated (hindering the addition of extra nucleotides) and the base incorporated is detected via CE. This method is mainly used in forensics for different small-scale panels for identification panels (Musgrave-Brown et al., 2007), DNA phenotyping (Fondevila et al., 2013; Phillips et al., 2012b), ancestry inference (Mengel-From et al., 2010; Ruiz et al., 2013; Walsh et al., 2011) or even for the analysis of methylated DNA for age estimation (Ambroa-Conde et al., 2022; Jung et al., 2019); the last three applications explained in more detail in sections 1.5.3, 1.5.4 and 1.4.5, respectively.

1.5.2.2 Next Generation Sequencing

The terms Next Generation Sequencing (NGS) and Massively Parallel Sequencing designate a set of different sequencing approaches that represent a shift from first-generation sequencing and target the simultaneous sequencing of up to millions of DNA molecules. The chemistry used by most of the MPS platforms rely on a sequencing-by-synthesis approach, in which the detection of the nucleotide sequence is based on the detection of the polymerase activity when synthesises the complementary strand of our target DNA molecule. Pyrosequencing was one of the first detection techniques used in MPS, that was implemented in the 454 Life Sciences GS20 platform, introduced in 2005, and was the first commercial MPS platform that became broadly available and relatively affordable. This method is based on the use of the firefly luciferin enzyme that, in combination with an ATP sulfurylase and an apyrase transform the PP_i subproduct that is released by the DNA polymerase during DNA synthesis into a light pulse (Nyren et al., 1993) that was adapted for MPS analyses (Margulies et al., 2005). Nonetheless, the chemistry used by the sequencing platforms that dominate the MPS analyses in forensics field, the Ion S5 from Thermo Fisher Scientific (TFS) and the MiSeq FGx by Verogen/Qiagen, use a different approach.

The Ion Torrent sequencers, including the Ion S5 and its predecessors (Ion PGM and the Ion Proton) based their sequencing technology on the use of semiconductors that detect pH variation produced by the liberation of protons during DNA synthesis (Rothberg et al., 2011). In simple terms, each sequencing reaction occurs in a well in which a semiconductor that acts as a pH-meter is located, which detect the H^+ that is released when the polymerase adds a new nucleotide into the nascent chain. The nucleotides do not differ from traditional dNTPs, but the key for the analysis lies in the fact that they are added in a one-by-one basis in a sequential way. In each step of the process, one type of dNTP is added together with the rest of the reagents required to complete sequencing reaction. Wells in which the pH variation is detected means

that the base has been incorporated and thus the next in the sequence and if multiple dNTPs are incorporated, the signal recorded by the semiconductors will be proportional according to it. Then, nucleotides not incorporated are washed away and a different dNTP is used.

The chemistry utilized by the Illumina platforms, the MiSeq and the HiSeq platforms present a detection method that employs nucleotides labelled with removable fluorophores – each nucleotide with a different dye (Bentley et al., 2008). In each cycle of the sequencing process, all four nucleotides are added, and since the 3'OH end is blocked, only one can be added to the nascent strand. The fluorophore is then excited with a laser, the signal emitted is recorded and then the dyes are cleaved, liberating the 3' hydroxyl group for the next cycle. This chemistry was also used by Qiagen's MPS platform, the GeneReader (now discontinued).

While both platforms offer a good performance, the error rates recorded for the Ion S5 are higher than for the MiSeq (≥ 1% vs 0.01%), mainly comprising insertions and deletions as the most common error found in this platform – especially for homopolymer stretches. On the MiSeq on the other hand the most common errors are base substitutions and, despite its higher accuracy, presents longer runs and does not offer automatic library preparation and chip loading. Overall, and despite these differences, both platforms (and NGS analysis in general) offer several advantages over CE typing; including (i) the capability to simultaneously analyse several types of markers, (ii) increased STR informativeness thanks to capability to detect isoalleles (alleles that share the same length in CE but differ in their sequence), which in turn could aid in mixture analysis, (iii) a larger pool of markers can be analysed using the same input DNA; (iv) the possibility to analyse other types of markers, such as MHs; or (v) that the amplicon size can be reduced to the minimum, as size separation of the alleles is not performed (Ballard et al., 2020).

So, it is important to ask why MPS analyses are not the main technology used nowadays in routine forensic analyses? For most analysis carried out routinely, STRs are informative enough. MPS would require increased analysis time, both hands-on and sequencing time, and increase the cost per sample. Additionally, the lack of scalability can also pose a problem for small- and medium-scale laboratories; as a large pool of samples would be required for each run to make the analyses more affordable, and the investment for the transition from CE to MPS might also constitute a deterrent, as the initial investment to acquire an MPS platform can be high. Furthermore, the transition from CE to MPS involves a substantial initial investment, as acquiring an MPS platform can be expensive. For this reason, most of MPS technologies are the exception rather than the norm and are reserved for the occasions when STR resolution does not provide conclusive results.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.5.2.2.1 MPS sequencing workflow

The first step in an MPS workflow is usually the library preparation, i.e., preparing the DNA sample for the sequencing process. The strategies that can be used for MPS analyses are different and depend on the type of analysis that is pursued. When a species genome is being sequenced *de novo* or for the identification of a disease-causing mutation whole genome sequencing (WGS) may constitute the most suitable approach. In forensic applications however WGS is generally not the most indicated methodology, as the initial quantity and quality of DNA are high; and also entails a large amount of resources, which increase the cost of the analyses. Targeted sequencing approaches concentrate instead on the typing of a just the ROIs that are selected, using enrichment procedures to isolate the desired regions, so a greater depth of coverage is obtained for those regions compared to WGS and a more straightforward data analysis is required. Two main strategies for targeted sequencing exist, namely PCR amplification and hybridization capture target enrichment. Target enrichment PCR consists in a multiplex PCR to amplify those ROI, usually utilizing primers modifications to permit higher multiplexing; whereas hybrid capture relies on a DNA fragmentation followed by a, isolation of these ROIs using probes that are complementary to them, often labelled with biotin for a biotin-streptavidin separation – which is posteriorly amplified via PCR to increase its concentration (Mertes et al., 2011; Tucker et al., 2009).

During library preparation adapter sequences are added to the amplicons that will be required for the sequencing process. Sometimes, the barcoding of the libraries is also performed, in which a different barcode sequence (i.e., a DNA tag, sometimes also referred to as index) is added to the libraries. This allows to combine in the same run several samples, as in posterior bioinformatic analyses these tags can be recognised and reads can be de-multiplexed and be assigned to its corresponding sample. UMIs (Unique Molecular Indices) are another type of tagging sequence that is incorporated to DNA or RNA molecules prior to PCR amplification. During downstream analyses, reads bearing the same UMI can be grouped, creating a single consensus read (UMI read) which allows to remove PCR or sequencing errors. In addition to improving the accuracy of the results, the use of UMIs serves other purposes, such as accurately counting the number of molecules that were originally present in the sample (e.g., in RNA-sequencing) (Islam et al., 2014) or detecting low-frequency variants, such as analysing circulating tumour DNA (Kinde et al., 2011; Newman et al., 2016).

After library preparation and before sequencing process MPS platforms require, as previously mentioned, the clonal amplification of the libraries so that the signal emitted can be detected by the sensors. Yet again, several strategies are used for this step, and usually each commercial company uses a different approach: while TFS sequencers use a clonal amplification based on an emulsion PCR, Illumina platforms employ a bride PCR amplification. In emulsion PCR, DNA molecules from the libraries are captured in aqueous micelles which contain a bead (termed Ion Sphere™ particles in TFS instruments) together with the rest of the reagents required for the PCR reaction (ddNTPs, polymerase and primers). The bead is covered with adapters that are complementary to the ones added to the libraries and in

USC

the PCR process, which allows the copying of the library fragment all over the surface of the bead. Bride amplification, on the other hand, is carried out on the surface of the flow cell in which the sequencing is to be carried out. The flow cell is covered with both forward and reverse complementary primers for the adapters. Library templates bind first to one of the adapters, generating a copy of the library on the surface. Then, during the amplification, the free end can interact with the primers located nearby (forming a bridge structure when hybridising) and generate another copy of the library, which eventually leads to the creation of clusters each from the same original library molecule (Goodwin et al., 2016). Figure I3 illustrates the described sequencing workflow of an MPS sequencer.

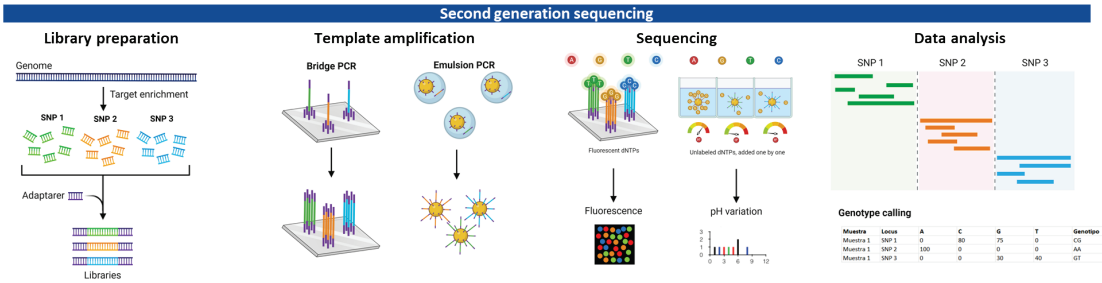


Figure I3. Schematic representation of an MPS workflow for a targeted sequencing approach. Created with Biorender.

1.5.2.3 Third-generation sequencing

Third-generation sequencing (TGS) is a term that has been used in the recent years to design long-read sequencing technologies that, as in second-generation chemistries, encompasses different sequencing chemistries – being the sequencers from Pacific Biosciences and Oxford Nanopore Technologies (ONT) the most representatives so far from this generation. An additional feature of this generation aside from the increased read length is that TGS platforms can perform real-time PCR-free sequencing. Nonetheless, as stated by Plesivkova et al., the absence of clonal amplification can cause a drastic reduction in coverage and a subsequent decrease in accuracy (Plesivkova et al., 2019), which could not be suitable for many forensic samples.

Unlike the PacBio sequencing chemistry, that also relies on a sequencing-by-synthesis methodology, the unique approach by ONT uses a distinctive nanopore-based detection strategy. In this approach, DNA, which is turned into single strand, is translocated through the nanopores that are embedded within a synthetic membrane. This causes characteristic disruptions in electrical current that are used to determine the sequence of the passing molecule (Figure I4). The MinION, which constitutes the smallest ONT platform, stands out as a highly portable, pocket-sized sequencer. Its compact design and versatility make

it ideal for ad hoc laboratory setups, enabling scalable solutions for the adaptation to specific laboratory necessities, while also constituting one of the most accessible sequencing devices available.

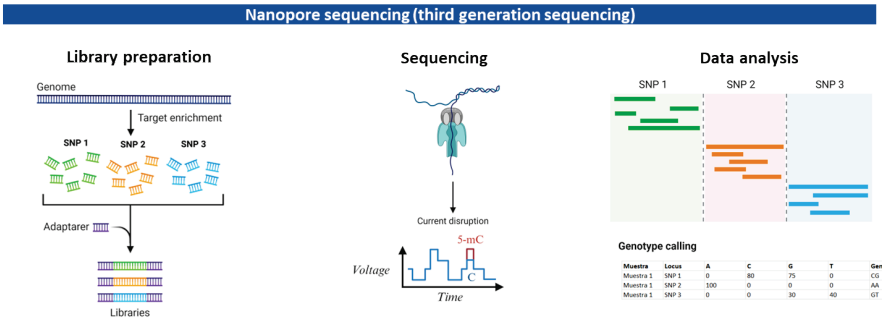


Figure I4. Schematic representation of an MPS workflow of a targeted sequencing approach. Created with Biorender

The main handicap of ONT sequencers is the high error rate compared to short-read technologies. Upon release, the accuracy was lower than 60%, but the improvements in sequencing chemistry and base calling algorithms have led to significant improvements (Rang et al., 2018). The recently launched R10.4.1 flowcells offer a modal accuracy higher than 99% – comparable with the 99% and >99.9% accuracy that the Ion S5 and MiSeq offer, respectively. In addition, the versatility of nanopore sequencing does not only allow for just standard DNA sequencing, but its applicability could be extended to analyse multiple types of biomolecules including RNA, DNA modifications such as methylation analysis (Logsdon et al., 2020) or even modifications on protein structures (Martin-Baniandres et al., 2023). The use of these platforms in forensic genetics is still scarce, but several studies have already begun to adapt already-existing forensic panels for ONT sequencing (Casanova-Adan et al., 2023; Sapan et al., 2024).

1.5.3 Forensic DNA phenotyping

Forensic DNA phenotyping (FDP) or simply DNA phenotyping is known as the methodologies employed to infer EVCs from an unknown DNA donor, acting like a sort of "biological witness" (Kayser, 2015). The first and most simple EVC that could be determined using DNA markers was the biological sex. Amelogenin is the primary protein found in the extracellular matrix in the developing tooth bud, responsible for proper enamel development (Termine et al., 1980). The homologous amelogenin genes located in the sex chromosomes present a 6 bp indel, allowing the detection of male DNA (Mannucci et al., 1994). The

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

amelogenin indel is a common marker present in most STR kits nowadays, although it is habitual to find additional markers for sex determination, e.g., Y-STRs or other indels.

Other EVCs, unfortunately, are not so easily determined as they are more complex and cannot be defined in discrete categories, but they rather represent a spectrum showing a gradient of several states. From the genetic point of view, they are also challenging, as several genes intervene in these characteristics, and they are also influenced by additional elements including environmental factors or the individual's age. Some of the most common and well-established phenotypic characteristics, aside from the biological sex, include the pigmentation traits, most notably eye, hair and skin colour, as the number of genes involved in this is relatively small and involved in the melanogenesis (melanin biosynthesis).

The differences in iris pigmentation rely on different factors like the melanin type (eumelanin or pheomelanin), quantity and distribution (D'Mello et al., 2016). The *HERC2-OCA2* region contains most of the SNPs with association with eye colour variation, including rs12913832, a key SNP for the eye colour prediction that shows strongest association given its role in the regulation of *OCA2* transcription (Visser et al., 2012). The eye colour was the first one to be addressed for forensic purposes and implemented into a SNaPshot multiplex, encompassing 6 SNPs associated with iris pigmentation (Walsh et al., 2011). The approach used for the classification was choosing extreme values (brown and blue) and another category for intermediate phenotypes. Further studies have incorporated number of markers to increase the accuracy of the prediction of intermediate (green-hazel) phenotype including several HERC2 SNPs (Freire-Aradas et al., 2014; Ruiz et al., 2013). Despite the classification of extreme values show high accuracy, intermediate phenotypes still pose a challenge and a better understanding of the molecular basis of eye formation may be key for enhancing the prediction models (Brancato et al., 2023).

Hair pigmentation represented the next step in EVC prediction, as it posed a greater challenge given the greater number of categories and genes involved. The red hair was the first phenotype that could be predicted with high accuracy owing to the discovery of several variants with high correlation of near the *MC1R* gene (Grimes et al., 2001). The HIrisplex system, nonetheless, was the first forensic panel that targeted the whole spectrum of hair colours and their shade by analysing 24 SNPs, including as well the 6 SNPs comprised in the previously developed Irisplex SNP set (some of which were also used in the model for hair colour prediction). Following a similar approach as before, four hair colour categories were defined (blonde, red, black or brown) (Walsh et al., 2013) and, as with the intermediate eye colour, the differentiation between certain categories (blonde and brown) proved to be more challenging. A subsequent study by Söchtig and collaborators pursued as well the identification of variants associated with hair colour (Sochtig et al., 2015), successfully identifying a more compact set of 12 SNPs, partially overlapping with the HIrisplex markers. Similarly to what was observed with the HIrisplex study, this exploration highlighted once again the difficulty of the differentiation of brown colour.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Additional studies have also explored the prediction of skin colour. Unlike the eye and hair colour mentioned previously, whose variability is mainly restricted to the European population, skin colour presents a global variation. So, the identification of markers for skin colour prediction had the added challenge of identifying markers that were not biased due to the population. The work by Maroñas et al. was the first forensic assay presented for skin colour analysis and prediction, comprising a total of 6 SNPs (Maronas et al., 2014). A posterior study by Chaitanya et al. presented a new version of the HIrisplex system (termed HIrisplex-S) that complemented the previous SNPs 41 SNPs for the simultaneous prediction of hair, eye and skin colour (Chaitanya et al., 2018). Both tools demonstrated high accuracy, especially when accounting for the extreme values, although the model by Maroñas et al. comprised 3 categories (white, intermediate and black) whereas the model developed for the HIrisplex-S markers contemplated five categories (very pale, pale, intermediate, dark and very dark).

Beyond pigmentation studies, several other markers for EVC have been identified thanks to genome-wide association studies (GWAS). For instance, in the literature there are numerous studies targeting height prediction (Liu et al., 2014), hair morphology (Pospiech et al., 2015) or male-pattern baldness (Liu et al., 2016) prediction, with varying degree of success. Face morphology, despite its high level of complexity, may also be feasible in the near future or which several loci have been already identified (White et al., 2021).

1.5.4 Bio-geographic ancestry

1.5.4.1 The structure of human population

The worldwide human population is a complex entity that cannot be represented by a homogeneous group of individuals nor a collection of isolated populations. Instead, it reflects the result of a long and intricate evolutionary process, with influencing factors including: adaptation to selective forces, both positive and negative; Hominin introgression (both Neanderthal and the divergent Denisovan-Neanderthal); migration; trade; conflict and conquest; and on a surprisingly large scale, slavery. All these factors have shaped human global population and regional sub-population structure found today, and this is reflected in the distribution of genome variability which accounts for the very small percentage of differences between population groups (see below) - exploited in forensic tests for BGA.

Over the years, numerous population studies have sought to elucidate the structure of human populations. The first study dates back to 1972, when Lewontin estimated that the differences observed within a population accounted for ~85% of the variation, while the differences between groups and population groups were responsible for the remaining 10% and 5%, respectively. (Lewontin, 1972). Given the technical limitations of the time however, this study was performed with a limited number of autosomal markers. A posterior study by Rosenberg et al. in 2002 (Rosenberg et al., 2002) analysed a 377 STR collection to measure the

USC DE SANTIAGO DE COMPOSTELA

population structure on the 1056 samples from 52 different populations comprised in the Human Genome Diversity Project-Centre Etude Polymorphism Humain sample set (HGDP-CEPH) (Cann et al., 2002). The analyses using the STRUCTURE clustering algorithm (Pritchard et al., 2000) (explained with more detail in section 3.4) estimated a within-population variation of ~93-94%, a ~4% between-groups and a ~ 2.5 between-population, within-group – lower values for the between-group variation estimated by Lewontin. In addition, the STRUCTURE-based classification of the sample set defined clusters that corresponded to continentally-defined population groups when using a 5-group separation (Eurasia, sub-Saharan Africa, East Asia, America and Oceania) and a subdivision of the Eurasian continent using a 7-group separation (sub-Saharan Africa, Europe, Middle East, Central/South Asia, East Asia, America and Oceania; herein AFR, EUR, ME, SAS, EAS, AME and OCE, respectively). A subsequent study by Li et al. re-analysed the same population samples using a panel of 650,000 SNPs (Li et al., 2008). Despite the patterns obtained from both studies being highly comparable, the results for the 7-group separation showed that South Asian and Middle Eastern populations did not show full cluster membership to a specific cluster in Li's assessment, which is an unsurprising finding, given the lower divergence of these population groups.

The Hardy–Weinberg principle dictates that, in absence of natural selection or other forces such as mutations, migration or population substructure, the allele frequencies of a population remain constant. However, the assumptions made from this model – including random mating, no overlap between generations and an infinitely large population size – represent unrealistic conditions that do not apply in real populations. Instead, the effect of different evolutionary forces will act on these populations, modifying their allele frequences and shaping the different populations. In other words, the Hardy–Weinberg principle is used to measure the effect on various factors on allele frequency changes between generations, and allows the measurement of many of these factors from observing changes in allele frequency over time. For human populations the main factors influencing any one marker's allele frequencies are mass movements between continents, and across previous geographic barriers, resulting from migration, trade and slavery. Other factors have a greater time depth so have not shaped allele frequencies to such a large extent, although it has been speculated that Lactase persistence has evolved in a matter of a few thousand years (Segurel & Bon, 2017).

Mutations (i.e., the introduction of random changes in the genetic code) are responsible for the introduction of variation in the population. The effect of mutations does not have *per se* a great impact on the population frequencies but will depend on the action of other evolutionary forces. Natural selection, for instance, will act on any variant that has positive on survival or reproductive success an effect on the individuals and conversely reducing the frequency of the variations that carry disadvantageous or pejorative effects. Several factors may create different selection pressures on certain traits, including climate, disease or diet (Coop et al., 2009). Positive selection of lactase persistence in adults marks a clear example of natural selection, as the not once, nor twice, but thrice independent events where this trait has been selected due (Ingram et al., 2009) – although not frequent enough for the European population,

USC DE SANTIAGO DE COMPOSTELA

unfortunately). In addition, frequencies can also be influenced by an aleatory sampling process known as genetic drift. Every population has an infinite number of individuals, so the alleles that will be passed to each new generation will be randomly selected from those present in the parental generation. Genetic drift is highly dependent on the population size: while large populations may maintain similar allele frequencies with small fluctuations, in smaller populations it may cause the fixation (i.e., reach a frequency of 1) or loss of an allele after a few generations (Wright, 1931). Finally, migratory events can constitute a significant source of variation. Similarly to the genetic drift, migration constitutes a random process that changes allele frequencies of a population. The key difference between genetic drift and migration is that migration cannot alter the frequences species-wide, but it may alter the allele frequencies of populations. In addition, both have opposite effects on allele changes: while the genetic drift tends to increase the variation among different groups, the genetic flow from migration events acts as an opposing force by reducing the genetic differences among populations (Jobling et al., 2013). Figure I5 illustrates how these evolutionary forces can affect allele frequencies.

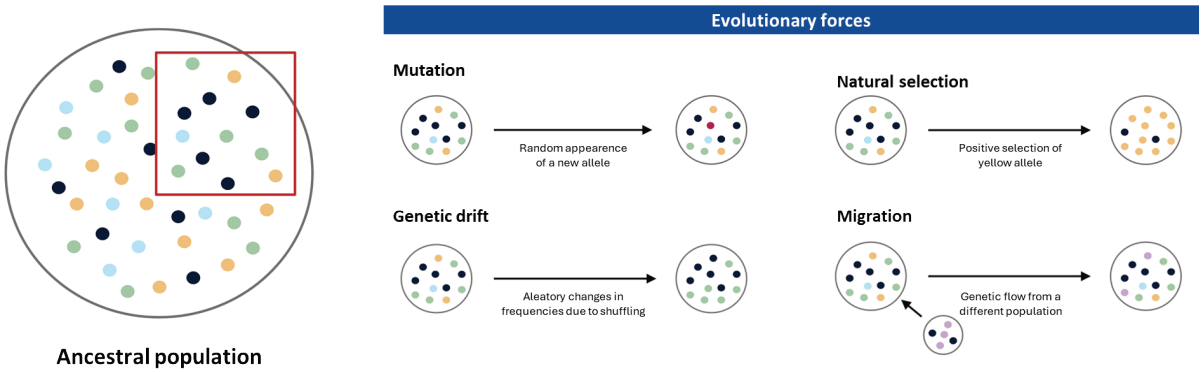


Figure I5. Schematic representation of the effect of evolutionary forces on allele frequency variation. Created with Biorender

The Out-of-Africa model marks the most accepted theory for the explanation of the origin of modern humans and their expansion to the rest of the continents 90-120 thousand years ago, a theory that is supported by multiple types of evidence, including a greater genetic variability found in African populations. This expansion entailed a serial founder effect, where several consecutive population bottlenecks led to a progressive reduction in genetic variability, denoted in the genetic diversity cline observed across an axis that spanned Africa-to-Eurasia, and then in two directions to Oceania, and America via Beringia (Friedlaender et al., 2008; Wang et al., 2007).

Several additional factors also explain the current shape of human populations. The interbreeding among human beings is not random, but the presence of geographical barriers like mountains, deserts or oceans or the geographical distance itself hinders the free movement

of individuals and, consequently, the gene flow between the populations isolated by the above factors (Phillips, 2015) . In addition, the action of natural selection in the adaptation to different bio-geographical factors, recent migration events followed by admixture of ancestral populations (Pickrell & Reich, 2014) and events of archaic introgression between modern humans and archaic Hominins, represented by the recent detailed characterization of Neanderthal and Denisovan genetic variation present in non-Africans and Oceanians, respectively (Gibbons, 2011; Green et al., 2010; Pickrell & Reich, 2014; Reich et al., 2010; Sankararaman et al., 2012).

1.5.4.2 Inference of biogeographic ancestry using genetic markers

As discussed in the previous section, human population genetic patterns of variation are not characterised by discrete groups of individuals but are better represented by geographic clines. Large sets of randomly chosen markers can be used to group the human population into main continentally defined populations. However, these approaches are susceptible to ascertainment bias (Enoch et al., 2006), and would unfeasible for most forensic samples. The use of small sets of genetic markers with strong contrasts in allele frequences between population groups, termed as Ancestry Informative Markers (AIMs), represents a more suitable and sensitive approach for the detection population stratification (Phillips, 2015). Small sets of AIMs implemented in SNaPshot multiplexes, with the capacity of differentiate up to 5 different population groups have been for a long period of time a recurrent tool used for BGA prediction (de la Puente et al., 2016; Phillips et al., 2012b). The use of larger sets of AIMs however, are still an emerging application in forensic analyses – and constitute one of the main areas of focus of this Thesis.

The use of ancestry markers such as AIMs provides a global view of the biogeographic ancestry of the individuals. However, as previously outlined, lineage markers (the Y-chromosome and the mitochondrial genome) constitute two powerful tools to study the paternal and maternal lineages that can provide more background information on admixed individuals or hint the BGA when no autosomal information is available. Due to the lack of recombination, they are transmitted as haplotype blocks, and are relatively stable, only affected by mutation events. Given their transmission pattern, they have a lower effective population size compared to the autosomes or the X chromosome. Thus, as previously detailed, a greater genetic drift effect in their haplotypes. Together with the lack of recombination, this confers a high stability for the haplotypes in the mtDNA and Y-chromosome haplogroups (Jobling & Tyler-Smith, 2004) and they have been shown to have strong correlations with specific geographical regions.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.5.5 Chronological age prediction

The prediction of chronological age (i.e., the time elapsed since an individual's birth) has been a topic of great interest in the forensic sciences. Not only can it provide valuable information about a suspect and help narrow the field of suspects, but it can also be beneficial to enhance the prediction models for other tools influenced by age, including male baldness patters, face morphology or hair colour.

Distinct approaches have been proposed and tested to ascertain the chronological age from a biological sample. The accumulation of deletions in the mitochondrial genome resulting from oxidative damage by reactive oxygen species for instance, was explored by Meissner et al. (Meissner et al., 1999), but technical limitations and low precision constituted a deterrent of its implementation in forensic cases. The shortening of the telomeres (the repetitive sequences located in the extremes of chromosomes), caused by the decrease of the telomerase activity due to aging has also been considered as a potential tool, but the results of this approach yielded high prediction errors and required a substantial amount of DNA (Tsuji et al., 2002). Other studies focused on the analysis of aspartic acid D-aspartic acid accumulation caused by the amino acid racemization process (Waite et al., 1999). In mammals only L-amino acids are used for protein synthesis, but with age the racemization of the amino acids gives rise to an increasing population of D-amino acids. Waite et al. developed a model measuring the proportion of D-aspartic acid in the dentin, obtaining a prediction error of just 3 years. However, the application of this technique in forensic studies is very limited as it is restricted to a single tissue. Additional studies variations in tissue colouration caused by the glycation of proteins, that produce Advanced Glycation End-products (AGEs), whose accumulation causes a yellow-brown coloration. Their use as a forensic test would be however very challenging given the lack of standardization and the restricted age interval prediction (45 years or older) (Pilin et al., 2007). Another approach assessed the use of signal-joint T-cell receptor excision circles (sjTRECs), a circular DNA residue secreted by T cells during their maturation in the thymus. Throughout life the thymus activity starts to fade, reducing the number of sjTRECs in the system. Several works have attempted to use this principle for the inference of age, but models based on the sole use of this marker yielded prediction errors of approximately 10 years, while also being restricted to the analysis of blood samples (Zubakov et al., 2010).

A major breakthrough for this application however came with the publication of Horvath's epigenetic clock, one of the first and most complete approaches based on the use of methylation markers for age prediction. Horvath developed a model with 353 CpG sites suitable for most tissues and cell types (Horvath, 2013). While the number of markers analysed constituted a limitation for its application in forensic analyses, it sat the basis for the development of posterior assays for age prediction. Up to date, multiple models have been described for the most relevant tissues in forensics, including blood (Freire-Aradas et al., 2016), cartilage (Freire-Aradas et al., 2023), saliva (Ambroa-Conde et al., 2022), bone (Lee et al., 2020) or semen (Lee et al., 2024).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.5.6 Investigative Genetic Genealogy

Investigative Genetic Genealogy (IGG), also known as Forensic Genetic Genealogy (FGG) refers to the use of methodologies that make use of high-density SNP data to analyse distant kinship (Kling et al., 2021). In recent years, the interest towards IGG has surged, especially following the identification of Joseph DeAngelo as The Golden State Killer. The relevance of this investigation was not only due to the high profile of the case, involving a serial killer and serial rapist from the 70's and 80's, but to the controversy of the approach used by the authorities for his identification and the ethical concerns that it entailed (Syndercombe Court, 2018). The main evidence that led to the detention of DeAngelo was the match of a distant relative of in the GEDmatch database, based on a high-density SNP profile (mimicking the ones used from direct-to-consumer (DTC) companies) generated from the DNA retrieved from a victim (Phillips, 2018). These events were the catalyst to the implementation of multiple changes to the company's policies, such as the access regulation of police investigators, or the implementation of an opt-in policy to consent the use of their genetic information in police investigations. Despite this, the interest in IGG methodologies is still rising; and has become an emerging topic in the forensic community, gathering interest from both researchers and private companies – reflected by the recent acquisition of GEDmatch by a large multinational corporation.

Given the usual level of forensic DNA, it is important to review how IGG SNP typing methodologies work. Two main methods exist for the identification of distant relatives using high-density SNP data: the likelihood approach and exploratory approaches. The likelihood approach relays on Bayes' Theorem for the calculation of the likelihood ratio given two hypotheses (e.g., related vs unrelated), similar to the procedure used for standard kinship calculation. This method, however, requires prior knowledge of the relationship between the tested individuals and generally entails computationally intense calculations. In contrast, exploratory methods estimate the degree of relatedness measuring the total length of IBD segments, which is methodology commonly adopted by the major DTC companies. Closely related individuals are expected to share long stretches of IBD segments, while the length of these segments decreases as the relationship becomes more distant (Kling et al., 2021). Although long-distance relatives may not share detectable IBD segments, the study by Henn et al. showed that IBD analysis can identify up to 9th cousins (Henn et al., 2012).

The main limitation of the implementation on forensic contexts is that, generally, the dense SNP data is generated with SNP arrays that require large quantities of non-degraded DNA and therefore of little use when working with degraded or trace DNA samples. Nonetheless, recent studies have begun to successfully analyse large SNP datasets using strategies compatible with the analysis of forensic samples (Antunes et al., 2024; Phillips et al., 2020; Tillmar et al., 2020; Tillmar et al., 2021) – further reviewed in the Discussion section.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2. OBJECTIVES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

2. OBJECTIVES

The main objective of this Thesis was the design and evaluation of large-scale genetic marker panels for forensic purposes, to exploit the capability of massively parallel sequencing (MPS) platforms and upgrade the tools that are currently available, throughout the selection of more informative markers and the expansion of panel sizes. Research worked on both objectives: tools aiming to enhance the inference of bio-geographical ancestry and individual human identification. From this main objective, the following sub-objectives were defined:

i) Evaluating the VISAGE Basic Tool for ancestry inference, a tool designed for the simultaneous analysis of single nucleotide polymorphisms to predict ancestry (aiSNPs) and phenotype (EVC SNPs).

ii) Evaluating the VISAGE Enhanced Tool, an improved and more precise version of the VISAGE Basic Tool, that incorporates newly selected and expanded numbers of aiSNPs, not only aiming to differentiate a higher number of populations, but also incorporating microhaplotypes (MHs), Y- and X-SNPs.

iii) Assessing the ancestry informativeness of a previously designed 113-MH panel, originally intended for identification purposes, to expand its applicability in other forensic contexts.

iv) Exploring the creation of panels for ancestry inference based on the use of *Signature AIMs*, SNPs showing absolute specificity for a single population group – defined as SNPs with zero specific allele frequency in all populations except that targeted, and testing this concept by compiling a SAS specific *Signature AIM* panel.

v) Performing a systematic search of tri-allelic SNPs in the 1000 Genomes Phase III data for the development of a large-scale forensic panel for the identification of missing persons.

vi) Testing and validating the tri-allelic SNP panel on three of the most utilized MPS platforms in the forensic field: the Ion S5, the MiSeq and the GeneReader.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3. METHODOLOGY

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3. METHODOLOGY

3.1 RETRIEVING GENOTYPES FROM WGS PROJECTS

Throughout the different studies compiled in this Thesis, the genotyping results from different whole-genome sequence projects were collected to gather data from as many geographical regions as possible, to gauge the variability of the markers in different population groups; and arranging the use of this data for several purposes, including the selection of candidate loci for the multiplexes, compiling the allele frequency estimates from, and performing a genotype concordance analysis. The whole genome sequence (WGS) projects that were consulted were: (i) The 1000 Genomes Project, including Phase III results (The 1000 Genomes Project Consortium, 2015) and the high coverage sequencing analyses performed by the New York Genome Centre (NYGC) (Byrska-Bishop et al., 2022); (ii), the HGDP-CEPH human diversity panel (Bergström et al., 2020); (iii) Simons Foundation human genome diversity project (Mallick et al., 2016); (iv) Estonian Biocentre genome diversity project (Pagani et al., 2016); and, (v) a 137 sample collection from 8 Middle East populations by Almarri et al. (Almarri et al., 2021). The genotype data from Genome in a Bottle Consortium (Wagner et al., 2022) was also consulted for concordance analysis and the Genome Aggregation Database (gnomAD) was used for contrasting population frequencies from the independent population frequency estimates at this database (Chen et al., 2024).

The VCF (Variant Call Format) files from the above sequencing projects, containing both the genotypes for each individual sample and variant frequency estimates per population, were queried using the *BCFtools* query tool, adapting the parameters of the filtering to the desired characteristics – e.g. selecting specific SNPs for specific samples in order to perform genotype concordance, or filtering the SNPs having an allele frequency for the alternative allele for a population higher than 0.1 and zero (or lower than 0.01) for the rest of the populations, to compile *Signature AIMs*.

3.2 SELECTING MARKERS FOR MPS PANEL CONSTRUCTION

MPS for forensic analysis usually focuses on the selection of a reduced set of highly informative loci, which are usually analysed by conducting a targeted sequencing approach instead of pursuing whole genome sequencing, to maintain the sensitivity of the assay for low level or degraded DNA, typical of forensic traces, and increase the coverage of those regions while also reducing the cost of the analyses. The first step towards the construction of an MPS panel

USC DE COMPOSTELA

requires the evaluation of markers to select the most informative ones for the desired multiplex, and the criteria used for the selection will vary depending on the type of panel design required for a particular forensic DNA analysis.

3.2.1 Selecting markers for ancestry inference

Ancestry inference is based on the use of markers that present a contrast in allele frequencies between the target population and other populations to differentiate. Different metrics measure those differences and can be used for pairwise calculations to determine the informativeness of the markers and how to rank them for selection according to their Population Specific Divergence (PSD). These values are calculated by performing pairwise comparisons, taking into account the target population and the other populations merged into a single allele frequency group (e.g., the European populations versus the rest of the world populations). Some examples of metrics traditionally used to measure PSD are the absolute allele frequency difference (δ) and the fixation index (F_{ST}). In the research described here, the main metric used to measure the PSD was Rosenberg's Informativeness (I_n, informativeness for assignment) (Rosenberg et al., 2003), which is calculated as: $I_n = \sum_i(-p_{i,j} \ln p_{i,j} + \frac{1}{K}\sum_k(p_{i,j,k}\ln p_{i,j,k})$, where $p_{i,j,k}$ represents the frequency of allele *i* in locus *j* for population *k*., whereas $p_{i,j}$ represents the frequency of the same allele across all populations. These calculations were performed through the *PCA graphics and analysis of population data* tool in the USC Snipper portal[3]. In addition to the selection of the most informative markers for each population, a balance of the cumulative PSD among the different populations was sought (i.e., ensuring there was equal informativeness for each population), to avoid potential bias of the panel towards the selection of any one population. Careful manual re-selection of markers was performed to achieve this balance, due to differences in the divergence between populations.

In the opposite way to the above strategy, the selection of *Signature AIMs* SNPs with absolute population specificity did not rely on ranking the markers by their divergence, but instead they were identified to have the desirable frequency patterns by performing a frequency intersect by restricting the SNPs that had relatively high alternative allele frequency (> 0.1) in the target population, but a frequency lower than 0.01 in the other populations.

Additionally, two other metrics were also used to measure the genetic variation within and between populations, the mean number of pairwise differences (π) and F_{ST}. The π metric averages the number of differences between all pairs of sequences, while F_{ST} is a metric from the family of fixation indices first defined by Wright (Wright, 1950), which is a measure of the genetic diversity in the subpopulations in contrast to the genetic diversity in the whole population. Nei's adaptation of the F_{ST} metric (Nei, 1973) is calculated as $F_{ST} = \frac{(H_T - H_S)}{H_T}$, where H_T is the expected heterozygosity in the total population and H_S in the mean expected

[3] http://mathgene.usc.es/snipper/

heterozygosity in the subpopulations. These two measures were calculated using the *Arlequin* software (Excoffier & Lischer, 2010).

.

3.2.2 Selecting markers for Human Identification

The selection of markers for human identification aimed to maximize the informativeness of the markers by selecting loci with a high level of polymorphism. Thus, candidate SNPs were ranked according to their heterozygosity (*Het*), calculated as: $Het = 1 - \sum_i p_{i,j}$, where p is the allele frequency for the allele i in the locus j. In addition, candidate loci were sought that were informative regardless of the population of origin. In this sense, the selection of SNPs with allele frequencies not showing high contrasts between different population groups.

3.2.3 Other considerations for marker selection

Even if markers presented the desired characteristics for ancestry inference or identification, the adequacy of the candidate loci to be implemented in an MPS assay was evaluated before they were included in the selection. This included aspects such as the contrast of different databases and/or projects, so as to confirm that the allele frequency patterns observed were maintained (i.e., to determine if a SNP allele specific to a population is near identical across different population samples in different databases); the scrutiny of the context sequence to inspect the presence of elements that may hinder the correct alignment of the reads or primer binding; or the assessment of distance between markers is taken into consideration, meaning markers must me adequately spaced across each chromosome to avoid linkage between them.

3.3 DNA TYPING

3.3.1 Sample selection

Several types of samples were used for sequencing depending on the aim of the investigation. Population studies included DNAs from the HGDP-CEPH Human Genome Diversity Panel (Cann et al., 2002) as well as in-house population samples (ethical approval CAAE: 16542613.8.0000.0030, CAAE: 72917916.3.0000.0030, and reference no. 17–416) to expand data on under-represented geographical regions. In addition, other samples such as forensic DNA controls, including 2800M Control DNA (Promega) and NIST Standard Reference Material samples were also sequenced to conduct different types of studies for the evaluation of the panel, e.g., sensitivity or concordance studies.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.3.2 Library preparation and genotyping

For genotyping, the main MPS platform utilized was the Ion S5 (TFS). Libraries were prepared manually using the corresponding primer pool, quantified with qPCR using a 7500 Real-Time PCR System (TFS) and subsequently pooled at equimolar concentrations. Automated template preparation was conducted with the Ion Chef System (TFS) using either an Ion 520 or an Ion 530 sequencing chip. Sequencers MiSeq FGx (Verogen) and GeneReader (Qiagen), were also included for some studies, for which the corresponding kits and protocols were used.

3.3.3 Bioinformatic analysis

Sequencing data from the Ion S5 was mainly analysed through the *Ion Torrent Suite* and *HID_SNP_Genotyper* plugin using the default parameters, which allowed the direct retrieval of SNP genotypes. In other cases, when MHs were analysed, or other information had to be taken into account (i.e., UMIs from the reads), FASTQ files were exported, aligned to the reference genome and the information on the target loci was extracted using either custom pipelines based on open-source software or the CLC Genomic Workbench software (Qiagen). In those instances, allele calling was performed with custom R scripts, and using IGV (Robinson et al., 2011) alignments were inspected when necessary to ascertain the correct genotype assignment or uncover possible artifacts.

3.4 Population data analysis

Different methodologies were used for the analysis of population data, in order to evaluate the degree of separation that could be accomplished with a set of markers and the capability to infer the ancestry from unknown donors. A constrained set of populations were used as reference data to represent each geographic area, seeking a balance of the number of samples in each of them trying, in turn, to avoid any ascertainment bias. This implied the collection of the genotype data for the target populations, which was conducted as described in section 3.1. Generally, Africans were represented by the Yoruba from Nigeria (YRI) from 1000 Genomes (1KG); Europeans with NW Europeans from Utah (CEU) from 1000 Genomes; East Asians by Han Chinese from Beijing (CHB) from 1000 Genomes; South Asians by Gujarati from Houston (GIH) from 1000 Genomes; Middle East by the HGDP-CEPH Israeli Arabs from Palestinian, Druze and Bedouin populations plus Algerian Mozabite (although Algerians were considered as a separated group representing North Africa for several analyses); Oceanians by the HGDP-CEPH Papuans from Bougainvillea and Papua New Guinea; Native Americans samples, by the HGDP-CEPH samples from Maya, Pima, Colombian and Amazonian Surui and Karitiana populations, and 1KG Peruvians from Lima, Peru (PEL), the latter limited to those in which no

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

indication of co-ancestry patterns had been previously detected, based on the analysis using the *Affimetrix Human Origins array* SNPs (Phillips et al., 2021).

The first method used for the analyses was a naïve Bayes classifier, a methodology that uses the Bayes theorem to calculate the odds of an unknown sample belonging to each population based on their allele frequency estimates. The Snipper App portal was used for these analyses, providing the LR value using the ratio between the two populations with the highest probabilities. In addition, the *PCA graphics and analysis of population data* tool from Snipper was used to perform a leave-one-out cross-validation – that is, an analysis that consists of iteratively selecting one sample from the reference set and using the rest of the data to classify the selected sample, so as to assess the efficiency of the ancestry markers by measuring the number of times that the sample is classified correctly. Bayes classifiers however provide an 'absolute' classification of the profile and therefore as the highest likelihood is always a large value when data from 100 or more SNPs are used, such classifiers are not suitable for admixed individuals, as the second highest likelihood (for the minor co-ancestry component) can be much lower in value.

STRUCTURE software (Pritchard et al., 2000) was widely applied to the population data analysis, particularly to detect and analyse admixture from the presence of multiple ancestries in an individual's background. STRUCTURE is a program that uses a Bayesian approach, relying on a Markov Chain Monte Carlo (MCMC) algorithm, allowing the estimation of the cluster memberships of samples. The program performs an initial random assignation of the samples into a number of pre-defined clusters (the 'K' number) and then an iterative process begins, in which the allele frequency estimations for each group are used to re-allocate the samples (burnin period); followed by the MCMC iterations, where the estimates of the cluster memberships are measured. When the run is performed using the admixture model, to make adjustments to parameters applied to the cluster algorithm, it provides a matrix containing the membership coefficients for each of the K groups for each sample (otherwise, using the non-admixture model would assign the sample into a single group and provide a matrix containing the posterior probabilities). Typically, a range of different K values were run, each of which is run as replicates to explore the degree of differentiation that can be achieved. The output from STRUCTURE, which comprised the different replicates for each K value, was exported to CLUMPAK (Kopelman et al., 2015), a software suite that combines two different utilities: CLUMPP (Jakobsson & Rosenberg, 2007), that merges the results from the different replicates run for each K value, and Distruct (Rosenberg, 2004), which provides a bar plot for each sample representing their cluster membership proportions. Additionally, Structure Harvester software (Earl & vonHoldt, 2012) was employed to aid in the determination of the K value that optimally fits the data.

Principal component analysis (PCA) and multi-dimensional scaling (MDS) analysis were also used for the analysis of population data. Both methodologies follow similar strategies, aiming for the reduction of the number of dimensions of the data in order to reduce its complexity and allow their visualization so that an inspection of whether the clustering of the

USC UNIVERSIDADE DE SANTIAGO

data is observed. PCA converts a set of variables that are correlated into another set of uncorrelated orthogonal variables. Each of these new variables are termed a Principal Component (PC), and are ranked in order according to the amount of variation they explain (i.e., PC1 will account for most of the variance, then PC2, and so on). MDS serves a similar purpose, but MDS prioritizes maintenance of optimal separation distances between the points instead of maximizing the variance, and then provides a representation in a lower-dimensional space. In both instances, the two or three first PCs or dimensions are plotted in sets of 2D plots (i.e. representing PC1 vs PC2, PC2 vs PC3, etc.) to analyse the clustering of the samples, since 2D plots are easy to visualize and, despite using a reduced number of PCs/Dimensional combinations, these account for most of the variability in all cases. The PCA analyses were performed using Snipper, whereas MDS analyses were made with the the *pegas* R software package (Paradis, 2010). The construction of neighbour-joining trees (NJT) (Saitou & Nei, 1987) was also used to explore data clustering. This approach consists of the construction of an unrooted phylogenetic tree, which is done by minimizing the branch length at each step.

Finally, the Genetic Distance Algorithm (GDA) developed by Cheung et al. (Cheung et al., 2017) and implemented in the Snipper portal, was also employed to analyse the population data. The algorithm measures the distance in genotype frequency between the reference populations, allowing the inference of co-ancestry proportions from unknown samples.

3.5 Pedigree analyses

Famlink software (Kling et al., 2012) was used to simulate profiles from different kinship scenarios, for when the relationship between the individuals was true (hypothesis 1 or H1) and when they were not related (H2) – seeking to evaluate the capability of the panel used to correctly differentiate the tested relationship. This approach provided LR distributions for each scenario, calculated as LR=Pr(E|H1)/Pr(E|H2); where E|H1 represents the probability of the profiles given H1 and E|H2 the probability of the profiles when H2 is true. The simulations used the allele frequency estimates for the selected markers along with estimates of centimorgan values (cM) for genetic distances, obtained from the HapMap data (International HapMap Consortium, 2007; Phillips et al., 2012a). The *quickped* suite (Vigeland, 2022) was used to draw the pedigrees.

3.6 Forensic validation of an MPS panel

Validation is defined by the Scientific Working Group on DNA Analysis Methods (SWGDAM) as a «process by which a procedure is evaluated to determine its efficacy and reliability for forensic casework and/or database analysis» (Scientific Working Group on DNA Analysis Methods, 2016). The ISO/IEC 17025 norm, which contains the general requirements for testing and calibration laboratories, also establishes that «The laboratory shall validate non-standard

methods, laboratory-developed methods and standard methods used outside their intended scope or otherwise modified. The validation shall be as extensive as is necessary to meet the needs of the given application or field of application» (ISO/IEC, 2017), implying that testing laboratories that work under the ISO/IEC 17025 accreditation must use validated methods.

Two types of validation can be distinguished: the developmental and the internal validations. While the internal validation is expected to be performed in each laboratory during the implementation of a new procedure to confirm its correct performance, the developmental validation is the process performed for new methodologies that allows the determination of the conditions and limitations of these new methodologies for forensic applications. This entails that a series of studies have to be carried out before it is used for real casework. Based on the SWGDAM guidelines, several studies were performed to validate newly developed MPS panels. These included the sequencing of forensic DNA controls and reference samples, sensibility studies, to determine the limit of detection of the assay; mixture studies, to evaluate the capacity to detect mixtures; and concordance studies to test the accuracy of the results.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

4. RESULTS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

4. RESULTS

Article 1. Development and Evaluation of the Ancestry Informative Marker Panel of the VISAGE Basic Tool

Reference: Genes (2021). 12(8):1284

ISSN: 2073-4425

DOI: 10.3390/genes12081284

Web link: https://www.mdpi.com/2073-4425/12/8/1284

Authors: de la Puente M[1], Ruiz-Ramírez J[1], Ambroa-Conde A[1], Xavier C[2], Pardo-Seco J[3], Álvarez-Dios J[4], Freire-Aradas A[1], Mosquera-Miguel A[1], Gross TE[5,6], Cheung EYY[5], Branicki W[7], Nothnagel M[8,9], Parson W[2,10], Schneider PM[5], Kayser M[11], Carracedo A[1,12], Lareu MV[1], Phillips C[1].

Author information:
1. Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain
2. Institute of Legal Medicine, Medical University of Innsbruck, 6020 Innsbruck, Austria
3. Genetics, Vaccines, Infectious Diseases and Pediatrics Research Group (GENVIP Group), Instituto de Investigación Sanitaria de Santiago de Compostela, 15706 Santiago de Compostela, Spain.
4. Faculty of Mathematics, University of Santiago de Compostela, 15705 Santiago de Compostela, Spain.
5. Institute of Legal Medicine, Faculty of Medicine and University Clinic, University of Cologne, 50823 Cologne, Germany.
6. Hessisches Landeskriminalamt, 65187 Wiesbaden, Germany.
7. Malopolska Centre of Biotechnology, Jagiellonian University, 30-387 Kraków, Poland.
8. Cologne Center for Genomics, University of Cologne, 50823 Cologne, Germany.
9. University Hospital Cologne, 50937 Cologne, Germany.
10. Forensic Science Program, The Pennsylvania State University, University Park, State College, PA 16802, USA.
11. Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, 3015 CN Rotterdam, South Holland, The Netherlands.
12. Fundación Pública Galega de Medicina Xenómica (FPGMX), 15706 Santiago de Compostela, Spain.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Quality metrics (JCR, 2021):

- Journal Impact factor: 4.141
- Quartile: Q2 (genetics & heredity)
- Rank by JCR: 72/175 (genetics & heredity)

Contribution: methodology, MPS sequencing, curation of the genotypes and collection of population data.

Authorization:

© 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (**https://creativecommons.org/licenses/by/4.0/**).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

genes

MDPI

Article

Development and Evaluation of the Ancestry Informative Marker Panel of the VISAGE Basic Tool

María de la Puente [1], Jorge Ruiz-Ramírez [1], Adrián Ambroa-Conde [1], Catarina Xavier [2], Jacobo Pardo-Seco [3], Jose Álvarez-Dios [4], Ana Freire-Aradas [1], Ana Mosquera-Miguel [1], Theresa E. Gross [5,6], Elaine Y. Y. Cheung [5], Wojciech Branicki [7], Michael Nothnagel [8,9], Walther Parson [2,10], Peter M. Schneider [5], Manfred Kayser [11], Ángel Carracedo [1,12], Maria Victoria Lareu [1], Christopher Phillips [1,*] and on behalf of the VISAGE Consortium [†]

1 Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain; m.delapuente.vila@gmail.com (M.d.l.P.); ruizramirez.jorge@gmail.com (J.R.-R.); ambroa.adrian@gmail.com (A.A.-C.); ana.freire3@hotmail.com (A.F.-A.); ana.mosquera@usc.es (A.M.-M.); angel.carracedo@usc.es (Á.C.); mvictoria.lareu@usc.es (M.V.L.)
2 Institute of Legal Medicine, Medical University of Innsbruck, 6020 Innsbruck, Austria; catarinagaxavier@hotmail.com (C.X.); walther.parson@i-med.ac.at (W.P.)
3 Genetics, Vaccines, Infectious Diseases and Pediatrics Research Group (GENVIP Group), Instituto de Investigación Sanitaria de Santiago de Compostela, 15706 Santiago de Compostela, Spain; j.pardoseco@gmail.com
4 Faculty of Mathematics, University of Santiago de Compostela, 15705 Santiago de Compostela, Spain; joseantonio.alvarez.dios@usc.es
5 Institute of Legal Medicine, Faculty of Medicine and University Clinic, University of Cologne, 50823 Cologne, Germany; gross.theresa@gmail.com (T.E.G.); elaine.yy.cheung@outlook.com (E.Y.Y.C.); peter.schneider@uk-koeln.de (P.M.S.)
6 Hessisches Landeskriminalamt, 65187 Wiesbaden, Germany
7 Malopolska Centre of Biotechnology, Jagiellonian University, 30-387 Kraków, Poland; wojciech.branicki@uj.edu.pl
8 Cologne Center for Genomics, University of Cologne, 50823 Cologne, Germany; mnothnag@ini-koeln.de
9 University Hospital Cologne, 50937 Cologne, Germany
10 Forensic Science Program, The Pennsylvania State University, University Park, State College, PA 16802, USA
11 Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, 3015 CN Rotterdam, South Holland, The Netherlands; m.kayser@erasmusmc.nl
12 Fundación Pública Galega de Medicina Xenómica (FPGMX), 15706 Santiago de Compostela, Spain
* Correspondence: c.phillips@mac.com
† A complete list of the centers and investigators in the VISAGE GROUP is provided in Appendix A.

check for updates

Citation: de la Puente, M.; Ruiz-Ramírez, J.; Ambroa-Conde, A.; Xavier, C.; Pardo-Seco, J.; Álvarez-Dios, J.; Freire-Aradas, A.; Mosquera-Miguel, A.; Gross, T.E.; Cheung, E.Y.Y.; et al. Development and Evaluation of the Ancestry Informative Marker Panel of the VISAGE Basic Tool. *Genes* **2021**, *12*, 1284. https://doi.org/10.3390/genes12081284

Academic Editor: Niels Morling

Received: 21 July 2021
Accepted: 18 August 2021
Published: 22 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: We detail the development of the ancestry informative single nucleotide polymorphisms (SNPs) panel forming part of the VISAGE Basic Tool (BT), which combines 41 appearance predictive SNPs and 112 ancestry predictive SNPs (three SNPs shared between sets) in one massively parallel sequencing (MPS) multiplex, whereas blood-based age analysis using methylation markers is run in a parallel MPS analysis pipeline. The selection of SNPs for the BT ancestry panel focused on established forensic markers that already have a proven track record of good sequencing performance in MPS, and the overall SNP multiplex scale closely matched that of existing forensic MPS assays. SNPs were chosen to differentiate individuals from the five main continental population groups of Africa, Europe, East Asia, America, and Oceania, extended to include differentiation of individuals from South Asia. From analysis of 1000 Genomes and HGDP-CEPH samples from these six population groups, the BT ancestry panel was shown to have no classification error using the Bayes likelihood calculators of the *Snipper* online analysis portal. The differentiation power of the component ancestry SNPs of BT was balanced as far as possible to avoid bias in the estimation of co-ancestry proportions in individuals with admixed backgrounds. The balancing process led to very similar cumulative population-specific divergence values for Africa, Europe, America, and Oceania, with East Asia being slightly below average, and South Asia an outlier from the other groups. Comparisons were made of the African, European, and Native American estimated co-ancestry proportions in the six admixed 1000 Genomes populations, using the BT ancestry panel SNPs and 572,000 Affymetrix Human Origins array SNPs. Very similar co-ancestry proportions were observed down to a minimum value of 10%,

below which, low-level co-ancestry was not always reliably detected by BT SNPs. The *Snipper* analysis portal provides a comprehensive population dataset for the BT ancestry panel SNPs, comprising a 520-sample standardised reference dataset; 3445 additional samples from 1000 Genomes, HGDP-CEPH, Simons Foundation and Estonian Biocentre genome diversity projects; and 167 samples of six populations from in-house genotyping of individuals from Middle East, North and East African regions complementing those of the sampling regimes of the other diversity projects.

Keywords: bio-geographical ancestry; massively parallel sequencing; ancestry informative markers; SNPs; 1000 Genomes; Human Origins SNP array

1. Introduction

In the last ten years, forensic DNA analysis has been extended beyond identification tests that link a suspect to crime-scene evidence using STR profiling, into areas of genetic analysis designed to predict a range of physical characteristics of unidentified trace donors, mainly defined by single nucleotide polymorphisms (SNPs). Such SNP tests, applicable when no suspect has been identified by investigators, have been termed forensic DNA phenotyping (FDP), and until recently, consisted of tests to infer bio-geographical ancestry (BGA) [1], or to predict certain externally visible characteristics (EVCs) [2]. These SNP-based tests have now been supplemented with DNA analyses that measure the methylation status of carefully chosen CpG sites showing age-correlated methylation patterns in order to estimate a donor's chronological age from the same DNA samples used for BGA and EVC tests [3]. The array of FDP tests now being adopted can serve to progress a major criminal investigation lacking strong leads, resurrect cold cases by providing new genetic data about old crime-scene DNA samples, and aid in the identification of the missing or historical remains [4]. While FDP tests have been gaining traction as supplementary analysis regimes available to forensic laboratories, the emergence of massively parallel sequencing (MPS) now provides major enhancements for forensic DNA analysis, including improved sensitivity to analyse minimal amounts of evidential material, larger multiplexes that can combine conventional identification STRs with SNPs for FDP analysis, and greater detail about the detectable variation within a sequence. The VISible Attributes through GEnomics (VISAGE) Consortium was initiated in 2017 to develop new MPS-based tools for predicting the BGA, appearance based on common EVCs, and age of an unidentified crime-scene DNA donor.

VISAGE adopted a two-step program to create MPS tools for FDP: developing a Basic Tool as a pilot assay, followed by the development of an Enhanced Tool with more markers and extended capabilities. The VISAGE Basic Tool (herein BT) is the first forensic MPS test combining markers for predicting eye, hair, and skin colour with those for BGA inference in a single assay, along with a saliva-blood focused age estimation assay requiring a parallel assay and workflow due to the bisulphite conversion steps needed to measure methylation. The VISAGE BT was founded on marker panels which have been shown to be informative for the three pigmentation traits [5–7] and for continental-scale population differentiation [8–12]. The level of multiplexing necessary to accommodate sufficient SNPs to yield optimum predictions of both EVCs and continental BGA was in part dictated by experience of building multiplexes for forensic MPS analysis [8,12,13]. Therefore, for the BT assay, which acts as a pilot test, we decided to design a moderate PCR scale of 150 to 160 SNPs by combining the 41 well-established HIrisPlex-S markers, which have been successfully applied in previous SNaPshot and MPS assays [5–7,13], with up to 120 ancestry-informative marker (AIM) SNPs. A panel of 120 AIM SNPs represents a reduced multiplex size compared to the one commercially available forensic ancestry panel for MPS, i.e., the Thermo Fisher Precision ID Ancestry Panel, comprising 165 AIMs [14], although similar in scale to another panel of 127 AIMs [8]. For this reason, we selected markers with the maximum possible ancestry informativeness for any given population,

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

and then balanced these to ensure the major population groups had broadly comparable levels of differentiation. In contrast to the previous development of the 127-SNP Global AIMs panel (gAIMs [8]), South Asians, i.e., individuals from the Indian subcontinent, were treated as a separate population group.

This paper describes the compilation of component SNPs into the ancestry informative marker set forming the major portion of the VISAGE Basic Tool for appearance-ancestry. The development of the AIMs set and its evaluation involved four stages: i. Selection of markers and balancing of their relative population informativeness proportions in the multiplex to ensure comparable levels of continental population group differentiation; ii. Generation of a large-scale reference dataset built on open-access, genome-wide variant datasets, plus our own MPS genotyping of in-house population samples from which to apply the statistical analysis pipelines accompanying the BT assay; iii. Genotype concordance studies and an audit of existing genome-wide variant datasets for the selected AIM SNPs by analysing whole-genome sequence data released by the 1000 Genomes Project in 2020 [15–17]; and iv. Evaluation of the panel's capacity to make efficient population differentiations using existing likelihood-based ancestry inference tests.

2. Materials and Methods

2.1. Selection and Balancing of Component Ancestry Informative SNPs

The BT assay design targeted an AIMs panel of approximately 120 SNPs, based on the fact that ~25% of the full multiplex would be occupied with the 41 markers of HIrisPlex-S, and shared use of key SNPs rs16891982 (in *SLC45A2* gene), rs1426654 (*SLC24A5*), rs12913832 (*HERC2*) for both ancestry and appearance predictive purposes. Because tri-allelic SNPs have been found to be useful for the detection of mixed DNA [18], but when carefully selected have also been informative for ancestry to comparable levels as binary SNPs [19], we expanded their number in the multiplex above those of the previous gAIMs panel (which had six tri-allelic SNPs of a total 127 loci [8]). Twenty tri-allelic SNPs showing the highest levels of population divergence were selected from a very large candidate pool created from a genome-wide compilation of this type of human variation [20].

Careful consideration was made to balance the differentiation of the five main continental population groups of sub-Saharan Africa, Europe, East Asia, Oceania and America (herein, AFR, EUR, EAS, OCE and AMR) with differentiation of the South Asian sub-group (SAS). We decided to target 60–70 SNPs for continental differentiation, plus 20 SNPs to distinguish South Asia from the neighbouring population groups of Europe and East Asia.

SNPs for continental differentiation were compiled from four well-established AIM sets: i. the 127-SNP gAIMs ancestry panel; ii. the 446-SNP LACE (Latin American Cancer Epidemiology) ancestry panel [21]; iii–iv. the 165-SNP Thermo Fisher Precision ID Ancestry Panel (PIAP) [14], which includes the widely used 56-SNP Kiddlab ancestry panel [22]. SNPs differentiating South Asian and European populations were compiled from the above panels and three dedicated AIM sets for such geographic distinctions: v. the best European differentiating SNPs from the EUROFORGEN NAME panel [23]; vi. the Shriver group's US admixture mapping panel, with many highly European-informative loci [24]; and vii. the *Eurasiaplex* South Asian informative panel [9].

The continental SNPs in BT were balanced in terms of their individual population differentiation power to equilibrate as far as possible the five continental groups, ensuring that no one population group had excessive numbers of indicative SNPs, which could potentially bias calculations of co-ancestry proportions in persons with admixed backgrounds (e.g., excessive numbers of African-informative SNPs in such a small panel could overestimate this co-ancestry compared to the others). For each SNP, this process records the population specific divergence ($I_{n\ AFR}$; $I_{n\ EUR}$; $I_{n\ E\ ASN}$; etc., the metric divergence is capitalised and measures the degree of differentiation per marker) to guide the removal or addition of SNPs for a specific population comparison in order to improve overall I_n balance. To reach this balance, the *Snipper* online ancestry analysis portal [25] was used to estimate cumulative I_n values, as previously described for the development of the gAIMs

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

panel [8]. As BT had a smaller number of SNPs than gAIMs, we did not expect to reach a comparably narrow range of finely balanced I_n values.

2.2. Compilation of a 4132-Sample Population Dataset

A comprehensive population dataset for the BT ancestry SNPs was generated by compiling online whole-genome-sequencing variant data from 3965 samples and adding in-house genotyping of 167 samples from six populations, chosen to cover geographic gaps in data for several under-represented regions, i.e., 30 from Eritrea, 16 from Somalia, 32 from Morocco, 32 from Central Iraq, 28 from the Kurdistan region of Iraq and 29 from Turkey (Turkish resident in Germany). An additional 277 DNA samples from the HGDP-CEPH human diversity panel were genotyped with the BT assay with the goal of further extending the geographic coverage of Oceanian, Native American and Middle East (herein ME) population variation, i.e., regions not covered by 1000 Genomes Phase 3 populations. However, the 1000 Genomes Consortium completed whole-genome-sequencing of the full HGDP-CEPH panel in 2019 soon after our in-house genotyping had been completed, so we compiled the complete dataset from 1000 Genomes and used the overlapping sample data to measure the MPS genotyping concordance of BT. This was accomplished for the whole marker set of BT (i.e., including the 41 appearance markers) to gauge sequencing performance across the complete multiplex.

Of the total 4132 SNP genotype profiles compiled, 520 were used as a standardised six population group reference dataset for uploading as training sets to the *Snipper* ancestry analysis portal, with the aim of providing the closest possible equivalence of sample sizes amongst the continental and South Asian regions. The standardised reference dataset consisted of: sub-Saharan Africans represented by 108 1000 Genomes 30X sequence coverage (1KG-30) Yoruba from Nigeria (YRI); Europeans by 99 1KG-30 CEPH Europeans from Utah (CEU); East Asians by 103 1KG-30 Han Chinese from Beijing (CHB); South Asians by 103 1KG-30 Gujarati from Houston (GIH); Oceanians by 28 HGDP-CEPH Papuans from Bougainvillea and Papua New Guinea; Native Americans by 79 samples, comprising 61 HGDP-CEPH samples from Maya, Pima populations, Colombians and Amazonian Surui and Karitiana populations, supplemented by 18 1KG-30 Peruvians from Lima, Peru (PEL) which we had previously analysed and showed no detectable non-American co-ancestry (in-silico analysis of 572,743 Affymetrix Human Origins SNPs, see Table 10.5 of [26]).

Although we established the above standardised reference dataset, the additional populations compiled can be flexibly included in *Snipper* or can replace those in the reference set at the user's discretion by adjusting the Excel-based training set file uploaded to the 'multiple profiles' classifier portal (http://mathgene.usc.es/snipper/analysismultipleprofiles.html, accessed on 20 August 2021), which generates a Bayes likelihood ratio statistical test for ancestry and a matched 2D principal component analysis (PCA) plot. For maximum flexibility, the other population data was arranged into a series of worksheets in the same core Excel reference file, allowing the end-user to select alternative reference populations from 1000 Genomes or HGDP-CEPH (e.g., HGDP-CEPH Pakistani South Asian samples rather than 1KG-30 Gujarati, or other populations from the Indian sub-continent). The same user-driven options apply to the VISAGE in-house population samples provided in a separate worksheet. The admixed populations in 1000 Genomes, comprising 96 African Caribbean individuals in Barbados (ACB), 61 Americans of African Ancestry in SW USA (ASW), 64 individuals with Mexican Ancestry from Los Angeles USA (MXL,) 94 Colombians from Medellin, Colombia (CLM), 104 Puerto Ricans from Puerto Rico (PUR), 67 of 85 PEL (i.e., with detected co-ancestry) were also compiled in a separate worksheet but would not be expected to be useful as training set data from which representative allele frequency estimates could be obtained, given their high levels of individual admixture.

Genome-wide variant data from Simons Foundation human genome diversity project (herein SGDP) has been available for several years [27], and offers additional information on geographic areas outside of those covered by 1000 Genomes or the HGDP-CEPH sampling regimes, in particular, Northeast Asia (broadly, Eastern Siberia east towards the Bering

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Straits); Southeast Island Asia; and Central South Asia (broadly, the Caucasus' east towards the central Asian Steppe immediately north of South Asia). There are 278 genome-wide SNP datasets, of which 22 overlap with 1000 Genomes and 126 overlap with the HGDP-CEPH panel samples, leaving 130 unique samples, from 67 populations. The small sample size of SGDP highlights the very limited number of 1, 2 or 3 samples per region. Nevertheless, we compiled the 130 SGDP-unique BT AIM profiles into a dedicated worksheet and completed population analyses using the standardised reference set.

Genome-wide variant data from the Estonian Biocentre genome diversity project (herein, EGDP) largely mirrors the sampling regimes of SGDP [28], with a maximum of 16 samples per region, but mainly 1, 2 or 3 per region. EGDP has 402 whole-genome datasets from the sampling of 126 populations that almost all complement those of SGDP and provides extensive coverage of regions in Siberia, Northeast Asia, Eastern Europe, and Central South Asia. As with SGDP, the 402 EGDP BT AIM profiles were compiled into a dedicated worksheet and population analyses made with the standardised reference set. Note that the tri-allelic SNPs forming part of the BT AIMs panel are not reported by EGDP and an additional four ancestry SNPs do not have genotypes, so data completeness is ~83.5% (96/115 SNPs).

2.3. Genotyping Concordance amongst MPS Sequence Data and Online Databases

Genotyping concordance for the BT SNPs was assessed in two ways. First, a direct comparison was made between the MPS genotypes generated during developmental runs genotyping 277 HGDP-CEPH DNA samples for all 153 SNPs in BT, and the SNP genotypes reported by 1000 Genomes for the same DNA samples [16], providing 42,381 comparisons. Second, the genotypes listed by 1000 Genomes Phase 3 database for 2504 worldwide population samples (and available in the Ensembl browser at: http://www.ensembl.org/Homo_sapiens/Info/Index, accessed on 20 August 2021) were compared to the high coverage re-sequencing dataset for the same samples, released in early 2020 by the New York Genome Centre [17]. The high sequence coverage genotypes correspond to an average 30X coverage, while the current Phase 3 genotypes were produced from an average 2-3X coverage. We considered the high coverage data to represent more reliable genotype calls for the SNPs of BT, so these were compiled as the major part of the reference population data we used for statistical analysis of ancestry, as described above.

The Coriell control DNA samples originally used as the genotyping concordance standards in the inter-laboratory evaluations of the BT assay [29] were also run as positive controls on each sequencing run performed for the in-house population studies.

2.4. Evaluation of Ancestry and Co-Ancestry Inference Efficiency of the BT AIMs

The efficiency of the BT AIMs to infer ancestry was assessed for a six-group differentiation of African, European, East Asian, South Asian, American and Oceanian ancestries. To measure classification error, one-out cross-validation of the standardised reference dataset of 520 samples was performed using the verbose cross validation option in *Snipper* (http://mathgene.usc.es/Snipper/analysispopfile2_new.html, accessed on 20 August 2021). For all other samples, STRUCTURE, PCA and neighbour joining tree (NJT) analyses were performed on each of the separate population sets, comprising admixed and unadmixed 1000 Genomes samples not used as reference data; HGDP-CEPH samples not used as reference data (i.e., no Oceanian, American population data); SGDP samples, with the six in-house populations included in the same analysis run; and EGDP samples (for 100 binary SNPs of 115 ancestry SNPs in BT). STRUCTURE v.2.3.4 [30] was used for runs consisting of five iterations per K (genetic cluster values K:2 to K:8 evaluated) consisting of 100,000 burnin steps and 100,000 MCMC steps, using correlated allele frequencies under the Admixture model. The optimum 'K' cluster number was inferred by plotting log probability of K and magnitude of ΔK, following the analyses of Evanno et al. [31]. Allele frequencies were updated only using individuals marked POPFLAG = 1 at the optimum genetic cluster number of K:6. Cluster membership proportion plots were constructed

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

with CLUMPAK v.1.1 [32]. Multi-dimensional scaling (MDS) analysis and construction of neighbour-joining trees was implemented using R software v.3.5.0 [33] over an allele-distance matrix computed with the *pegas* R package [34].

The reliable identification of co-ancestry in DNA donors with admixed backgrounds can be considered important information that can be reported to investigators. Although there is likely to be some imprecision in the estimation of co-ancestry ratios when using relatively small-scale AIMs panels like BT, if the markers have been carefully balanced, they are more likely to be free of bias towards one particular co-ancestral population group and therefore match more closely the actual co-ancestry proportions in such individuals. To measure the reliability of co-ancestry estimation from identifying genetic clusters using the BT AIMs, we ran all 504 admixed 1000 Genomes samples described above (including all 85 PEL), together with the unadmixed populations from AFR, EUR, AMR and EAS, to keep the identified genetic clusters simple (K:4). The ACB and ASW samples represent typical African Caribbean and African American co-ancestry patterns of majority African co-ancestry and 0–30% European co-ancestry. The MXL, CLB, PUR and PEL samples represent more complex three-way co-ancestry patterns of major European co-ancestry, with Native American and African co-ancestries at varied ratios, with a few PEL individuals showing detectable East Asian co-ancestry [15]. We compared the BT data to the co-ancestry ratio estimates from the detailed genetic cluster analysis of the same 1000 Genomes-CEPH populations using 572,743 Affymetrix Human Origins SNPs (see Section 2.2. Although the parallel analyses with the Human Origins SNPs used the ADMIXTURE algorithm [35] to identify clusters, because data input for this algorithm uses only binary variant data, we applied STRUCTURE to ensure the tri-allelic SNP genotypes were assessed. Co-ancestry ratios were grouped into 10-percentiles (i.e., bins of one tenth of the samples from any one population) and the mean ratios of each of these 10-percentile groups in both analyses were assessed for goodness-of-fit using linear regression. Finally, we included a comparison with the genetic distance algorithm (GDA) co-ancestry estimator available in *Snipper*, developed from the work of Cheung et al. 2017 [36].

2.5. Evaluation of Ancestry and Co-Ancestry Inference Efficiency of the BT AIMs

The current gnomAD version 3.1 genome database compiles variation in almost 100,000 individuals from ten populations, so it provides a very efficient system for the detection of rare variation which could be potentially population informative. We screened the gnomAD 3.1 database to make a genomic audit of rare variations associated with the selected BT ancestry SNPs, comprising two types: i. rare, but potentially informative 'flanking' SNPs present on the same amplified fragment as the target SNP; and ii. additional polymorphic alleles in the binary ancestry SNPs, in the form of a third nucleotide substitution—creating a tri-allelic SNP where the allele-3 is either very rare or linked to a specific population group and hitherto undetected. Cataloguing rare third nucleotide substitutions is an important MPS quality control step, as it forewarns the user that the detected nucleotide at the target site is not a spurious sequencing result, but has been detected as a low frequency allele. We confined out searches for flanking SNPs within an arbitrary +/− 25 nucleotides of the target SNP position.

3. Results and Discussion

3.1. Ancestry Informative SNPs in BT

One hundred binary SNPs were compiled into the final BT ancestry panel, comprising 7 SNPs informative for sub-Saharan African populations, 15 European, 16 East Asian, 20 South Asian, 13 Oceanian and 17 American populations. The extra 12 ancestry SNPs of the total 100 loci were selected as they showed a high level of differentiation between European and other Eurasian population groups, namely East and South Asian and Middle Eastern populations. Table 1 outlines the source of each of these binary SNPs from the previous forensic AIMs panel studies described in Section 2.1 [8,9,21–24], with just one

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

OCE-informative SNP added from a separate study that developed the *Pacifiplex* panel [11]. Over a quarter of SNPs (28 of 100) have been selected as forensic AIMs in multiple panels.

Balancing population specific divergence (PSD) values (excluding the 12 Eurasian divergent SNPs) was successful to a large extent and the cumulative PSD values for each of the six population groups are shown in Figure 1 (individual values per SNP in Table S1). AFR, EUR, OCE and AMR were well balanced with cumulative PSD values of 13.54, 13.30, 11.99 and 11.8, respectively. EAS with a cumulative value of 9.43 did not have enough informative SNPs to reach these levels of divergence, while SAS, with much less differentiation from EUR and EAS, only reached a cumulative value of 2.7. It is noteworthy that two BT appearance SNPs, rs6437783 (in *MYH15* gene) and rs1800414 (*OCA2*), shown in Table 1, had been selected for previous forensic ancestry panels as an EAS-informative SNP in gAIMs and Kiddlab AIM sets, respectively. Given the lower cumulative PSD value for EAS compared to other population groups, these two markers would be worth considering as joint ancestry-appearance SNPs along with the three already used in this way.

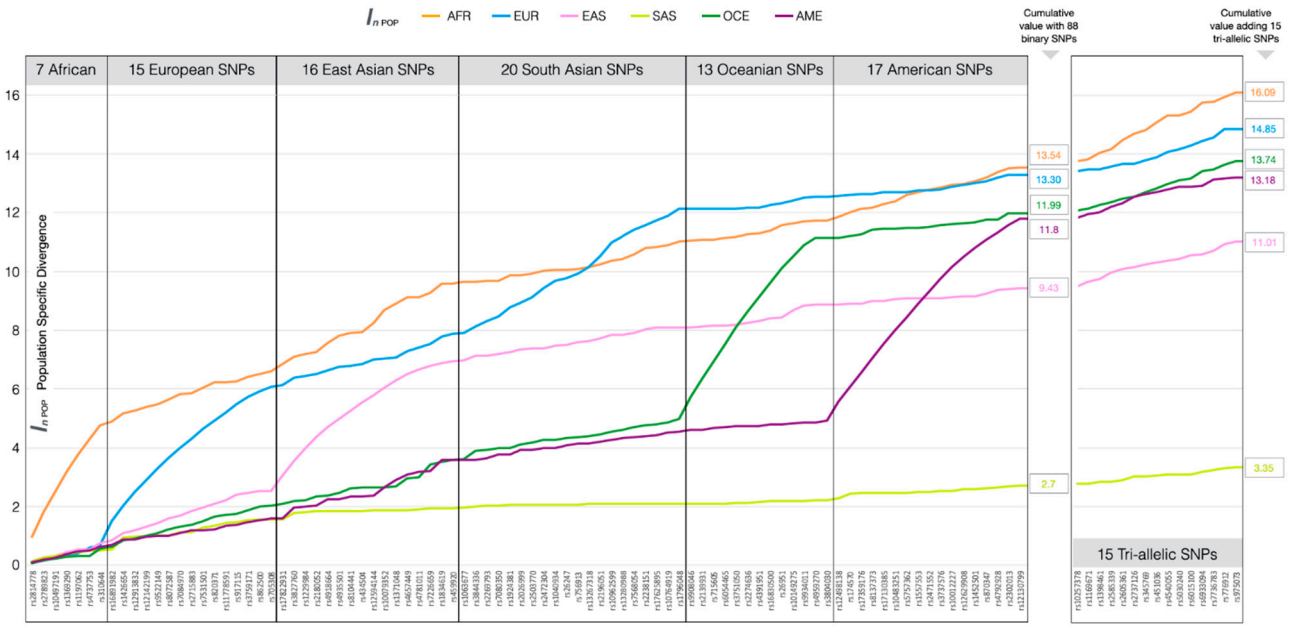


Figure 1. Accumulating population specific divergence values for each of the main population groups and the six sets of SNPs informative for their differentiation. The cumulative values are shown obtained from 88 binary SNPs (i.e., excluding 12 Eurasian-informative SNPs) and from the addition of the 15 tri-allelic SNPs in BT (VISAGE Basic Tool).

From a candidate list of 20 tri-allelic SNPs, rs11238577, rs11990024, rs9845503, rs7853487 and rs809540 were rejected from the assay design, bringing the total number of AIMs in BT to 115, including the three shared appearance SNPs. When the 15 tri-allelic SNPs were added to the cumulative PSD plot of Figure 1 the balance was marginally improved for EAS with a cumulative PSD value of 11.01 closer to the average value for the other population groups.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Table 1. Commonality between SNPs selected for the BT ancestry panel and established forensic ancestry panels: Kiddlab 56; Thermo Fisher Precision ID Ancestry Panel (PIAP); Euroforgen Global AIMs (gAIMs); the LACE panel; the NAME panel; Eurasiaplex; and Shriver et al. US admixture mapping panel [24]. The two markers lower right are BT appearance SNPs previously selected as AIMs.

No.	Pop.	SNP	KK/PIAP	gAIMs	LACE	Other
1	AFR	rs10497191	Kiddlab	-	LACE	
2	AFR	rs1197062	-	gAIMs	LACE	
3	AFR	rs1369290	-	gAIMs	-	
4	AFR	rs2789823	-	gAIMs	-	
5	AFR	rs2814778	Kiddlab	gAIMs	-	
6	AFR	rs310644	Kiddlab	gAIMs	-	
7	AFR	rs4737753	-	-	-	NAME
1	EUR	rs11778591	-	gAIMs	LACE	
2	EUR	rs12142199	-	gAIMs	-	
3	EUR	rs12913832	Kiddlab	gAIMs	-	
4	EUR	rs1426654	Kiddlab	gAIMs	-	
5	EUR	rs16891982	Kiddlab	gAIMs	-	
6	EUR	rs2715883	-	gAIMs	-	
7	EUR	rs3759171	-	gAIMs	LACE	
8	EUR	rs705308	PIAP	-	-	
9	EUR	rs7084970	-	gAIMs	-	
10	EUR	rs7531501	-	gAIMs	-	
11	EUR	rs8072587	-	gAIMs	-	
12	EUR	rs820371	-	gAIMs	LACE	
13	EUR	rs862500	-	gAIMs	LACE	
14	EUR	rs917115	Kiddlab	gAIMs	-	
15	EUR	rs9522149	Kiddlab	gAIMs	-	
1	OCE	rs10149275	-	gAIMs	-	
2	OCE	rs16830500	-	gAIMs	-	
3	OCE	rs2139931	-	gAIMs	-	
4	OCE	rs2274636	-	gAIMs	-	
5	OCE	rs26951	-	gAIMs	-	
6	OCE	rs3751050	-	gAIMs	-	
7	OCE	rs3804030	-	gAIMs	-	
8	OCE	rs4391951	-	gAIMs	-	
9	OCE	rs4959270	-	x	-	Pacifiplex *
10	OCE	rs6054465	-	gAIMs	-	
11	OCE	rs715605	-	gAIMs	-	
12	OCE	rs9908046	-	gAIMs	-	
13	OCE	rs9934011	-	gAIMs	-	

No.	Pop.	SNP	KK/PIAP	gAIMs	LACE	Other
1	AME	rs10012227	-	gAIMs	-	
2	AME	rs10483251	-	gAIMs	LACE	
3	AME	rs12130799	PIAP	-	-	
4	AME	rs12498138	Kiddlab	gAIMs	-	
5	AME	rs12629908	PIAP	-	-	
6	AME	rs1452501	-	gAIMs	LACE	
7	AME	rs1557553	-	gAIMs	LACE	
8	AME	rs17130385	-	gAIMs	LACE	
9	AME	rs17359176	-	gAIMs	LACE	
10	AME	rs174570	Kiddlab	gAIMs	LACE	
11	AME	rs2302013	-	gAIMs	-	
12	AME	rs2471552	-	gAIMs	-	
13	AME	rs3737576	Kiddlab	-	-	
14	AME	rs4792928	-	gAIMs	-	
15	AME	rs5757362	-	-	LACE	
16	AME	rs8137373	-	gAIMs	-	
17	AME	rs870347	Kiddlab	-	-	
1	EAS	rs10079352	-	gAIMs	LACE	
2	EAS	rs1229984	Kiddlab	gAIMs	-	
3	EAS	rs12594144	-	gAIMs	-	
4	EAS	rs1371048	-	gAIMs	-	
5	EAS	rs17822931	-	gAIMs	-	
6	EAS	rs1834619	Kiddlab	gAIMs	-	
7	EAS	rs2180052	-	gAIMs	-	
8	EAS	rs3827760	Kiddlab	gAIMs	-	
9	EAS	rs434504	-	gAIMs	-	
10	EAS	rs459920	Kiddlab	-	-	
11	EAS	rs4657449	-	gAIMs	-	
12	EAS	rs4781011	PIAP	-	-	
13	EAS	rs4918664	Kiddlab	gAIMs	-	
14	EAS	rs4935501	-	gAIMs	-	
15	EAS	rs7226659	Kiddlab	-	-	
16	EAS	rs8104441	-	gAIMs	-	

No.	Pop.	SNP	KK/PIAP	gAIMs	LACE	Other
1	Eurasia	rs1495085	-	-	-	NAME
2	Eurasia	rs1757928	-	-	-	NAME
3	Eurasia	rs2337024	-	-	-	NAME
4	Eurasia	rs6989963	-	-	-	NAME
5	Eurasia	rs6990312	Kiddlab	-	-	
6	Eurasia	rs7148809	-	-	-	NAME
7	Eurasia	rs12203115	-	-	-	NAME
8	Eurasia	rs2227203	-	-	-	NAME
9	Eurasia	rs39897	-	-	-	Eurasiaplex
10	Eurasia	rs4308478	-	-	-	NAME
11	Eurasia	rs7570971	-	-	-	NAME
12	Eurasia	rs984038	-	-	-	NAME
1	SAS	rs1040934	-	-	-	Shriver
2	SAS	rs1063677	-	-	-	Shriver
3	SAS	rs10764919	-	-	LACE	
4	SAS	rs10962599	-	-	-	Eurasiaplex
5	SAS	rs13267318	-	-	-	Shriver
6	SAS	rs13280988	-	-	LACE	
7	SAS	rs17625895	-	-	-	Eurasiaplex
8	SAS	rs1796048	-	-	-	Shriver
9	SAS	rs1924381	-	gAIMs	LACE	
10	SAS	rs2026999	-	-	-	Shriver
11	SAS	rs2196051	Kiddlab	-	-	Eurasiaplex
12	SAS	rs2238151	Kiddlab	-	-	
13	SAS	rs2269793	PIAP	-	-	
14	SAS	rs2472304	-	-	-	Eurasiaplex
15	SAS	rs2503770	-	gAIMs	LACE	
16	SAS	rs26247	-	-	-	Shriver
17	SAS	rs3844336	-	-	-	Shriver
18	SAS	rs7080350	-	-	LACE	
19	SAS	rs7568054	-	-	LACE	
20	SAS	rs756913	-	-	-	Eurasiaplex
MYH15 °	EAS	rs6437783	-	gAIMs	-	-
OCA2 °	EAS	rs1800414	Kiddlab	-	-	-

* A single SNP developed for the Pacifiplex panel was not incorporated into gAIMs. ° Gene locations of the two BT appearance SNPs previously selected for other ancestry panels.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Allele frequency estimates of the 115 BT AIMs based on the standardised reference grid outlined in Section 2.2 (YRI, CEU, CHB, GIH, plus HGDP-CEPH Americans and Oceanians) are given in File S1. Middle East allele frequency estimates are included based on 134 HGDP-CEPH samples from the Israeli Bedouin, Druze and Palestinian populations. Figure S1A–G summarise in simple bar charts the allelic variation in each binary AIM, arranged in sets of the AIMs informative for each population group and ranked by a simple delta allele frequency differential calculation comparing the informative-allele frequency in each target population against the average frequency in the other populations (except SAS-informative SNPs, ranked by SAS-EUR delta values). Figure S1G shows the allele frequencies of 12 Eurasian-informative SNPs in the relevant population groups. For comparison purposes we included North African (NAF) allele frequency estimates based on HGDP-CEPH Mozabite from Algeria, and delta values were estimated for EUR compared to a NAF-ME-SAS average).

3.2. Compilation of Reference and Test Population Datasets

File S2 contains all the Excel worksheets necessary for generating Bayes likelihood calculations and linked PCA analyses by uploading the BT ancestry SNP genotype data to *Snipper*. The active worksheet of reference and query SNP profiles uploaded to *Snipper* is in position 1 in Excel and has the label '1' applied to all reference population profiles in the rightmost column, and '0' to unknown profiles such as those compiled from a forensic sample. The values in cells A1, B1, C1 provide numbers of reference/query profiles, markers, and populations in the worksheet, respectively. All other SNP profiles compiled in the six supporting worksheets can be treated as test samples (i.e., marked with a '0'), or used as alternative or additional reference population datasets (e.g., the HGDP-CEPH or VISAGE in-house ME population genotypes can be added to the reference profiles as a seventh population group).

3.3. Genotyping Concordance

3.3.1. HGDP-CEPH in-House BT Genotypes vs. 1000 Genomes Whole-Genome-Sequence Data

File S3A lists all the pairwise genotype comparisons between in-house data for HGDP-CEPH samples and the data from 1000 Genomes published in 2020 [16]. This file examines all SNPs in BT except for one appearance marker, rs796296176, which did not have genotype data from the in-house analyses of HGDP-CEPH samples. Five HGDP-CEPH Middle East samples had excessive numbers of data gaps from in-house sequencing and were not included in genotype comparisons. Of the 153 total markers in BT, two ancestry SNPs had higher than average discordancy rates, indicating a problem with genotyping either from 1000 Genomes sequence analysis, or the BT MPS assay. First, tri-allelic SNP rs2737126 had 15/277 (5.4%) genotype discordancy, seemingly arising from failure to detect a G allele, when present as a GT or CG heterozygote, but not for all such genotypes. This phenomenon was not explained by detailed scrutiny of sequence data at this variant site. Second, the binary SNP rs3804030 had a disproportionately high genotype discordancy rate of 29/277 (10.47%), which appeared to be due to a failure to detect the rare, OCE-indicative C allele outside of Oceanian populations examined by 1000 Genomes. Interestingly, no discordances were seen for this SNP amongst pairwise comparisons of 27 HGDP-CEPH Oceanians.

File S3A shows that for the 151 SNPs of BT without apparent problems, the genotyping concordance was very high, reaching 99.96% concordance with 17 discordances largely appearing at random across the data grid, rather than systematically as detected in the above two SNPs. This high level of genotype concordance between the BT assay developed vs. 1000 Genomes whole genome sequence-based variant detection indicates the robustness of the MPS method and the BT multiplex developed by VISAGE. The no-call rate (unreported genotypes) was also low at 0.2% for 1000 Genomes, and 0.6% for our MPS genotyping.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.3.2. 1000 Genomes Phase 3 Genotypes vs. 1000 Genomes High Sequence Coverage Genotypes

File S3B gives the pairwise comparisons for all BT SNPs between the current 1000 Genomes Phase 3 genotypes listed online by Ensembl, and the genotypes published following high coverage sequence analysis of the same samples by the New York Genome Centre [17]. As we have relied extensively on the online genotype data produced by 1000 Genomes Phase 3 for several years, the cross-checks of genotype calls comparing 30X and 2-3X sequence coverage that this concordance analysis provides are important. Amongst all 153 BT SNPs, 73 had zero genotype discordances and 66 showed largely inconsequential discordances numbering 1–5 genotype differences between each sequencing dataset (a locus-specific concordance of 99.96–99.8%). However, 14 SNPs had higher levels of six or more discordances, comprising 11 ancestry SNPs. Three ancestry SNPs: rs2789823, rs9522149 and rs2196051 had concordance rates $\leq$ 98.2% due to 25, 27 and 46 discordances, respectively. The BT ancestry SNPs genotype discordances of two or more in number are summarised in Figure 2. It can be assumed that higher sequence coverage provides more secure genotyping for a small, but significant proportion of SNPs in the 1000 Genomes variant database, of which 14 with higher-than-average levels of discordancy are part of BT. For this reason, all reference population data compiled from 1000 Genomes for VISAGE ancestry analyses with BT are based on the high coverage (30X) genotype datasets, which have yet to be made available in the Ensemble variant browser for 1000 Genomes.

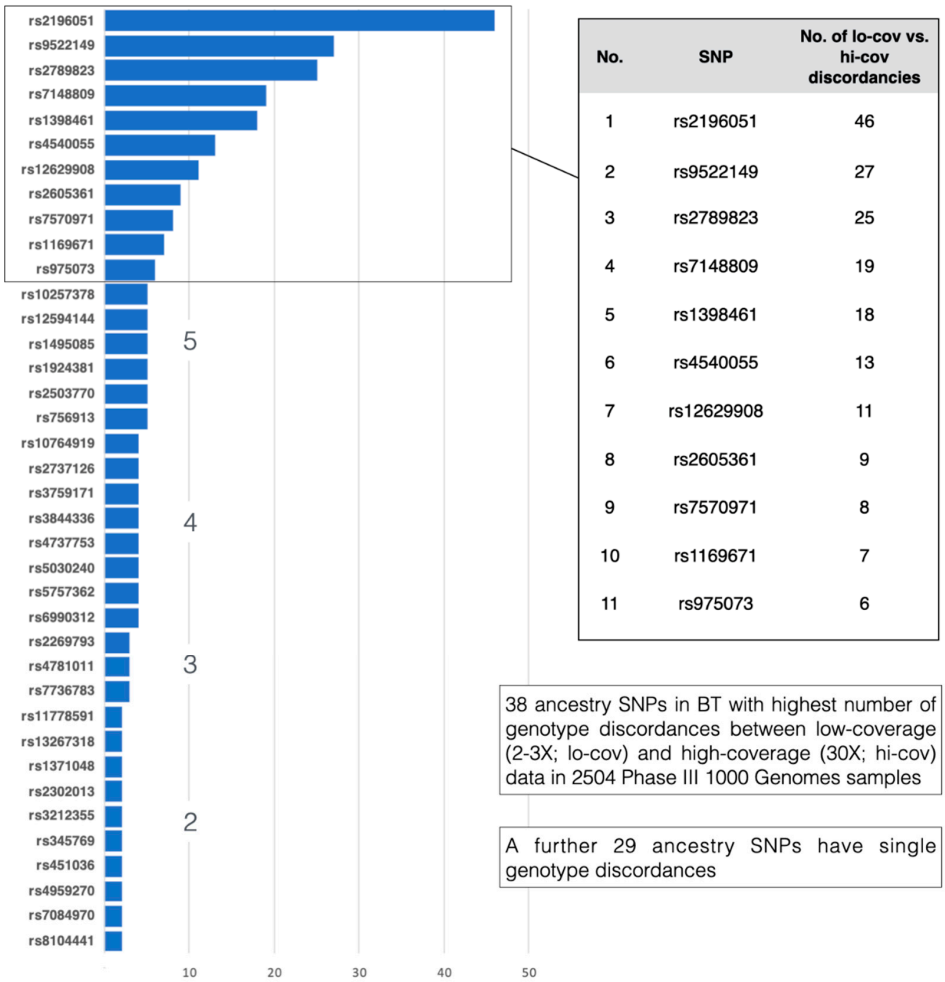


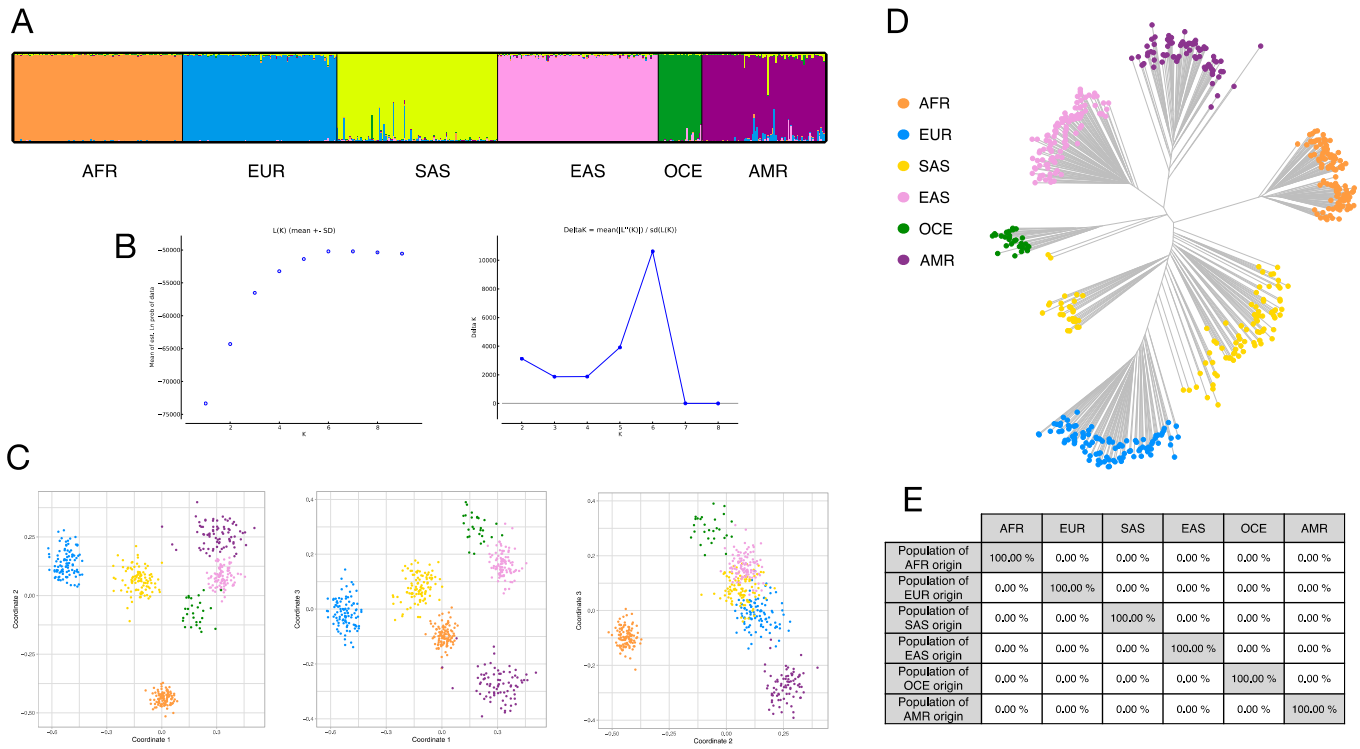
Figure 2. The 38 BT ancestry SNPs with the highest numbers of discordant genotypes between 1000 Genomes low coverage vs. high coverage sequence data. An additional 29 had only one discordant genotype. The full set of genotype comparisons for all BT SNPs and all 1000 Genomes samples are given in File S3B.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.4. Ancestry Inference Efficiency of the BT Ancestry SNPs

One-out cross validation of the reference dataset gave 100% correct classification in all cases, i.e., all 520 reference profiles were successfully assigned to their true population of origin. The likelihood ratio values produced by cross validation were consistently high and most samples gave likelihoods greater than 'The profile is more than 1 billion times more likely to be from the true population of origin than another population' (data not shown).

Figure 3A shows the STRUCTURE analysis K:6 cluster plots for the reference populations (1–6) in worksheet 1 of File S1. Figure 3B shows the accompanying probability of data and Delta K statistical analyses indicating an optimum cluster number of 6 for these analyses. Figure 3C shows colour-matched MDS plots for the same samples for coordinate 1 vs. 2 (left plot), 1 vs. 3 (mid) and 2 vs. 3. Only a small degree of overlap between the OCE and EAS MDS clusters-of-points remains when comparing all three coordinate plots. Figure 3D shows the Neighbour Joining Tree (NJT) distance matrix results, which indicate a clear arrangement of branch lengths and spatial separation, apart from two SAS outliers close to the OCE set. The summary of cross-validation classification results in Figure 3E completes the ancestry analysis of these sample sets.



	AFR	EUR	SAS	EAS	OCE	AMR
Population of AFR origin	100.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Population of EUR origin	0.00 %	100.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Population of SAS origin	0.00 %	0.00 %	100.00 %	0.00 %	0.00 %	0.00 %
Population of EAS origin	0.00 %	0.00 %	0.00 %	100.00 %	0.00 %	0.00 %
Population of OCE origin	0.00 %	0.00 %	0.00 %	0.00 %	100.00 %	0.00 %
Population of AMR origin	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	100.00 %

Figure 3. Ancestry analysis with 115 BT ancestry SNPs of the six populations of the standardised reference dataset (AFR: brown; EUR: blue; SAS: yellow; EAS: pink; OCE: green; AMR: purple), with consistent colours across three statistical analyses of A, C and D. (**A**) STRUCTURE cluster membership proportions at K = 6, indicated to be the optimum K number of genetic clusters by: (**B**) the mean L(K) (log probability of data) and ΔK plots from STRUCTURE runs, following the analyses of Evanno et al. [31]. (**C**) Multi-dimensional scaling (MDS) analysis showing principal component (PC) 1 vs. PC2 coordinates, PC1 vs. PC3 and PC2 vs. PC3 two-dimensional plots. (**D**) Neighbour joining tree (NJT) analysis. (**E**) Summary classification success table of cross validation of the standardised reference set.

Figure S2.1 summarises the same type of population analyses as Figure 3 applied to unadmixed (8–11, 14–25) and admixed (12–13, 26–29) 1000 Genomes samples (30X coverage genotypes) analysed against the reference population set. In nearly all cases, unadmixed samples show low levels of joint cluster membership. The admixed African populations give similar patterns to those found in other studies (but see Section 3.5) with European co-ancestry ranging from 5–30% K2 cluster membership. The admixed

American samples from 1000 Genomes also show the expected proportions from K:2 and K:6 in the STRUCTURE analysis, with K:1 cluster membership forming the third African co-ancestry component, most evident in Puerto Ricans. The carefully curated Peruvian reference Native American samples are clearly discernible on the left of the columns of the K6 cluster. Equally consistently, admixed samples occupy the middle ground between the MDS clusters-of-points formed by the unadmixed samples, depending on the degree of their separation in coordinates 1 vs. 3 or 1 vs. 2 MDS plots, particularly the admixed Africans in 1 vs. 2.

Figure S2.2–S2.4 show identical analysis output and using the same reference populations for HGDP-CEPH samples, SGDP-VISAGE in-house samples and EGDP samples, respectively. The HGDP-CEPH samples show a degree of joint cluster membership proportions in regions whose populations occupy continental margins, e.g., Mozabite show African-European co-ancestry patterns (joint membership proportions) and Pakistani populations show some European co-ancestry. The SGDP and EGDP sample analyses are the most varied, reflecting the more widespread sampling locations and emphasis on continental margin regions such as Central South Asia. The SGDP-EGDP Oceanian and American samples give clear, unequivocal cluster memberships, underlining the careful choice of informative SNPs for these population groups. However, it should be noted that BT and all other equivalent MPS ancestry panels show Oceanian ancestry to be only a small proportion of the STRUCTURE cluster patterns in the single SGDP Maori and Hawaiian samples, indicating that AIMs selected from analysis of Melanesians (i.e., HGDP-CEPH Papuans) are not sufficiently informative to infer Polynesian ancestry.

3.5. Co-Ancestry Analysis of Known Admixed Individuals in 1000 Genomes with BT Ancestry SNPs

3.5.1. Comparison of Co-Ancestry Patterns Obtained with the Human Origins Panel and BT

Figure 4 shows a direct pairwise comparison of the co-ancestry patterns (cluster membership proportions) of 504 admixed samples from 1000 Genomes (see File S1C), analysed in combination with the standardised reference set with the >572,000 Human Origins array SNPs (using ADMIXTURE due to SNP data density) and the 115 ancestry SNPs of BT (using STRUCTURE). File S4 gives the genetic cluster proportions from both analyses on which the Figure 4 cluster plots are based. These plots show a very good match of co-ancestry patterns in the two African admixed ACB and ASW populations. The correlation analysis and r^2 values aligned below the cluster plots in Figure 4 indicate high levels of correlation between Origins and BT co-ancestry proportion estimates for ACB and ASW, notably for the AFR proportions (although they are combined with AMR co-ancestry in these assessments). The EUR co-ancestry proportions tend to be slightly underestimated by BT SNPs or go undetected when very small. The four admixed populations with AMR co-ancestry are arranged in descending overall proportions of this co-ancestry and all show the same tendency for BT SNPs to underestimate the EUR proportions, but the PEL, MXL and CLM patterns still show a good match and relatively high r^2 values for both main EUR and AMR co-ancestries. When three co-ancestries are detected, the BT SNPs have less precision to detect AMR and AFR co-ancestry at levels below 25%, particularly evident in the PUR non-EUR co-ancestry proportion estimates. Overall, the patterns match very well for simple two-way co-ancestry across most ratios, with BT only failing to detect very minor co-ancestry when it falls below 5%. The match of patterns only tends to reduce when 3-way co-ancestry is more prevalent, notably when the two minor co-ancestries are both below 10%. BT SNPs can detect two-way co-ancestry using STRUCTURE with sufficient accuracy for proportions to be reliably reported down to about 10% minor co-ancestry, although when three-way co-ancestry is detected, more caution would be required in reporting minor ratios below this 10% threshold and the lack of accuracy in such cases should be highlighted in the report.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

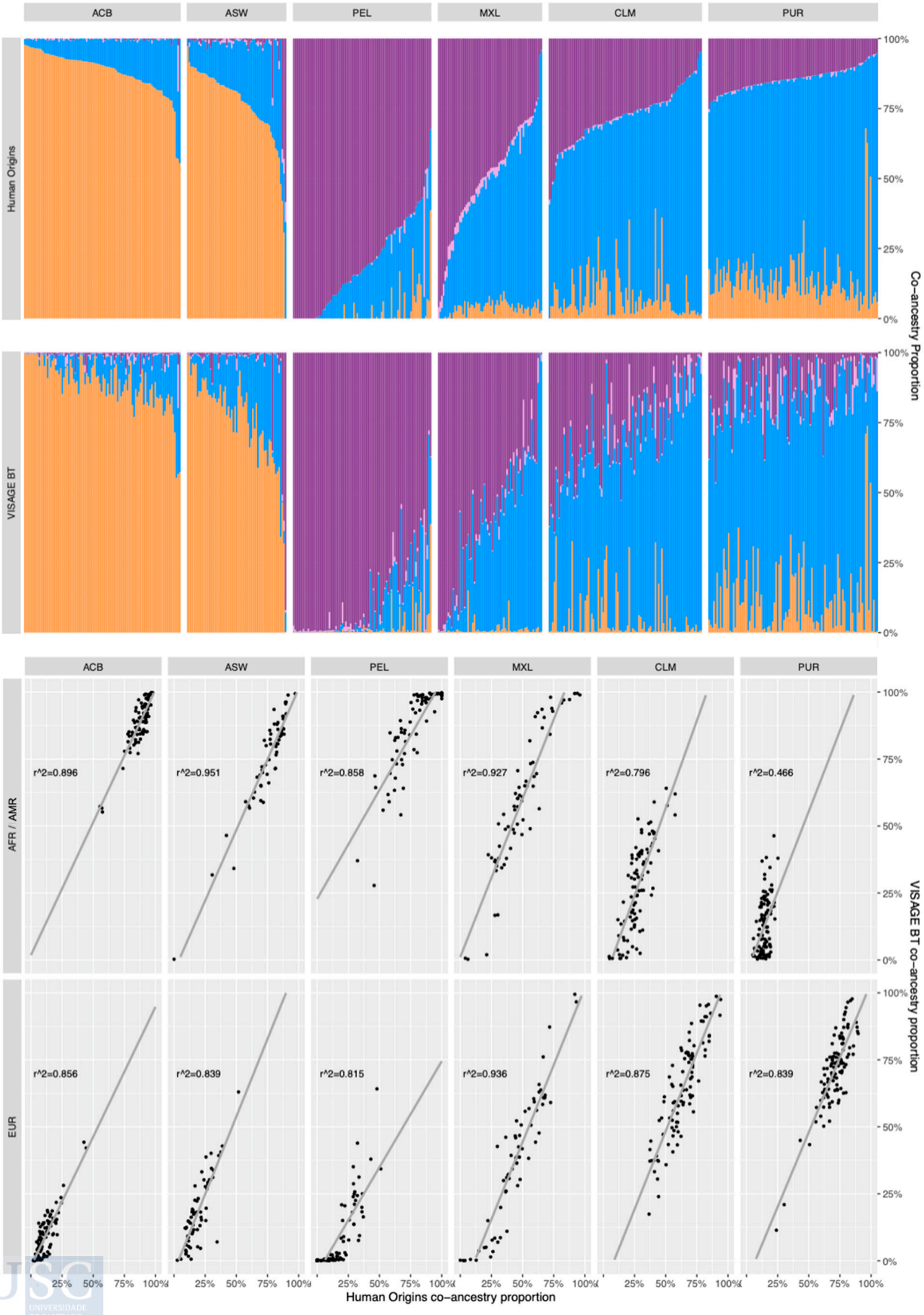


Figure 4. Pairwise comparison of individual co-ancestry proportions (cluster membership proportions) in 504 admixed samples from 1000 Genomes, analysed using the standardised reference set with Human Origins array, comprising

>572,000 SNPs (using ADMIXTURE) and 115 VISAGE BT ancestry SNPs (using STRUCTURE). Co-ancestry proportions in four genetic clusters representing, African, European, Native American and East Asian, ancestries are given in File S4. The r^2 correlation analysis plots of each admixed population group are shown below the corresponding cluster patterns.

3.5.2. Using GDA as a Simple Snipper-Based Evaluation of Admixture

From the STRUCTURE analysis of the BT SNPs, co-ancestry proportion estimates were compared to equivalent proportions provided by GDA in *Snipper*. This intended to test the efficiency of GDA to detect admixture and alert the user to the need to perform STRUCTURE analysis, if this initial analysis of the sample data does not suggest straightforward unadmixed ancestry. Figure S3A shows cluster plots for the STRUCTURE run and for the co-ancestry proportions based on GDA values, with r^2 plots below arranged as in Figure 4. The r^2 values indicate a good match between each approach, but GDA underestimates the EUR co-ancestry to a significant degree in ACB and ASW, to the point where GDA did not detect this co-ancestry in a large proportion of individuals, and it was underestimated by more than 50% of the true value (from Human Origins SNP array data) in most of the others. The trend of GDA underestimating European co-ancestry continues across all the PEL, MXL, CLM and PUR samples, where additionally, AFR co-ancestry is consistently overestimated in each population. The EUR underestimation is partly explained by detection of significant proportions of EAS co-ancestry by GDA, likely to be falsely created by a close relationship between AMR and EAS population groups, which STRUCTURE differentiates more efficiently as distinct genetic clusters compared to GDA's measure of distance between EUR and both AMR and EAS.

3.5.3. Comparisons of 10-Percentile Co-Ancestry Ratio Patterns

A final visual comparison of co-ancestry patterns was made by plotting the mean co-ancestry ratios per 10-percentile group of individuals to smooth the data and allow easier assessment of trends in both the use of the small-scale BT ancestry panel to detect co-ancestry and GDA as an initial measure of the presence of admixture. Mean STRUCTURE co-ancestry proportions obtained with Origins SNPs and BT SNPs, plus GDA proportions, are plotted in Figure S3B. When admixture patterns are summarised into percentile groups, it is much easier to see the very close match in estimated co-ancestry ratios between Origins and BT SNP-based STRUCTURE runs. This is particularly evident in the 10th ASW percentile with a three-way co-ancestry pattern, and the 10th PEL percentile with a fourth EAS co-ancestry. However, co-ancestry proportions below 10% regularly fail to be detected by BT (1st ACB percentile, 3rd–5th PEL, 1st–2nd MXL, 9th–10th CLM percentiles). Therefore, it remains prudent to recognise that BT will not consistently detect co-ancestry ratios $\leq$10%. The presence of EAS co-ancestry at very small proportions in many of the MXL-CLM-PUR BT percentile groups indicates this component should be ignored when it is less than 10% and AMR co-ancestry is present above this ratio at the same time. The percentile group patterns in GDA emphasise that it is not an efficient technique for detecting admixture accurately when co-ancestral ratios are below 20%, so it does not provide an accurate measure of individuals with these proportions and STRUCTURE would be required in all cases, particularly as the bulk of the PEL-MXL-CLM percentile groups show spurious EAS co-ancestry ratios above this threshold value).

3.6. Audit of Additional Variation in BT SNPs from the GnomAD 3.1 Database

File S5 lists the co-amplified flanking SNPs and extra third-substitution alleles amongst both ancestry and appearance binary BT SNPs detected in the gnomAD 3.1 database. Twelve ancestry SNPs and three appearance SNPs had a third substitution allele, which reached the highest allele frequency of only 0.0002 in rs2302013 and rs917115 (both in East Asians). Ancestry SNP rs1063677 and appearance SNP rs12821256 showed all four substitution alleles, with the highest frequency for allele T of 0.0005 in Europeans in the GC SNP rs1063677.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

The power of gnomAD to detect low frequency variants across the genome enabled sequence screens that indicated a large number of flanking SNPs within 25 nucleotides of the target site of many BT SNPs. Amongst the 115 ancestry SNPs, 49 had flanking SNPs, while in the 41 appearance (EVC) SNPs, there were 23, notably markers targeted in the SNP-rich MC1R gene. Four ancestry SNPs showed informative combinations of alleles in close proximity which show unlinked allele frequencies between the two loci—in other words, forming population-informative 2-SNP microhaplotypes (i.e., diplotypes). These four are detailed in Figure S4. The Chr-3 pair rs1169671-rs76444263 (target SNP in bold), Chr-8 pair rs75064826-rs13280988 and Chr-19 pair rs2337025-rs2337024 all show SAS-informative diplotypes, while the Chr-20 pair rs4809578-rs310644 has a range of diplotypes, which are most differentiating for AFR and EAS population groups. Microhaplotype rs1169671-rs76444263 combines a tri-allelic and binary SNP, but the latter has low levels of polymorphism so is less informative than the other binary SNP pairs. Diplotype SNP allele combinations can be analysed in *Snipper* by assigning A, C, G, T allele nomenclature to the alphabetic arrangement of diplotypes, which, for the above binary SNP pairs, would be A=AA, C=AG, G=CA, T=CG combinations; A=AA, C=AG, G=GA, T=GG combinations; and A=AC, C=AT, G=TC, T=TT combinations, respectively. While the phase of each SNP pair's alleles needs to be detected from the sequence data to do this, it indicates the potential extra variant information, which is accessible from MPS analysis of forensic SNPs.

4. Concluding Remarks

The primary aim of the VISAGE Consortium of creating an all-in-one forensic intelligence tool based on MPS technology requires the development of a full analytical pipeline. This should include population reference data and predictive algorithms for the inference of age, ancestry and appearance characteristics associated with the genotyped SNPs, plus an optimum sequencing protocol which is applicable to the two main forensic MPS platforms in current use. The strategy of starting with a prototype tool of simple age predictors in one multiplex, and established pigmentation phenotype and continental ancestry predictors in the second, allowed us to build a strong foundation on which to develop the more ambitious VISAGE Enhanced Tool (ET). ET represents substantial expansion of all three 'AAA' predictive systems in the same parallel MPS multiplex setup as BT (i.e., separate DNA workflows dedicated to methylation analysis and SNP genotyping). The ancestry panel of BT therefore took the best ancestry informative SNPs from several well-established forensic ancestry panels, with particular regard to known MPS sequence reliability as well as optimum ancestry informativeness. For this reason, BT has provided very good sequencing performance with forensic DNA [29,37], and enables ancestry inferences to be made for populations of continental and South Asian origin with the highest levels of statistical likelihood.

Although we dedicated considerable resources to genotyping BT SNPs in populations addressing several geographic gaps in coverage left by large-scale genomic sequencing projects, many global regions remain to be characterised with new studies. The dynamic system in *Snipper*, i.e., compiling the largest possible array of population data and allowing the end-user to select which samples extend or modify the standardised reference set, provides flexibility for statistical analyses of BT SNP data. It also allows new population data to be uploaded to the webpages holding reference sets (http://mathgene.usc.es/snipper/forensic_mps_aims.html, accessed on 20 August 2021) and the additional populations shown in File S1. Although genetic cluster analyses using STRUCTURE can be challenging for forensic practitioners analysing population data for the first time, in our view, it is the best way to detect and analyse individual co-ancestry in persons with admixed backgrounds. The comparisons of GDA and STRUCTURE made in this study indicate that GDA is not accurate enough to evaluate low level co-ancestry when admixture patterns are complex and an individual has more than two co-ancestries. However, when the BT ancestry SNPs are compared to the much larger panel of Human Origins SNPs, the level of detail obtained using STRUCTURE is well matched and indicates

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

BT SNPs are informative enough to accurately detect co-ancestral components present above 10% proportions.

Lastly, the care taken in selecting the BT ancestry SNPs based on their recorded reliability and sensitivity in previous MPS-based forensic ancestry panels [8,12,18,20,23] has helped to ensure predictable performance going forward with evaluation of both forensic MPS platforms. This extends to the sequencing chemistry, library preparation and sequence read balance (between strands, between alleles, within loci, between loci) observed so far in the forensic validations the VISAGE Consortium has made following SWGDAM guidelines for newly developed forensic DNA techniques (AmpliSeq chemistry [29], PowerSeq chemistry [37], and ForenSeq chemistry with a manuscript in submission). These same principals have been applied to the development of the ET ancestry panel, providing the means to increase the overall multiplex size of the tool and extending the range of populations among which ET ancestry SNPs can differentiate.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12081284/s1, Files S1: Allele frequency estimates in seven population groups for the 115 ancestry SNPs of BT; S2A: Snipper Reference Grid; S2B: 1KG Unadmixed Populations; S2C: 1KG Admixed Populations; S2D: HGDP-CEPH Populations; S2E: VISAGE In-house Populations; S2F: SGDP Test Samples; S2G: EGDP Test Samples, S3A: Genotype comparisons of 1000 Genomes HGDP-CEPH data (red) vs VISAGE in-house HGDP-CEPH data (black); S3B: Genotype comparisons of 1000 Genomes Phase III low-coverage data based on 2-3X sequence coverage (black) vs. 1000 Genomes high coverage data based on New York Genome Centre's 30X resequencing of the same samples (blue); S4: Genetic cluster membership proportions from the four major contributing populations to American admixture with 572K SNPs of Human Origins array analyses using ADMIXTURE vs 115 BT SNPs using STRUCTURE, for 1511 unadmixed and 504 admixed 1000 Genomes samples; S5: Screen of all 153 BT SNPs of gnomAD genome database information for current GRCh37-38 positions, gene identifiers, third alleles and high frequency flanking SNPs (present on the same sequence strand as the target site within +/− 25 nucleotides). Figures S1A: 7 African-informative SNPs; S1B: 15 European-informative SNPs; S1C: 16 East Asian-informative SNPs; S1D: 20 South Asian-informative SNPs; S1E: 13 Oceanian-informative SNPs; S1F: 17 American-informative SNPs; S1G: 12 Eurasian divergent SNPs, S2.1: Bio-geographical ancestry analysis based on the six reference populations (1–6, summarised in Figure 1) plus unadmixed (8–11, 14–25) and admixed (12–13, 26–29) populations available from 1000 Genomes 30X high coverage data (1KG-30); S2.2: Bio-geographical ancestry analysis based on the six reference populations presented in Figure 1, plus populations available from HGDP-CEPH Whole Genome Sequencing Data; S2.3: Bio-geographical ancestry analysis based on the six reference populations presented in Figure 1, plus populations available from in-house genotyping with the VISAGE BT tool and SGDP data; S2.4: Bio-geographical ancestry analysis based on the six reference populations presented in Figure 1, plus populations available from EGDP data, with ~85% (95/115 SNPs) profile completeness. S3.1: (A) Co-ancestry proportions sample-by-sample in 504 1000 Genomes admixed individuals based on STRUCTURE and GDA analyses. (B) Correlation analyses and r2 values comparing the combined African-American and European co-ancestry proportion estimates with both techniques. S3.2: Mean co-ancestry proportions summarised as 10-percentile columns (10 bins of 10% of each set of individuals), of varyingnumbers per bin: ACB: 10, ACB: 10, ASW: 6, PEL: 8, ACB: 10, MXL: 6, CLM: 9, PUR: 10. S4: Four BT ancestry SNP target sites and population informative flanking SNPs within 25 nucleotides (nt) on the 5′ or 3′ side. Table S1: Individual and cumulative population specific Divergence (PSD) values for the 115 ancestry SNPs arranged in descending order of PSD value per population.

Author Contributions: Conceptualization, C.P., M.d.l.P., M.K., M.N., W.B., W.P., P.M.S., M.V.L. and J.Á.-D.; Methodology, C.P., J.R-R., A.A.-C., M.d.l.P., J.P.-S., A.M.-M., A.F.-A., C.X., E.Y.Y.C. and T.E.G.; Data Curation, C.P., A.A.-C., J.R.-R., M.d.l.P., A.F.-A., J.Á.-D., J.P.-S., E.Y.Y.C., P.M.S., T.E.G. and A.M.-M.; Writing and Editing Manuscript; C.P., M.d.l.P., M.V.L. and Á.C. All authors have read and agreed to the published version of the manuscript.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Funding: The study was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 740580 within the framework of the VISible Attributes through GEnomics (VISAGE) Project and Consortium. MdlP is supported by a post-doctorate grant funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de

Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481D-2021-008). J.R. is supported by the "Programa de axudas á etapa predoutoral" funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481A-2020/039).

Institutional Review Board Statement: In-house population samples were previously obtained from: i. Moroccan volunteers resident in Madrid, originally collected by the Comisaría General de Policía Científica, Madrid; ii. Eritrean, Somali, Central Iraqi, Kurdish Iraqi, and Turkish volunteers resident in Germany, collected according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of the Faculty of Medicine, University of Cologne, Germany, reference no. 17-416 (dated 16.05.2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All data generated in these studies, except the cross-validation likelihoods, are included in the Files.

Acknowledgments: STRUCTURE and ADMIXTURE runs were performed by the FinisTerrae II supercomputer at the Centro de Supercomputación de Galicia, Santiago de Compostela (CESGA), Spain.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Centres and investigators of the VISible Attributes through GEnomics (VISAGE) Consortium, Website: http://www.visage-h2020.eu/ (accessed on 20 August 2021)

- Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands: Manfred Kayser, Vivian Kalamara, Arwin Ralf, Athina Vidaki
- Jagiellonian University, Krakow, Poland: Wojciech Branicki, Ewelina Pośpiech, Aleksandra Pisarek
- Universidade de Santiago de Compostela, Santiago de Compostela, Spain: Ángel Carracedo, Maria Victoria Lareu, Christopher Phillips, Ana Freire-Aradas, Ana Mosquera-Miguel, María de la Puente
- Medizinische Universität Innsbruck, Innsbruck, Austria: Walther Parson, Catarina Xavier, Antonia Heidegger, Harald Niederstätter
- Universität zu Köln, Cologne, Germany: Michael Nothnagel, Maria-Alexandra Katsara, Tarek Khellaf
- King's College London, London, UK: Barbara Prainsack, Gabrielle Samuel
- Klinikum der Universität zu Köln, Cologne, Germany: Peter M. Schneider, Theresa E. Gross, Jan Fleckhaus, Elaine Cheung.
- Bundeskriminalamt, Wiesbaden, Germany: Ingo Bastisch, Nathalie Schury, Jens Teodoridis, Martina Unterländer
- Institut National de Police Scientifique, Lyon, France: François-Xavier Laurent, Caroline Bouakaze, Yann Chantrel, Anna Delest, Clémence Hollard, Ayhan Ulus, Julien Vannier
- Netherlands Forensic Institute, The Hague, Netherlands: Titia Sijen, Kris van der Gaag, Marina Ventayol-Garcia
- National Forensic Centre, Swedish Police Authority, Linköping, Sweden: Johannes Hedman, Klara Junker, Maja Sidstedt
- Metropolitan Police Service, London, United Kingdom: Shazia Khan, Carole E. Ames, Andrew Revoir
- Centralne Laboratorium Kryminalistyczne Policji, Warsaw, Poland: Magdalena Spólnicka, Ewa Kartasinska, Anna Woźniak

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

References

1. Phillips, C. Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci. Int. Genet.* **2015**, *18*, 49–65. [CrossRef]
2. Kayser, M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci. Int. Genet.* **2015**, *18*, 33–48. [CrossRef] [PubMed]
3. Freire-Aradas, A.; Phillips, C.; Lareu, M.V. Forensic individual age estimation with DNA: From initial approaches to methylation tests. *Forensic Sci. Rev.* **2017**, *29*, 121–144. [PubMed]
4. King, T.E.; Fortes, G.G.; Balaresque, P.; Thomas, M.G.; Balding, D.; Delser, P.M.; Neumann, R.; Parson, W.; Knapp, M.; Walsh, S.; et al. Identification of the remains of King Richard III. *Nat. Commun.* **2014**, *5*, 5631. [CrossRef]
5. Walsh, S.; Liu, F.; Ballantyne, K.N.; van Oven, M.; Lao, O.; Kayser, M. IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci. Int. Genet.* **2011**, *5*, 170–180. [CrossRef]
6. Walsh, S.; Liu, F.; Wollstein, A.; Kovatsi, L.; Ralf, A.; Kosiniak-Kamysz, A.; Branicki, W.; Kayser, M. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci. Int. Genet.* **2013**, *7*, 98–115. [CrossRef] [PubMed]
7. Walsh, S.; Chaitanya, L.; Breslin, K.; Muralidharan, C.; Bronikowska, A.; Pospiech, E.; Koller, J.; Kovatsi, L.; Wollstein, A.; Branicki, W.; et al. Global skin colour prediction from DNA. *Hum. Genet.* **2017**, *136*, 847–863. [CrossRef]
8. Phillips, C.; Parson, W.; Lundsberg, B.; Santos, C.; Freire-Aradas, A.; Torres, M.; Eduardoff, M.; Børsting, C.; Johansen, P.; Fondevila, M.; et al. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. *Forensic Sci. Int. Genet.* **2014**, *11*, 13–25. [CrossRef]
9. Phillips, C.; Freire Aradas, A.; Kriegel, A.K.; Fondevila, M.; Bulbul, O.; Santos, C.; Serrulla Rech, F.; Perez Carceles, M.D.; Carracedo, A.; Schneider, P.M.; et al. *Eurasiaplex*: A forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Sci. Int. Genet.* **2013**, *7*, 359–366. [CrossRef]
10. Carvalho Gontijo, C.; Porras-Hurtado, L.G.; Freire-Aradas, A.; Fondevila, M.; Santos, C.; Salas, A.; Henao, J.; Isaza, C.; Beltrán, L.; Nogueira Silbiger, V.; et al. PIMA: A population informative multiplex for the Americas. *Forensic Sci. Int. Genet.* **2020**, *44*, 102200. [CrossRef]
11. Santos, C.; Phillips, C.; Fondevila, M.; Daniel, R.; van Oorschot, R.A.H.; Burchard, E.G.; Schanfield, M.S.; Souto, L.J.; Uacyisrael, J.; Via, M.; et al. Pacifiplex: An ancestry-informative SNP panel centred on Australia and the Pacific region. *Forensic Sci. Int. Genet.* **2016**, *20*, 71–80. [CrossRef] [PubMed]
12. Phillips, C.; McNevin, D.; Kidd, K.K.; Lagacé, R.; Wootton, S.; de la Puente, M.; Freire-Aradas, A.; Mosquera-Miguel, A.; Eduardoff, M.; Gross, T.; et al. MAPlex-A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. *Forensic Sci. Int. Genet.* **2019**, *42*, 213–226. [CrossRef]
13. Breslin, K.; Wills, B.; Ralf, A.; Ventayol Garcia, M.; Kukla-Bartoszek, M.; Pospiech, E.; Freire-Aradas, A.; Xavier, C.; Ingold, S.; de La Puente, M.; et al. HIrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms. *Forensic Sci. Int. Genet.* **2019**, *43*, 102152. [CrossRef] [PubMed]
14. Al-Asfi, M.; McNevin, D.; Mehta, B.; Power, D.; Gahan, M.E.; Daniel, R. Assessment of the Precision ID Ancestry Panel. *Int. J. Legal. Med.* **2018**, *132*, 1581–1594. [CrossRef]
15. The 1000 Genomes Project Consortium; Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74.
16. Bergström, A.; McCarthy, S.A.; Hui, R.; Almarri, M.A.; Ayub, Q.; Danecek, P.; Chen, Y.; Felkel, S.; Hallast, P.; Kamm, J.; et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **2020**, *367*, 1339–1349. [CrossRef]
17. Byrska-Bishop, M.; Evani, U.S.; Zhao, X.; Basile, A.O.; Abel, H.J.; Regier, A.A.; André Corvelo, A.; Clarke, W.E.; Musunuri, R.; Nagulapalli, K.; et al. High Coverage Whole Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 trios. Available online: https://www.internationalgenome.org/data-portal/data-collection/30x-grch38; http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/ (accessed on 20 June 2021).
18. Xavier, C.; de la Puente, M.; Phillips, C.; Eduardoff, M.; Heidegger, A.; Mosquera-Miguel, A.; Freire-Aradas, A.; Lagace, R.; Wootton, S.; Power, D.; et al. Forensic evaluation of the Asia Pacific ancestry-informative MAPlex assay. *Forensic Sci. Int. Genet.* **2020**, *48*, 102344. [CrossRef]
19. Cheung, E.Y.Y.; Phillips, C.; Eduardoff, M.; Lareu, M.V.; McNevin, D. Performance of ancestry-informative SNP and microhaplotype markers. *Forensic Sci. Int. Genet.* **2019**, *43*, 102141. [CrossRef] [PubMed]
20. Phillips, C.; Amigo, J.; Tillmar, A.O.; Peck, M.A.; de la Puente, M.; Ruiz-Ramírez, J.; Bittner, F.; Idrizbegović, Š.; Wang, Y.; Parsons, T.J.; et al. A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel. *Forensic Sci. Int. Genet.* **2020**, *46*, 102232. [CrossRef]
21. Galanter, J.M.; Fernandez-Lopez, J.C.; Gignoux, C.R.; Barnholtz-Sloan, J.; Fernandez-Rozadilla, C.; Via, M.; Hidalgo-Miranda, A.; Contreras, A.V.; Figueroa, L.U.; Raska, P.; et al. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet.* **2012**, *8*, e1002554. [CrossRef] [PubMed]
22. Kidd, K.K.; Speed, W.C.; Pakstis, A.J.; Furtado, M.R.; Fang, R.; Madbouly, A.; Maiers, M.; Middha, M.; Friedlaender, F.R.; Kidd, J.R. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci. Int. Genet.* **2014**, *10*, 23–32. [CrossRef]
23. Pereira, V.; Freire-Aradas, A.; Ballard, D.; Børsting, C.; Diez, V.; Pruszkowska-Przybylska, P.; Ribeiro, J.; Achakzai, N.M.; Aliferi, A.; Bulbul, O.; et al. Development and validation of the EUROFORGEN NAME (North African and Middle Eastern) ancestry panel. *Forensic Sci. Int. Genet.* **2019**, *42*, 260–267. [CrossRef] [PubMed]

24. Mao, X.; Bigham, A.W.; Mei, R.; Gutierrez, G.; Weiss, K.M.; Brutsaert, T.D.; Leon-Velarde, F.; Moore, L.G.; Vargas, E.; McKeigue, P.M.; et al. A genomewide admixture mapping panel for Hispanic/Latino populations. *Am. J. Hum. Genet* **2007**, *80*, 1171–1178. [CrossRef] [PubMed]
25. Available online: http://mathgene.usc.es/Snipper/ (accessed on 20 June 2021).
26. Phillips, C.; Amigo, J.; McNevin, D.; de la Puente, M.; Cheung, E.Y.Y.; Lareu, M.V. Online population data resources for forensic SNP analysis with Massively Parallel Sequencing: An overview of online population data for forensic purposes. In *Forensic DNA Analysis: Technological Development and Innovative Applications*; Pilli, E., Berti, A., Eds.; CRC Press: Boca Raton, FL, USA, 2021.
27. Mallick, S.; Li, H.; Lipson, M.; Mathieson, I.; Gymrek, M.; Racimo, F.; Zhao, M.; Chennagiri, N.; Nordenfelt, S.; Tandon, A.; et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **2016**, *538*, 201–206. [CrossRef]
28. Pagani, L.; Lawson, D.J.; Jagoda, E.; Mörseburg, A.; Eriksson, A.; Mitt, M.; Clemente, F.; Hudjashov, G.; DeGiorgio, M.; Saag, L.; et al. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **2016**, *538*, 238–242. [CrossRef]
29. Xavier, C.; de la Puente, M.; Mosquera-Miguel, A.; Freire-Aradas, A.; Kalamara, V.; Vidaki, A.; Gross, T.; Revoir, A.; Pośpiech, E.; Kartasińska, E.; et al. Development and validation of the VISAGE AmpliSeq Basic Tool to predict appearance and ancestry from DNA. *Forensic Sci. Int. Genet.* **2020**, *48*, 102336. [CrossRef]
30. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **2000**, *155*, 945–959. [CrossRef]
31. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* **2005**, *14*, 2611–2620. [CrossRef]
32. Kopelman, N.M.; Mayzel, J.; Jakobsson, M.; Rosenberg, N.A.; Mayrose, I. Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* **2015**, *15*, 1179–1191. [CrossRef] [PubMed]
33. R: A Language and Environment for Statistical Computing. Available online: http://www.r-project.org/ (accessed on 20 June 2021).
34. Paradis, E. Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics* **2010**, *26*, 419–420. [CrossRef]
35. Alexander, D.; Novembre, J.; Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **2009**, *19*, 1655–1664. [CrossRef] [PubMed]
36. Cheung, E.Y.Y.; Gahan, M.E.; McNevin, D. Prediction of biogeographical ancestry from genotype: A comparison of classifiers. *Int. J. Legal Med.* **2017**, *131*, 901–912. [CrossRef] [PubMed]
37. Palencia-Madrid, L.; Xavier, C.; de la Puente, M.; Hohoff, C.; Phillips, C.; Kayser, M.; Parson, W. on behalf of the VISAGE Consortium, Evaluation of the VISAGE Basic Tool for Appearance and Ancestry Prediction Using PowerSeq Chemistry on the MiSeq FGx System. *Genes* **2020**, *11*, 708. [CrossRef] [PubMed]

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Article 2. Development and evaluations of the ancestry informative markers of the VISAGE Enhanced Tool for Appearance and Ancestry

Reference: Forensic Science International. Genetics (2023), 64, 102853

ISSN: 1878-0326

DOI: 10.1016/j.fsigen.2023.102853

Web link: https://www.sciencedirect.com/science/article/pii/S1872497323000285

Authors: Ruiz-Ramírez J[1*], de la Puente[1*], Xavier C[2], Ambroa-Conde A[1], Álvarez-Dios J[3], Freire-Aradas A[1], Mosquera-Miguel A[1], Ralf A[4], Amory C[2], Katsara MA[5], Khellaf[5], M. Nothnagel M[5,6], E.Y.Y. Cheung EYY[7], Gross TE[7], Schneider PM[7], Uacyisrael J[8], Oliveira S[9], Klautau-Guimarães MdN[9], Carvalho-Gontijo C[9], Póspiech E[10], Branicki W[11], Parson W[2,12], Kayser M[4], Carracedo A[13,14], Lareu MV[1], Phillips C[1]
* contributed equally to the study

Author information:
1. Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain
2. Institute of Legal Medicine, Medical University of Innsbruck, 6020 Innsbruck, Austria
3. Faculty of Mathematics, University of Santiago de Compostela, 15705 Santiago de Compostela, Spain
4. Department of Genetic Identification, Erasmus MC, University Medical Center Rotterdam, 3015 CN Rotterdam, South Holland, the Netherlands
5. Cologne Center for Genomics, University of Cologne, 50823 Cologne, Germany
6. University Hospital Cologne, 50937 Cologne, Germany
7. Institute of Legal Medicine, Faculty of Medicine and University Clinic, University of Cologne, 50823 Cologne, Germany
8. Fiji Police Forensic Biology and DNA Laboratory, Nasova, Suva, Fiji
9. Departamento Genética e Morfologia, Universidade de Brasília, Brasília, DF, Brazil
10. Malopolska Centre of Biotechnology, Jagiellonian University, 30-387 Kraków, Poland
11. Institute of Zoology and Biomedical Research, Jagiellonian University, 30-387 Kraków, Poland
12. Forensic Science Program, The Pennsylvania State University, University Park, State College, PA 16802, USA
13. Fundación Pública Galega de Medicina Xenómica (FPGMX), Instituto de Investigación Sanitaria (IDIS),15706 Santiago de Compostela, Spain
14. Genomics Group, CIBERER, CIMUS, University of Santiago de Compostela, Spain

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Quality metrics (JCR, 2023):

- Journal Impact factor: 3.2
- Quartile: Q1 (legal medicine) and Q2 (genetics & heredity)
- Rank by JCR: 1/23 (legal medicine) and 68/191 (genetics & heredity)

Contribution: methodology, MPS sequencing, collection of population data, data curation, formal analysis and writing (partial contribution)

Authorization:



Development and evaluations of the ancestry informative markers of the VISAGE Enhanced Tool for Appearance and Ancestry

Author:
J. Ruiz-Ramírez,M. de la Puente,C. Xavier,A. Ambroa-Conde,J. Álvarez-Dios,A. Freire-Aradas,A. Mosquera-Miguel,A. Ralf,C. Amory,M.A. Katsara,T. Khellaf,M. Nothnagel,E.Y.Y. Cheung,T.E. Gross,P.M. Schneider,J. Uacyisrael,S. Oliveira et al.

Publication: Forensic Science International: Genetics

Publisher: Elsevier

Date: May 2023

© 2023 The Authors. Published by Elsevier B.V.

Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

BACK CLOSE WINDOW

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Forensic Science International: Genetics 64 (2023) 102853

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Development and evaluations of the ancestry informative markers of the VISAGE Enhanced Tool for Appearance and Ancestry☆

J. Ruiz-Ramírez [a,1], M. de la Puente [a,*,1], C. Xavier [b], A. Ambroa-Conde [a], J. Álvarez-Dios [c], A. Freire-Aradas [a], A. Mosquera-Miguel [a], A. Ralf [d], C. Amory [b], M.A. Katsara [e], T. Khellaf [e], M. Nothnagel [e,f], E.Y.Y. Cheung [g], T.E. Gross [g], P.M. Schneider [g], J. Uacyisrael [h], S. Oliveira [i], M.d.N. Klautau-Guimarães [i], C. Carvalho-Gontijo [i], E. Pośpiech [j], W. Branicki [k], W. Parson [b,l], M. Kayser [d], A. Carracedo [m,n], M.V. Lareu [a], C. Phillips [a,*,1], on behalf of the VISAGE Consortium[2]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain
[b] Institute of Legal Medicine, Medical University of Innsbruck, 6020 Innsbruck, Austria
[c] Faculty of Mathematics, University of Santiago de Compostela, 15705 Santiago de Compostela, Spain
[d] Department of Genetic Identification, Erasmus MC, University Medical Center Rotterdam, 3015 CN Rotterdam, South Holland, the Netherlands
[e] Cologne Center for Genomics, University of Cologne, 50823 Cologne, Germany
[f] University Hospital Cologne, 50937 Cologne, Germany
[g] Institute of Legal Medicine, Faculty of Medicine and University Clinic, University of Cologne, 50823 Cologne, Germany
[h] Fiji Police Forensic Biology and DNA Laboratory, Nasova, Suva, Fiji
[i] Departamento Genética e Morfologia, Universidade de Brasília, Brasília, DF, Brazil
[j] Malopolska Centre of Biotechnology, Jagiellonian University, 30-387 Kraków, Poland
[k] Institute of Zoology and Biomedical Research, Jagiellonian University, 30-387 Kraków, Poland
[l] Forensic Science Program, The Pennsylvania State University, University Park, State College, PA 16802, USA
[m] Fundación Pública Galega de Medicina Xenómica (FPGMX), Instituto de Investigación Sanitaria (IDIS),15706 Santiago de Compostela, Spain
[n] Genomics Group, CIBERER, CIMUS, University of Santiago de Compostela, Spain

ARTICLE INFO

Keywords:
Bio-geographical ancestry
Massively parallel sequencing
Ancestry informative markers
Autosomal SNPs
Microhaplotypes
Y-SNPs
X-SNPs
1000 Genomes

ABSTRACT

The VISAGE Enhanced Tool for Appearance and Ancestry (ET) has been designed to combine markers for the prediction of bio-geographical ancestry plus a range of externally visible characteristics into a single massively parallel sequencing (MPS) assay. We describe the development of the ancestry panel markers used in ET, and the enhanced analyses they provide compared to previous MPS-based forensic ancestry assays. As well as established autosomal single nucleotide polymorphisms (SNPs) that differentiate sub-Saharan African, European, East Asian, South Asian, Native American, and Oceanian populations, ET includes autosomal SNPs able to efficiently differentiate populations from Middle East regions. The ability of the ET autosomal ancestry SNPs to distinguish Middle East populations from other continentally defined population groups is such that characteristic patterns for this region can be discerned in genetic cluster analysis using STRUCTURE. Joint cluster membership estimates showing individual co-ancestry that signals North African or East African origins were detected, or cluster patterns were seen that indicate origins from central and Eastern regions of the Middle East. In addition to an augmented panel of autosomal SNPs, ET includes panels of 85 Y-SNPs, 16 X-SNPs and 21 autosomal Microhaplotypes. The Y- and X-SNPs provide a distinct method for obtaining extra detail about co-ancestry patterns identified in males with admixed backgrounds. This study used the 1000 Genomes admixed African and admixed American sample sets to fully explore these enhancements to the analysis of individual co-ancestry. Samples from urban and rural Brazil with contrasting distributions of African, European, and Native American co-ancestry were also studied to gauge the efficiency of combining Y- and X-SNP data for this purpose. The small panel of Microhaplotypes incorporated in ET were selected because they showed the highest levels of haplotype diversity

☆ Dedication: This paper is dedicated to co-author Peter Matthias Schneider, our esteemed scientific colleague and friend, who sadly died during its submission.

* Corresponding authors.
E-mail addresses: mdelcarmendela.puente@usc.es (M. de la Puente), christopherpaul.phillips@usc.es, c.phillips@mac.com (C. Phillips).

[1] Contributed equally to the study

[2] A complete list of the institutions and investigators involved in the VISAGE Consortium is provided in Appendix A.

https://doi.org/10.1016/j.fsigen.2023.102853
Received 3 June 2022; Received in revised form 15 February 2023; Accepted 2 March 2023
Available online 5 March 2023
1872-4973/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

amongst the seven population groups we sought to differentiate. Microhaplotype data was not formally combined with single-site SNP genotypes to analyse ancestry. However, the haplotype sequence reads obtained with ET from these loci creates an effective system for de-convoluting two-contributor mixed DNA. We made simple mixture experiments to demonstrate that when the contributors have different ancestries and the mixture ratios are imbalanced (i.e., not 1:1 mixtures) the ET Microhaplotype panel is an informative system to infer ancestry when this differs between the contributors.

1. Introduction

The VISible Attributes through GEnomics (VISAGE) Consortium was initiated in 2017 specifically to develop new massively parallel sequencing (MPS) tools to genotype single nucleotide polymorphisms (SNPs) for the prediction of bio-geographical ancestry (BGA) [1] and a range of externally visible characteristics (EVCs) [2] that contribute to the appearance of an unidentified suspect who has left contact trace DNA at the crime-scene. The SNP genotyping tests for BGA and EVC prediction run in parallel to dedicated MPS assays for age estimation based on quantitative DNA methylation analysis [3]. VISAGE used a two-stage program to develop the MPS toolbox for DNA-based prediction of ancestry, appearance, and age. In the first stage, two prototype Basic Tools (BT) were created comprising the VISAGE BT for Appearance and Ancestry that combined in one MPS assay, 41 markers for predicting eye, hair, and skin colour with 115 ancestry-informative SNPs to analyse BGA [4–6]; and the VISAGE BT for age estimation from blood that combined in one MPS assay 32 CpGs from five genes [7].

Once the BT assays had been comprehensively optimised and their forensic performance evaluated on the Ion S5 (Thermo Fisher Scientific) and MiSeq (Illumina) MPS platforms, VISAGE moved to the second stage of MPS tool design with much more ambitious developmental targets for the Enhanced Tools (ET): The VISAGE ET for Appearance and Ancestry and two separate age tools: the VISAGE ET for age estimation from somatic tissue and the VISAGE ET for age estimation from semen. For the VISAGE ET for Appearance and Ancestry assay, new phenotyping SNPs were introduced for an expanded range of common EVCs beyond, but including, eye, hair and skin colour, which were combined with new BGA SNPs. Additional BGA SNPs focussed on the following objectives: i. the efficient differentiation of Middle East population variation from other Eurasian populations by selecting an expanded panel of SNPs focussed on Middle East regions; ii. the addition of gonosomal SNPs (X and Y) to obtain more detailed analysis of co-ancestry patterns in persons with admixed backgrounds; iii. the inclusion of markers providing a system to estimate the ancestry of the components in simple, 2-way mixed DNA, commonly encountered in forensic analyses. The ET toolbox expanded the age estimation MPS sequencing to eight combined CpG clusters analysing somatic tissue methylation patterns in blood, buccal cells and bones [8], and in a separate test, 13 CpG clusters for analysis of semen [9,10]. Therefore, the ET assays comprised a single combined appearance and ancestry MPS multiplex plus somatic or semen age estimation multiplexes running in parallel workflows in the same way as BT-based analyses. A key part of the development of the VISAGE toolbox was the design, optimisation and implementation of an integrated interpretation framework which includes software for combined statistical consideration of DNA information predicting appearance, age, and ancestry delivered by the ET assays.

For the ET ancestry panel, Middle East informative BGA SNPs were expanded from 12 to 29, but the overall number of binary autosomal BGA SNPs was reduced by ~25%. To analyse co-ancestry patterns in persons with admixed backgrounds, the two most informative marker sets complementing autosomal SNPs are Y-SNPs and mitochondrial DNA (mtDNA) SNPs. However, mtDNA was not considered for ET, as the target DNA copy number is substantially higher than genomic DNA extracted from the same forensic sample. Additionally, a very large number of SNPs would need to be genotyped. To compensate for the lack of mtDNA data, 16 X-SNPs were included to analyse the maternal lineage in admixed persons, alongside a core set of 85 Y-SNPs to analyse the paternal lineage in males. Both X- and Y-SNP sets provide highly informative data with which to compare the co-ancestry ratios estimated from autosomal BGA SNPs. Lastly, 21 Microhaplotypes (MHs) with ancestry informative properties [11] were included to improve the analysis of mixed DNA from the measurement of sequence imbalance and/or detecting more than two haplotypes per locus across multiple MHs, when such mixtures occur.

In the current study, we outline the selection of ancestry markers for the VISAGE ET for Appearance and Ancestry, the performance of these loci for ancestry inference using established statistical methodology, and the use of the specialist X-SNP, Y-SNP and MH marker sets added to the ET ancestry panel for co-ancestry analysis and ancestry-based deconvolution of simple DNA mixtures.

2. Materials and methods

2.1. Selection of ancestry markers for ET

2.1.1. Autosomal BGA SNPs

The previous targeted population differentiations of the BT BGA panel, which was composed entirely of autosomal SNPs, were Sub-Saharan Africa (herein Africa, unless specified as the geographically and genetically distinct North Africa or East Africa), Europe, East Asia, South Asia, America (i.e., Native American populations), and Oceania. These datasets are abbreviated to AFR, EUR, EAS, SAS, AMR and OCE, respectively. ET expanded the above population divisions to include Middle East populations (ME), located in regions ranging from North Africa bounded by the Sahara, eastwards to Iran and southwards towards the regions adjacent to the horn of Africa, where originally no distinction was made between North African variation and that shown by other Middle East populations when selecting candidate BGA SNPs. An additional 12 or more BT SNPs that had previously exhibited strong allele frequency contrasts between Middle East populations and Europeans or South Asians, so were also considered. The main source of ME-informative SNPs was the EUROFORGEN NAME panel [12] that previously compiled a total of 111 SNPs. Fig. 1 shows the proportion of autosomal binary and tri-allelic SNPs in both BT and ET ancestry panels, indicating autosomal SNPs comprised 46% of the ancestry markers in ET. Autosomal binary SNP numbers were reduced from BT to ET for all target population groups, ranging from a 21% reduction for SAS to over 87% reduction for OCE. The number of tri-allelic SNPs was increased, but in all markers, there was only limited commonality with BT BGA SNPs - i.e., no population used a simple subset of previously compiled BT BGA SNPs, but each was re-configured to include more powerful ancestry markers to compensate for a reduced number of autosomal SNPs overall, as outlined in Fig. 1. There was also a degree of adjustment for varied population informativeness, measured during searches by calculating Population Specific Divergence (i.e., Shannon's Divergence metric applied to the comparison of one population with all others in the classification system, herein denoted by: $\boldsymbol{I_n}_{\text{AFR}}$; $\boldsymbol{I_n}_{\text{EUR}}$; $\boldsymbol{I_n}_{\text{EAS}}$; etc.) using the Snipper SNP analysis portal, as previously described [13,14]. Notably, many SNPs specifically targeted to differentiate populations outside of Africa and Oceania also had informative patterns of variation in both of these populations. Many of the tri-allelic SNPs selected were chosen because of above-average levels of divergence between South Asia and Europe for allele-2 and/or allele-3.

The bulk of autosomal BGA SNPs selected for ET were identified from previous forensic ancestry panels, using HGDP-CEPH human diversity panel [15–17] and 1000 Genomes Phase III SNP data [18], (HGDP-CEPH population descriptions, grouping and sample sizes as outlined in [19]; and for 1000 Genomes populations in [18] – also see Section 2.2.1). Such population sample sets are increasingly being enhanced with more detailed and comprehensive whole-genome-sequence based variant catalogs. We took advantage of a series of recently published studies that provide high quality variant calls from higher levels of sequence coverage of the human genome [20–22] to compile the most up-to-date allele frequency estimates for each ET BGA SNP. At the same time, identical data was collected for the EVC SNPs of ET to explore whether additional SNPs can improve population differentiations beyond the three overlapping loci for appearance and ancestry analysis used in BT and ET (rs16891982 in *SLC45A2*, rs1426654 in *SLC24A5*, rs12913832 in *HERC2*). Lastly, we analysed genome-wide patterns of population variation in tri-allelic SNPs in the human genome from detailed scrutiny of a full dataset of these markers we had previously compiled [23]. The allele frequency data were then used to estimate and compile markers with the maximum $\boldsymbol{I_n}_{\text{POP}}$ values and then to balance the panel composition by adjusting relative numbers of BGA SNPs for the continental comparisons as previously described [4,13], but ignoring $\boldsymbol{I_n}_{\text{SAS}}$ and $\boldsymbol{I_n}_{\text{ME}}$ calculations and marker balance.

2.1.2. Y-SNPs

A total of 85 Y-SNPs were selected to create a set intended to achieve an optimal balance between detecting all broadly defined global Y-haplogroups and providing additional resolution within certain haplogroups, in a way that could be informative for forensic ancestry analysis, while at the same time, occupying the minimum multiplex space in ET. Supplementary Fig. S1 illustrates three examples of carefully selected Y-SNPs amongst the 85 that all belong to haplogroup R1a, but exhibit geographic frequency distributions that are very different, namely: R1a-Z284 = Northwest Europe; R1a-Z282 = East Europe; R1a-Z93 = West/Central/South Asia. The Y-SNP selection process also made use of compilations of the most informative Y-SNPs identified from the more extensive 859 Y-SNP MPS assays designed to analyse 640 Y-haplogroups [24]. The genomic details and geographic distribution summaries of the 85 Y-SNPs incorporated in ET are detailed in Table 1.

To interpret the Y-SNP data generated, a Y-haplogroup reference database was required, and creating such a database involved the challenging task of compiling disparate published Y-SNP population data. Although a lot of different geographic regions have been studied since Y-SNP genotyping became established, almost every published dataset has analysed different sets of Y-SNPs. Some studies have only focused on broad haplogroups, while others generated high-resolution Y-SNP data within a certain haplogroup. In order to make a reference database that was compatible with the Y-SNPs included in ET, the genotypes of each individual paper were inspected manually, the data that was compatible was included in the database, and incompatible data discarded. In some cases, the absence of certain haplogroups in a population sample could be inferred, for example, if 100 males were typed of which 70 belonged to haplogroup R1b and the remaining 30 to haplogroup I. By extension, the frequency of all Y-SNPs belonging to any other haplogroup was almost certain to be 0, even if those Y-SNPs had not been genotyped in the original study. Ninety Y-SNP studies, plus the data published by 1000 Genomes was used to create a Y-haplogroup database, these studies combined 84,269 genotyped males, of which 35,624 (42%) could be assigned to one of the haplogroups defined by the 85 ET Y-SNPs.

The compiled Y-SNP population database then formed the basis for a mapping module within the VISAGE ET interpretative software. The module generated charts which visualised the frequency distributions of the inferred haplogroup in populations or regions covered by reference studies and compatible with the 85 Y-SNPs. The distribution maps of Supplementary Fig. S1 illustrate the efforts to make a clear distinction

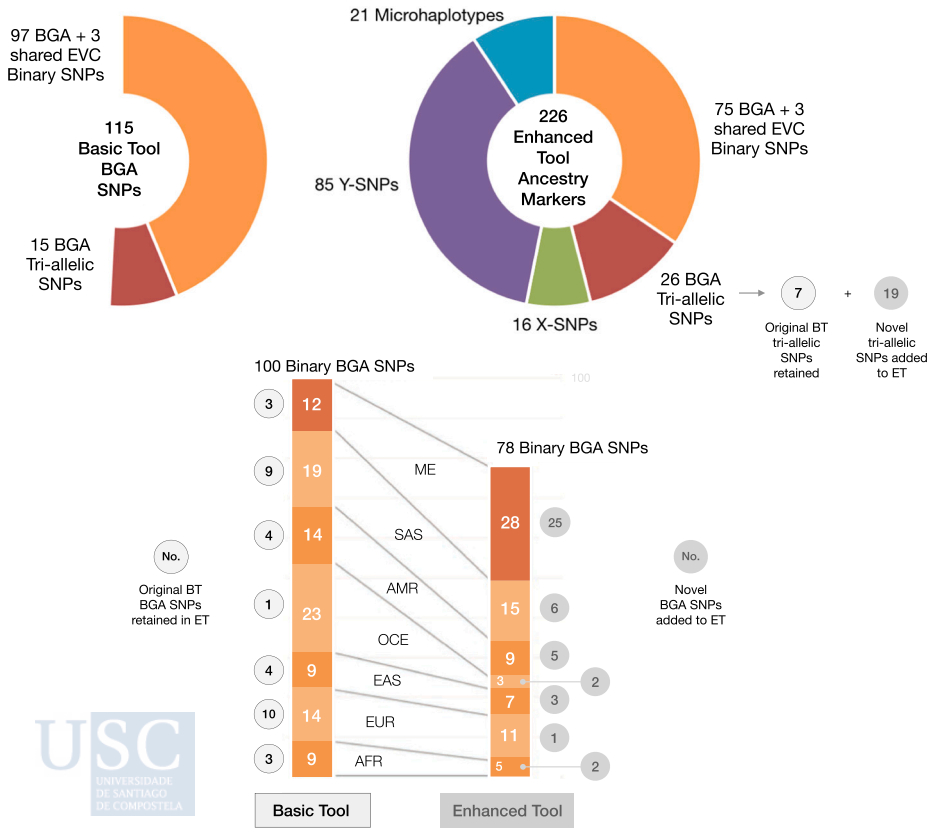


Fig. 1. Proportion of BGA SNPs and ancestry markers in the VISAGE Basic Tool (BT) and the VISAGE Enhanced Tool (ET). Amongst the binary autosomal BGA SNPs, all population-indicative sets were reduced in number, apart from Middle East (ME) informative SNPs, which were more than doubled in number. The expansion in multiplex space dedicated to ancestry markers in ET was occupied with ancestry-informative Microhaplotypes, Y-SNPs, X-SNPs, and more tri-allelic BGA SNPs. Light grey circles left denote BGA SNPs retained, dark grey circles right novel BGA SNPs introduced to ET to improve each population differentiation.

Table 1
Genomic details and geographic distribution summaries of the 85 Y-SNPs incorporated in ET. NA: no information available.

No.	Marker name	SNP-ID	Position GRCh37	Position GRCh38	Substitution	ISOGG Nomenclature	Geographic distribution	No.	Marker name	SNP-ID	Position GRCh37	Position GRCh38	Substitution	ISOGG Nomenclature	Geographic distribution
1	V148	rs181335666	6788191	6920150	G->A	A0	Central Africa, West Africa	43	M522	rs9786714	7173143	7305102	G->A	IJK	
2	L1086	NA	2826312	2958271	A->T	A00	Central Africa	44	M304	rs13447352	22749853	20587967	A->C	J	W Asia, North Africa, Horn of Africa, S Europe, Central Asia, South Asia
3	V168	rs191505182	17947672	15835792	G->A	A1		45	M267	rs9341313	22741818	20579932	T->G	J1	Northern Africa, Horn of Africa, West Asia, South Asia
4	M31	rs369315948	21739754	19577868	G->C	A1a	West Africa, North Africa	46	M172	rs2032604	14969634	12857709	T->G	J2	Southern Europe, West Asia
5	V50	rs189205028	6845936	6977895	T->C	A1b1a	Southern Africa, Central Africa	47	M9	rs3900	21730257	19568371	C->G	K	
6	M32	rs558241924	21740436	19578550	T->C	A1b1b	East Africa, Southern Africa	48	M526	rs2033003	23550924	21389038	A->C	K2	
7	M13	rs3904	21722098	19560212	G->C	A1b1b2b	Central Africa, East Africa	49	M20	rs3911	21733454	19571568	A->G	L	South Asia, West Asia
8	M42	rs2032630	21866840	19704954	A->T	BT		50	P326	rs372687543	8467290	8599249	T->C	LT [K1]	
9	M181	rs2032599	14851554	12739620	T->C	B	Central Africa, Southern Africa, East Africa	51	P256	P256	8685231	8817190	G->A	M or K2b1b	Near Oceania, Wallacea, Australia, Remote Oceania ???
10	M168	rs2032595	14813991	12702062	C->T	CT		52	M231	rs9341278	15469724	13357844	G->A	N	Northern Asia, Central Asia, Americas
11	M145	rs3848982	21717208	19555322	C->T	DE		53	M46	rs34442126	14922583	12810648	T->C	N1a1	Siberia / East Asia
12	M174	rs2032602	14954280	12842354	T->C	D	East Asia	54	VL29	rs752512309	14570424	12458624	T->C	N1a1a1a1a1a	NE Europe, Eastern Europe, Central Asia
13	F6251	NA	7681275	7813234	C->T	D1a	East Asia, Central Asia	55	B479	NA	26271075	24124928	C->A	N1a1a1a1a1c~	East Asia
14	M55	rs2032621	21872738	19710852	T->C	D1b	Japan	56	Z1936	rs774008164	21463326	19301440	C->T	N1a1a1a1a2	NE Europe, Eastern Europe, Central Asia
15	L1378	rs893924838	2828140	2960099	C->T	D2	SE Asia	57	F4205	rs1028202961	16331432	14219552	A->G	N1a1a1a1a3a	Mongolia
16	M96	rs9306841	21778998	19617112	C->G	E	Africa, West Asia,	58	B202	NA	2880546	3012505	T->C	N1a1a1a1a3b	Russian Far East

(*continued on next page*)

Table 1 (*continued*)

No.	Marker name	SNP-ID	Position GRCh37	Position GRCh38	Substitution	ISOGG Nomenclature	Geographic distribution	No.	Marker name	SNP-ID	Position GRCh37	Position GRCh38	Substitution	ISOGG Nomenclature	Geographic distribution
							Southern Europe								
17	M33	rs368762706	21740450	19578564	A->C	E1a	West Africa	59	M2118	rs571876713	23259624	21097738	A->G	N1a1a1a1a4	Russian Far East
18	V38	rs768983	6818291	6950250	C->T	E1b1a	Sub Saharan Africa	60	F2930	rs528311746	19080602	16968722	G->A	N1b	East Asia
19	M215	rs2032654	15467824	13355944	A->G	E1b1b		61	P186	rs16981290	7568568	7700527	C->A	O	East Asia, SE Asia, South Asia, Oceania
20	V32	rs371254614	6932821	7064780	G->C	E1b1b1a1a1b	East Africa	62	M119	rs72613040	21762685	19600799	T->G	O1a	SE Asia, East Asia, Oceania
21	V13	rs368031074	6842263	6974222	G->A	E1b1b1a1b1a	Southern Europe	63	P31	rs200861659	14495243	12383440	T->C	O1b	South Asia, SE Asia
22	M81	rs2032640	21892572	19730686	C->T	E1b1b1b1a	Northern Africa	64	M176	rs11575897	2655180	2787139	G->A	O1b2	East Asia
23	M123	rs371143248	21764586	19602700	C->T	E1b1b1b2a1	East Africa, West Asia	65	M122	rs78149062	21764674	19602788	A->G	O2	East Asia, Oceania
24	M75	rs2032639	21890177	19728291	G->A	E2	Sub Saharan Africa	66	JST-002611	rs2075181	7546726	7678685	G->A	O2a1b	East Asia
25	P143	rs4141886	14197867	12077161	G->A	CF		67	P201	rs2267801	2828196	2960155	T->C	O2a2	Oceania, East Asia
26	M130	rs35284970	2734854	2866813	C->T	C	Central, North & SE Asia, N America, East Asia, Near Oceania, Australia, Remote Oceania	68	P295	rs895530	7963031	8094990	T->G	P or K2b2	
27	M38	rs369611932	21742158	19580272	T->G	C1b3a	Oceania / Indonesia	69	M242	rs8179021	15018582	12906671	C->T	Q	Northern Asia, Central Asia, America
28	M347	rs868363758	2877479	3009438	A->G	C1b3b	Australia	70	M3	rs3894	19096363	16984483	G->A	Q1b1a1a	America
29	M217	rs2032668	15437333	13325453	A->C	C2	South Asia, Southern East Asia, Northern East Asia	71	M207	rs2032658	15581983	13470103	A->G	R	Europe, West Asia, Central Asia, South Asia, North Africa, Central Africa
30	P39	rs887450245	14484581	12363850	G->A	C2b1a1a1	Northern America	72	M173	rs2032624	15026424	12914512	A->C	R1	
31	M48	rs373681213	21749881	19587995	A->G	C2b1a1b	Siberia / Northern East Asia	73	M420	rs17250535	23473201	21311315	T->A	R1a	
32	M89	rs2032652	21917313	19755427	C->T	F		74	Z282	rs112563127	15588401	13476521	T->C	R1a1a1b1a	Eastern Europe, Balkan
33	M201	rs2032636	15027529	12915617	G->T	G	West Asia, South-West Asia, Europe, Central Asia	75	Z284	rs767265794	8717196	8849155	C->G	R1a1a1b1a3a	Northern Europe
34	M285	rs13447378	22741740	20579854	G->C	G1	South-West Asia Central Asia	76	Z93	rs566323605	7552356	7684315	G->A	R1a1a1b2	South Asia, Middle East, Central Asia

(*continued on next page*)

Table 1 (*continued*)

No.	Marker name	SNP-ID	Position GRCh37	Position GRCh38	Substitution	ISOGG Nomenclature	Geographic distribution	No.	Marker name	SNP-ID	Position GRCh37	Position GRCh38	Substitution	ISOGG Nomenclature	Geographic distribution
35	P287	rs4116820	22072097	19910211	G->T	G2	West Asia, South-West Asia, Europe, Central Asia	77	M343	rs9786184	2887824	3019783	C->A	R1b	Western Europe
36	L901	rs567848586	17844304	15732424	C->T	H	South Asia, Eastern Europe, South-West Europe, Western Europe	78	U106	rs16981293	8796078	8928037	C->T	R1b1a1b1a1a1	Western Europe
37	P96	rs1027017284	14869743	12757813	C->A	H2	Eastern Europe, South-West Europe, Western Europe	79	P312	rs34276300	22157311	19995425	C->A	R1b1a1b1a1a2	Western Europe
38	M170	rs2032597	14847792	12735858	A->C	I	Europe, West Asia	80	L21	rs11799226	15654428	13542548	C->G	R1b1a1b1a1a2c1	Western Europe
39	M253	rs9341296	15022707	12910796	C->T	I1	North-Europe, West Europe	81	CTS1078	rs567703217	7186135	7318094	G->C	R1b1a1b1b	Caucasus, Balkan, Middle East
40	M438	rs17307294	16638804	14526924	A->G	I2	South Europe, Central Europe, East Europe	82	V88	rs180946844	4862861	4994820	C->T	R1b1b	Sub Saharan Africa
41	M436	rs17315680	18747493	16635613	G->C	I2a1b	North-Europe, West Europe	83	M479	rs372157627	20834667	18672781	C->T	R2	South Asia
42	M429	rs17306671	14031334	11910628	T->A	IJ		84	B254	rs372295336	14102580	11981874	C->A	S	Oceania, East Asia, Australia
								85	M184	rs20320	14898163	12786229	G->A	T	West Asia, Horn of Africa, North Africa, Southern Europe, South Asia

between zero observations and missing data for those regions lacking genotype observations.

ET Y-SNP data was analysed in male samples in the VISAGE Study populations and compared to X-SNP data. Haplogroup assignments were made using the extensive population data compiled for the ET Y-SNP panel selection and used to generate the geographic distribution charts shown in Supplementary Fig. S1. We did not formally collect Y-SNP data from 1KG, CEPH or Sanger ME data as this was quite incomplete. Furthermore, we chose not to make the inference that all SNP data absent from each project's VCF files meant the male samples all had the *RefSeq* reference allele by default.

2.1.3. X-SNPs

In a previous unpublished survey of X chromosome SNP data which was made to compare variation across the major continental population groups of the HGDP-CEPH diversity panel from 650,000 genotyped SNPs [25], we identified a small number of X-SNPs with highly stratified allele frequency distributions. Sets of between two to four SNPs were compiled that were informative for AFR, EUR, EAS, AMR or OCE population differentiations to create a compact X-SNP panel of 16 markers distributed across the full length of the X chromosome. Five of these 16 SNPs were regularly spaced around the centromere but located in a region with very low recombination (Rc rates graphically summarised in Fig. 5 of [26]) and so were treated as a single haplotype block. The most recently published genomic data with genotypes for all the BGA SNPs of ET from high sequence coverage analysis of 1000 Genomes samples [21] has phased the SNP genotypes in all chromosomes, so X-SNP genotypes from females were collected as haplotypes for the centromeric 5-SNP haplotype block, and from males as single chromosome haplotype data (thus, phased by default). All other X-SNP data was compiled with the same approach used for autosomal variants, but accounting for hemizygosity in males when estimating allele frequencies.

In an operational setting, a forensic ancestry test using ET that analysed co-ancestry patterns would compare Y-SNP data and single chromosome X-SNP genotypes in male samples alone, so phasing into haplotype combinations would not be necessary. We collected the phased data from 1000 Genomes female samples in addition to male genotypes in order to provide the most complete analysis of population variation across the major population groups represented in 1000 genomes and added X-SNP genotype data for AMR and OCE from whole-genome-sequence analyses of the HGDP-CEPH diversity panel samples. An important parallel study was to assess the viability of X-SNP analysis in the admixed African and admixed American population samples of 1000 Genomes (labelled by this project as ACB, ASW African and MXL, CLM, PUR, PEL American [18]) - where X chromosomes of varied ancestral lineages are going to be present in a large proportion of these individuals and a degree of recombination may have disrupted the population stratification shown by the selected X-SNPs in the AFR, EUR and AMR admixture contributor populations.

2.1.4. Microhaplotypes

We chose Microhaplotypes for incorporation into ET from two sets we had previously designed for MPS sequence analysis [11,27] that had been selected and characterised for their ancestry informativeness properties. Carefully selected ancestry informative MH loci will have multiple haplotypes with contrasting population frequencies [28], and potentially allow simple mixed DNA deconvolution with the possibility to assign ancestry to components in simple 2-way mixtures, particularly if they are present in unequal ratios [29]. From 22 MHs originally chosen, 21 were successfully incorporated into the ET assay, comprising 8 from the MAPlex BGA panel [27], and 13 from a panel of 113 MHs designed for forensic identification but including several with ancestry informative haplotype distributions [11]. Six of the eight MAPlex MH loci were shortened from the original much longer loci containing more SNPs [29] to ensure forensic sensitivity analysing degraded DNA, by amplifying size-reduced sequences of comparable length to single-site SNP targets. Details of the SNP sets of the 21 MHs selected for ET and size reductions when made, are outlined in Table 2.

2.2. Reference and test population data

2.2.1. Public population data from human genome sequencing projects

A comprehensive population dataset for ET BGA markers was generated by compiling publicly available online whole-genome-sequencing variant data for 3570 samples, published by three major human genome projects [20–22]. This population data comprised 2504 1000 Genomes project samples (herein 1KG) now consisting of a revised, higher quality variant dataset based on an average 30x sequence coverage [21]; 929 HGDP-CEPH human diversity panel samples (CEPH [20]), and 137 Middle East samples from the analysis of 8 populations by Almarri et al. in 2021 [22], which we refer to collectively as the 'Sanger ME' dataset. We also added 130 samples from the Simons Foundation human genome diversity panel (SGDP [30]) excluding samples that overlap with those of 1KG or CEPH, and 402 samples from the Estonian Biocentre human genome diversity panel (EGDP [31]). Some genotype gaps exist in certain sample panels, notably all the tri-allelic SNP genotypes are missing from EGDP and there is a wide-scale absence of many MH component SNPs from EGDP data. The core ET BGA SNP dataset centred on 1KG, CEPH and Sanger ME SNP genotypes and haplotypes, and we used this data to create a standardised population reference set and to perform most of the evaluations of the ET BGA SNPs' population differentiation capabilities. SGDP and EGDP data is included as testing sample sets for users to make their own explorations.

2.2.2. VISAGE in-house study populations

A range of VISAGE participant laboratory in-house population sample sets (herein Study populations) were genotyped with the ET MPS assay. These sets were chosen to cover geographic gaps in under-represented regions, particularly the Middle East, comprising: 32 individuals from Morocco; 30 from Eritrea; 16 from Somalia; 30 from Central Iraq; 29 from the Kurdistan region of Iraq; 29 Turkish-origin individuals resident in Germany; 41 from Fiji; 19 from rural Brazil (Kalunga individuals, Goiás State), and 16 from urban Brazil (residents of the City of Brasília).

Informed consent was obtained from all Study population donors, which comprised samples previously obtained from: i. Moroccans resident in Madrid collected in 2008 by the Comisaría General de Policía Científica, Madrid, with written informed consent obtained from donors regarding the use of anonymised samples for the characterisation of population variation; ii. Eritrean, Somali, Central Iraqi, Kurdish Iraqi, and Turkish resident in Germany (co-authors P.M.S., T.E.G.) collected according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of the Faculty of Medicine, University of Cologne, Germany, reference no. 17–416 (dated 16.5.2018); iii. Fijian island samples obtained by Fiji Police Forensic Biology and DNA Laboratory, (co-author J.U.), with written informed consent obtained from donors regarding the use of anonymised samples for the characterisation of population variation; iv. Brazilian samples obtained in Brazil (co-authors S.O., M.K.-G., C.C.-G.) with ethical approval from Universidade de Brasília reference No. CAAE: 16542613.8.0000.0030 (rural) and CAAE: 72917916.3.0000.0030 (urban).

2.2.3. Compilation of standardised reference population datasets

A standardised seven-population group reference dataset was constructed to enable end-users to make population analyses independently of the VISAGE ET interpretative software. The reference dataset consisted of: Africans represented by 108 1KG Yoruba from Nigeria (YRI); Europeans by 99 1KG NW Europeans from Utah (CEU); East Asians by 103 1KG Han Chinese from Beijing (CHB); South Asians by 103 1KG Gujarati from Houston (GIH); Middle East by 161 HGDP-CEPH Israeli Arabs from Palestinian, Druze and Bedouin populations plus Algerian

Table 2
Genomic details of the 21 Microhaplotypes incorporated in ET. Original MH nomenclature lists in bold the six loci reduced in size in ET designs to enhance their forensic sensitivity.

Principal SNPs in the haplotype	Extra SNPs in MPS output	Internal MH name	Original MH nomenclature	Principal component SNPs	Extra SNPs in Ion S5 MPS sequence output	5′ coordinate: GRCh37	3′ coordinate: GRCh37	5′ coordinate: GRCh38	3′ coordinate: GRCh38	MH span in nucleotides	Original MH span
4	1	1pA		rs28503881-rs4648788-rs72634811-rs28689700	rs532405039	1529950	1529998	1594570	1594618	48	
3	2	MH01	**mh01KK-01**	rs6663840-rs58111155-rs6688969	rs199565833 / rs548721351	3743319	3743391	3826755	3826827	72	259
3	-	1pD		rs6702428-rs12031966-rs6687440	-	106770076	106770110	106227454	106227488	34	
3	-	MH03	**mh02KK-134**	rs12469721-rs3101043-rs3111398	-	161079411	161079450	160222900	160222939	39	103
3	2	MH04	mh02KK-136	rs6714835-rs6756898-rs12617010	rs530973697 / rs546011313	228092389	228092459	227227673	227227743	70	70
5	1	3pB		rs11129981-rs11129982-rs75361533-rs11129983-rs1896565	rs528474614	42924625	42924691	42883133	42883199	66	
5	5	3qC		rs6583335-rs9848767-rs843520-rs9833841-rs965140	rs559681042 / rs552643442 / rs550318827 / rs60667153 / rs183434367	196379897	196379993	196653026	196653122	96	
4	1	4qD		rs34521178-rs4533811-rs4450974-rs61132367	rs531239419	182795889	182795939	181874736	181874786	50	
4	3	7pB		rs6951954-rs6969555-rs2158900-rs73080042	rs139000977 / rs185814343 / rs552428908	25447589	25447640	25407970	25408021	51	
4	1	8pA		rs10097211-rs80063668-rs73660014-rs7007616	rs538206051	3306430	3306458	3448908	3448936	28	
5	6	8pB		rs34821009-rs7822905-rs7836134-rs7822909-rs6474278	rs577517386 / rs539800640 / rs113457629 / rs188201066 / rs113010596 / rs565537969	40664194	40664243	40806675	40806724	49	
5	1	9pA		rs1408329-rs11789647-rs12555748-rs1535838-rs1408330	rs567753466	2288647	2288718	2288647	2288718	71	
3	4	10pB		rs11816330-rs10828819-rs4749046	rs570240814 / rs536076967 / rs555668598 / rs572123381	25839394	25839446	25550465	25550517	52	
3	2	MH11	**mh11KK-180**	rs4752778-rs74047734-rs7112918-rs4752777	rs140892495 / rs555496836	1690950	1690984	1669720	1669754	34	193

(*continued on next page*)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Table 2 (*continued*)

Principal SNPs in the haplotype	Extra SNPs in MPS output	Internal MH name	Original MH nomenclature	Principal component SNPs	Extra SNPs in Ion S5 MPS sequence output	5′ coordinate: GRCh37	3′ coordinate: GRCh37	5′ coordinate: GRCh38	3′ coordinate: GRCh38	MH span in nucleotides	Original MH span
4	1	12qB		rs11177060-rs2111058-rs10878750-rs11835920	rs571889826	68508276	68508353	68114496	68114573	77	
4	-	15qD		rs1816771-rs74033914-rs5007156-rs4965040	-	98255928	98255978	97712698	97712748	50	
4	2	MH18	**mh16KK-255**	rs16956011-rs3934954-rs3934955-rs3934956-	rs576469239 / rs184092108	81970353	81970407	81936748	81936802	54	142
4	1	MH20	mh18KK-293	rs621320-rs621340-rs678179-rs621766	rs80093367	76089886	76089968	78329886	78329968	82	82
3	2	MH21	**mh21KK-315**	rs6517970-rs202132081-rs8131148-rs6517971	rs533846035 / rs538072435	21880158	21880231	20507846	20507919	73	145
3	2	MH22	**mh21KK-324**	rs2838868-rs7279250-rs8133697	rs537553521 / rs567533147	46714641	46714707	45294726	45294792	66	158
5	2	22qB		rs4925431-rs4925399-rs4925432-rs4925400-rs77899570	rs192804904 / rs537823715	49060976	49061028	48665164	48665216	52	

Mozabite - the latter sample divided into an eighth reference population representing North Africa in STRUCTURE analyses of Eurasians; Oceanians by 28 HGDP-CEPH Papuans from Bougainvillea and Papua New Guinea; Native Americans by 79 samples, comprising 61 HGDP-CEPH samples from Maya, Pima, Colombian and Amazonian Surui and Karitiana populations, supplemented by 18 1KG Peruvians from Lima, Peru (PEL) which we had previously analysed to indicate no detectable non-American co-ancestry (from analysis of 572,743 Affymetrix Human Origins SNPs, see Table 10.5 of [32]).

The 1KG admixed populations, comprising 96 African Caribbean individuals in Barbados (ACB), 61 Americans of African Ancestry in SW USA (ASW), 64 individuals with Mexican Ancestry from Los Angeles USA (MXL,) 94 Colombians from Medellin, Colombia (CLM), 104 Puerto Ricans from Puerto Rico (PUR), and 67 of 85 PEL (i.e., with detected co-ancestry), were used as the testing sample set for evaluating the admixture analysis capabilities of ET by comparison with co-ancestry estimates provided by the 1000 Genomes project (personal communication, Adam Auton, Albert Einstein College of Medicine, NYC, USA).

2.3. Evaluation of ancestry and co-ancestry analysis using ET BGA SNPs

The efficiency of the ET autosomal BGA SNPs to infer an individual's population of origin was assessed for a seven-group differentiation of African, European, East Asian, South Asian, American, Oceanian and Middle East ancestries. For BGA prediction within the ET integrated interpretation framework, VISAGE has implemented dedicated software using a strictly Bayesian approach that applies a flat prior probability model and multiple logistic regression to assign a forensic sample to one of the above seven possible ancestry classes. In the reported study we did not apply the alternative likelihood ratio analyses that form the core of the Snipper web portal [33], but instead relied on STRUCTURE analysis [34] to assess the ability of the ET ancestry markers to discern complex ancestry patterns in population samples from the Middle East, as well as populations representing regions where admixture to varying degrees is the predominant demographic pattern observed.

We developed a two-stage nested STRUCTURE analysis approach which analysed the test population sets (POPFLAG=0) with the reference population dataset (POPFLAG=1), which consisted of five continental populations of AFR, EUR, EAS, AMR and OCE at K:5, with a second Eurasian-focussed STRUCTURE analysis using a reference population dataset at K:6 consisting of AFR, EUR, SAS, EAS, ME and a sixth North African (NAF) population. The division of Middle East and North African ancestries followed the observation of consistent separation of the NAF Algerian Mozabite reference population from the ME HGDP-CEPH Israeli Arab reference populations at K:6. This approach was adopted after originally evaluating a dual K:5 run orientated towards west Eurasia (AFR, EUR, NAF, ME, SAS) and east Eurasia (EUR, NAF, ME, SAS, EAS), although using these two slightly different K:5 runs did not show any advantage over a K:6 Eurasian analysis. STRUCTURE was run with 100,000 burnin steps and 100,000 MCMC steps, using correlated allele frequencies under the Admixture model. Cluster membership proportion plots were constructed with CLUMPAK v.1.1 [35]. Optimum 'K' genetic cluster values were inferred by calculating mean ΔK and L(K) values using standard protocols [36,37].

The ability of ET X-SNPs to infer the ancestry of an individual's X-chromosome complement was evaluated using Principal Component Analysis (PCA), by uploading reference data from 1000 Genomes for simple three-way comparisons based on AFR-EUR-AMR, using the 'Classification of multiple profiles with a custom Excel file of populations' option in the Snipper web portal [33]. The multiple profiles classifier provides a Bayes likelihood ratio and PCA analysis, which is now based on the three 2D plots of principal component (PC) 1 vs PC2, PC1 vs PC3, and PC2 vs PC3. X-chromosome ancestries were assigned based on the position of an admixed study sample in relation to these three reference population PCA clusters, and unassigned if this lay in the region of minor overlap between clusters, Positions were judged to be equidistant from two adjacent cluster centroids by visual inspection of PCA chart data. Such points occupying intermediate positions where cluster overlap can occur, were interpreted to indicate recombination of contributor population X-SNP alleles, and alternatively the presence of

two X-chromosomes with different ancestries in females. Initial explorations of five-way analyses with PCA of the 16 X-SNPs used the above three populations plus CEPH Oceanians, and 1KG East Asians (CHB), and results were compared using a single Bayes likelihood ratio test in Snipper.

The 5-SNP centromeric haplotype was compiled in parallel to the full set of 16 SNPs to assess the effect of limited recombination rates on preserving haplotype structures in admixed individuals (with a focus on 1KG ACB and ASW, but also reviewing patterns in 1KG CLM, MXL, PUR, PEL). However, to generate PCAs we used the 'Naive Bayes (Hardy-Weinberg principle need not apply)' option in Snipper which adjusts the likelihood calculations to account for non-independence of variables, typically seen with syntenic marker sets.

2.4. Microhaplotype reconstruction from ET data and pilot experiments to evaluate ancestry-based deconvolution of simple mixed DNA

The haplotypes of each MH locus, identified as combinations of composite SNP alleles on the same sequence strand, need to be reconstructed from sequence data obtained from the ET MPS run. For this reason, we previously developed a custom MH calling pipeline for MPS sequence data from the Ion S5 platform [38]. In brief: i. a synthetic partial reference genome is constructed from 100 kb sequence segments extracted from the GRCh37/hg19 genome assembly that contain each

Table 3
Genomic details of autosomal ancestry informative SNPs incorporated in ET. SNP rs3857620 was the only redundant marker in terms of uninformative allele frequencies. SNPs are listed in each population set in descending order of differentiation power. BT: originally in the VISAGE BT ancestry panel; EVC: shared with the EVC informative SNP set; Chr: chromosome.

	No	SNP	Source	Chr	GrCh37 coordinate	GrCh38 coordinate		No	SNP	Source	Chr	GrCh37 coordinate	GrCh38 coordinate
African	1	rs2814778	BT	1	159174683	159204893	**American**	1	rs12498138	BT	7	83533047	83903731
	2	rs1871534	Novel	8	145639681	144414297		2	rs12594144	BT	20	62157718	63526365
	3	rs2789823	BT	9	136769888	133904766		3	rs7151991	Novel	3	121459589	121740742
	4	rs1197062	BT	17	58641118	60563757		4	rs17130385	BT	14	32635572	32166366
	5	rs9479657	Novel	6	153928396	153607261		5	rs3737576	BT	10	115196019	113436260
European	1	rs16891982	EVC BT	5	33951693	33951588		6	rs6088466	Novel	1	101709563	101244007
	2	rs1426654	EVC BT	15	48426484	48134287		7	rs9847307	Novel	20	32913534	34325728
	3	rs12913832	EVC BT	15	28365618	28120472		8	rs11960137	Novel	3	64525713	64540037
	4	rs12142199	BT	1	1249187	1313807		9	rs2024566	Novel	5	155338081	155911071
	5	rs8072587	BT	17	19211073	19307760	**South**	1	rs182857716	Novel	22	41697338	41301334
	6	rs10962599	BT	9	16795286	16795288	**Asian**	2	rs367953206	Novel	16	48221771	48187860
	7	rs9522149	BT	13	111827167	111174820		3	rs3857620	Novel	6	57496076	57629240
	8	rs2196051	BT	8	122124302	121112062		4	rs1757928	BT	4	130022161	129101006
	9	rs1924381	BT	13	72321856	71747724		5	rs2472304	BT	15	75044238	74751897
	10	rs2715883	BT	11	120133494	120262785		6	rs12405776	Novel	1	242431557	242268255
	11	rs1592672	Novel	12	80128593	79734813		7	rs2026999	BT	9	103140157	100377875
East Asian	1	rs3827760	BT	2	109513601	108897145		8	rs3844336	BT	8	62214766	61302207
	2	rs1545397	Novel	15	28187772	27942626		9	rs1796048	BT	2	97643576	96977839
	3	rs1229984	BT	4	100239319	99318162		10	rs1567803	Novel	2	101343018	100726556
	4	rs6437783	Novel	3	108172817	108453970		11	rs6754311	Novel	2	136707982	135950412
	5	rs1371048	BT	15	64161351	63869152		12	rs13280988	BT	8	112370516	111358287
	6	rs881929	Novel	2	145753166	144995599		13	rs17625895	BT	16	25775102	25763781
	7	rs4657449	BT	16	31079371	31068050		14	rs10764919	BT	10	131663651	129865387
Oceanian	1	rs4471745	Novel	1	165465281	165496044		15	rs1040934	BT	10	78066260	76306502
	2	rs3751050	BT	17	53568884	55491523	**Middle**	1	rs1024124	Novel	15	33617064	33324863
	3	rs10954737	Novel	11	9091244	9069697	**East**	2	rs12880237	Novel	14	68621818	68155101
Tri-allelic	1	rs1074689	Novel	16	52216074	52182162		3	rs1317026	Novel	6	161154955	160733923
SNPs	2	rs1150911	Novel	1	228494382	228306681		4	rs1495085	BT	8	15298515	15441006
	3	rs12629397	Novel	3	65814779	65829104		5	rs166054	Novel	16	11285202	11191345
	4	rs1382568	Novel	8	11351220	11493711		6	rs17086288	Novel	6	124210612	123889467
	5	rs1398461	BT	13	83839778	83265643		7	rs2156208	Novel	18	60131306	62464073
	6	rs17287498	Novel	10	54530788	52771028		8	rs234623	Novel	20	57488964	58913909
	7	rs2375771	Novel	4	187371930	186450776		9	rs262037	Novel	5	177990886	178563885
	8	rs2387842	Novel	12	38736442	38342640		10	rs2835133	Novel	21	37133457	35761159
	9	rs2585339	BT	14	49134978	48665775		11	rs310362	Novel	8	59925618	59013059
	10	rs2605361	BT	12	74903531	74509751		12	rs3852253	Novel	7	18866190	18826567
	11	rs2737126	BT	17	3618815	3715521		13	rs3862700	Novel	18	67862224	70194988
	12	rs392461	Novel	5	81720271	82424452		14	rs4308478	BT	5	136334314	136998625
	13	rs393953	Novel	21	43389036	41968927		15	rs4465645	Novel	17	50832843	52755483
	14	rs408046	Novel	15	80031510	79739168		16	rs4737753	BT	8	54701811	53789251
	15	rs4540055	BT	4	38803255	38801634		17	rs487750	Novel	9	138603740	135711894
	16	rs5030240	Novel	11	32424389	32402843		18	rs6496996	Novel	15	93402496	92859266
	17	rs556365	Novel	16	65927802	65893899		19	rs6701640	Novel	1	170696474	170727333
	18	rs6588145	Novel	1	65859784	65394101		20	rs6894681	Novel	5	127218995	127883303
	19	rs6933094	BT	6	150297603	149976467		21	rs7252391	Novel	19	44142771	43638619
	20	rs7171818	Novel	15	58855169	58562970		22	rs7594173	Novel	2	32900330	32675263
	21	rs776912	BT	1	10847784	10787727		23	rs7816786	Novel	8	101349662	100337434
	22	rs7989291	Novel	13	57572989	56998855		24	rs7975017	Novel	12	26428793	26275860
	23	rs809540	Novel	2	7879001	7738870		25	rs848461	Novel	7	77582265	77952948
	24	rs914468	Novel	20	62100463	63469110		26	rs9467370	Novel	6	24968682	24968454
	25	rs9845503	Novel	3	59700977	59715251		27	rs9817359	Novel	3	76473163	76424012
	26	rs6504633[a]	Novel	17	48112927	50035563		28	rs9899480	Novel	17	36185665	37826045

[a] Tetra-allelic SNP

MH amplicon; ii. raw reads in FASTQ format are aligned to the partial reference genome using the Burrows-Wheeler aligner (BWA) [39]; iii. alignments are then processed with SAMtools [40] to create the required input files for running the microhaplot R package [41], comprising a VCF file of composite SNPs of each MH and alignments in SAM format, sorted and filtering out short reads (<100 bp) and low-quality alignments (mapping quality < 30); iv. microhaplot output provides a raw table of allele strings and depth per MH, that are then filtered by minimum coverage per allele (min_cov) - set at 15 reads, and minimum allele read frequency (min_allele_frequency) - set at 0.02 for mixtures (0.1 for single-donor samples). All scripts and guidelines for processing raw MPS reads to obtain phased MH alleles are available at Github (https://github.com/MariadelaPuente/VISAGE_ET_Microhaplotyper).

As a pilot study of the viability of using MH data to infer the ancestry of components in simple 2-way mixed DNA, we combined two Coriell control DNA samples NA07000 and NA18498 at 1:1, 1:3 and 1:9 ratios. Coriell samples NA07000 and NA18498 have EUR and AFR ancestries, respectively; and the mixed DNA was run with the ET MPS genotyping assay using an optimised MPS protocol, with the sequence output processed as outlined above to reconstruct the haplotypes of the 21 MH loci. For the ancestry-based deconvolution of the multiple sequences observed in the mixtures, three analysts were asked to independently assign haplotypes based on the proportion of sequence reads recorded for each allele and with prior knowledge of the mixture proportions in each sample. A consensus 21 MH set of profiles was generated from the mixtures, adopting a conservative approach when there were discrepancies amongst analysts, i.e., the profile with less alleles assigned was used. Finally, each 21 MH profile was analysed in STRUCTURE, alongside a standard reference set of 1KG phased haplotypes for the samples from YRI, CEU, CHB, i.e., forming a simplified three-way AFR-EUR-EAS ancestry inference test that exploits the limits of differentiation for these continental population groups when analysing MH loci alone.

3. Results

3.1. Characteristics of autosomal BGA SNPs selected for ET

The genomic characteristics of the 104 autosomal BGA SNPs selected for ET are detailed in Table 3, with markers divided into the population differentiation they provide. The full genotype grids compiled from online datasets and in-house genotyping of population samples with ET are provided in Supplementary Table S1A. Allele frequency estimates for these SNPs are listed in Supplementary Table S1B, with summary frequencies for the 1KG and CEPH population groups, but individually for the Sanger ME and VISAGE study populations, as the gnomAD v.3.1.2 variant database [42] lists individual population allele frequencies for nearly all SNPs. One SNP that was originally thought to be tri-allelic, rs6504633, was in fact tetra-allelic - i.e., showing four common nucleotide substitution alleles in some populations [43].

We also list in Supplementary Table S1A the 184 EVC-SNP marker details and genotypes. Certain of the ET EVC SNPs showed potentially informative allele frequency distributions across the global population groups of this study (individual EVC-SNP's informativeness for the relevant populations are marked in Supplementary Table S1A) and we wished to explore whether they can improve ancestry inferences when combined with the dedicated BGA SNPs of ET (see Section 3.5.3).

Cumulative population-specific Divergence values were calculated (individual SNP data not shown) for the five main continental population groups and were quite comparable for $I_{n\ EAS}$= 6.789; $I_{n\ OCE}$= 6.857; $I_{n\ AME}$= 7.6324, indicating that despite quite different numbers of BGA SNPs targeting these populations, they were well balanced. However, $I_{n\ AFR}$= 13.154 and $I_{n\ EUR}$= 11.351 are much higher Divergence values and reflect a bias towards selecting BGA SNPs that could differentiate Middle East and South Asian populations from those of Europe most efficiently, while African-informative allele frequency distributions are seen in almost all BGA SNPs selected, particularly Middle East-informative markers, underlining the reason why only five SNPs specifically targeting African population differentiation were selected for the ET ancestry SNP set.

Only one selected SNP that was successfully incorporated into the ET MPS assay failed to produce the expected genotypes in the populations studied and this was highlighted when we reviewed the more detailed genomic datasets generated from high coverage sequencing of 1000 Genomes samples published by the New York Genome Centre [21]. The uninformative BGA SNP was rs3857620, with an average rs3857620-A allele frequency in 1000 Genomes SAS populations of ~46%, in contrast to 0% in EUR and 2% in EAS - suggesting a highly informative marker. However, the data from the high coverage sequencing variant calls indicates this SNP has almost no variation in the rs3857620-A allele with rs3857620-G present in all populations at 99–100%. The rs3857620 frequencies in 1KG and CEPH populations groups are summarised in Supplementary Fig. S2. Overall, this uninformative marker illustrates the importance of cross-checks between online SNP databases, as the main 1KG Phase III data portal of Ensemble [44] continues to show inaccurate allele frequency data for this SNP at the time of writing, while gnomAD correctly compiles all the HGDP-CEPH and 1KG high sequence coverage rs385762 allele frequencies that match what we observed.

3.2. Characteristics of X-SNPs selected for ET

The full set of X-SNP genotypes for all online genome datasets and study populations, apart from those of EGDP, are listed in Supplementary Table S1C. This data is arranged in an identical grid to the autosomal SNPs in Supplementary Table S1A for sample order and population grouping. Allele frequency estimates for the 16 ancestry-informative X-SNPs are provided in Supplementary Table S1D.

The chromosomal distribution and six-population summary allele frequencies of the 16 X-SNPs selected for inclusion in ET are shown in Fig. 2. A distinction is made between the five p-arm SNPs plus six q-arm SNPs compiled, and the five SNPs forming a tightly linked set of markers and consequent 5-allele haplotypes on the 3′ side of the X centromere. The Kosambi-adjusted recombination fractions (Rc values) are shown for the X-centromeric haplotype block (herein 5-SXC) indicating minimal recombination in this region, with close to zero recombination likely between the rs6655556-rs5937317 SNP pair at the 3′ end of the block.

The autosomal BGA SNPs of ET efficiently differentiate each of the seven population groups targeted by the SNP selection made for the panel (see Section 3.5), and this extends to the analysis of co-ancestry patterns detected in individuals with admixed backgrounds. However, the panel of 16 X-SNPs was selected to differentiate only the five main continental groups, and we found no indications of strong divergence between South Asian or Middle East populations and the other continental groups. Since there is no need to improve differentiation of unadmixed individuals by combining autosomal and X-SNPs, we advocate analysing X-SNP data separately (along with Y-SNP data in males), when an admixed background is inferred from patterns detected in autosomal BGA SNPs (see Section 3.5.2), so analyses will benefit from inference of contributor ancestries (which in male samples comprise matrilineal and patrilineal components). Given the limits of X-SNP differentiations, any analyses indicating South Asian or Middle East co-ancestry would not gain extra information on the likely admixture contributors from X-SNP genotypes.

3.2.1. 16 X-SNPs

We adopted an approach for analysing the ancestry of an individual's X-chromosome complement by treating the complete 16 X-SNP genotypes and the 5-SXC haplotypes as two separate datasets. The full set of 16 X-SNPs can be treated as a linked group of syntenic markers analysed with a likelihood ratio test in Snipper that is adjusted for association of the alleles tested. If such a test provides a high likelihood value (above

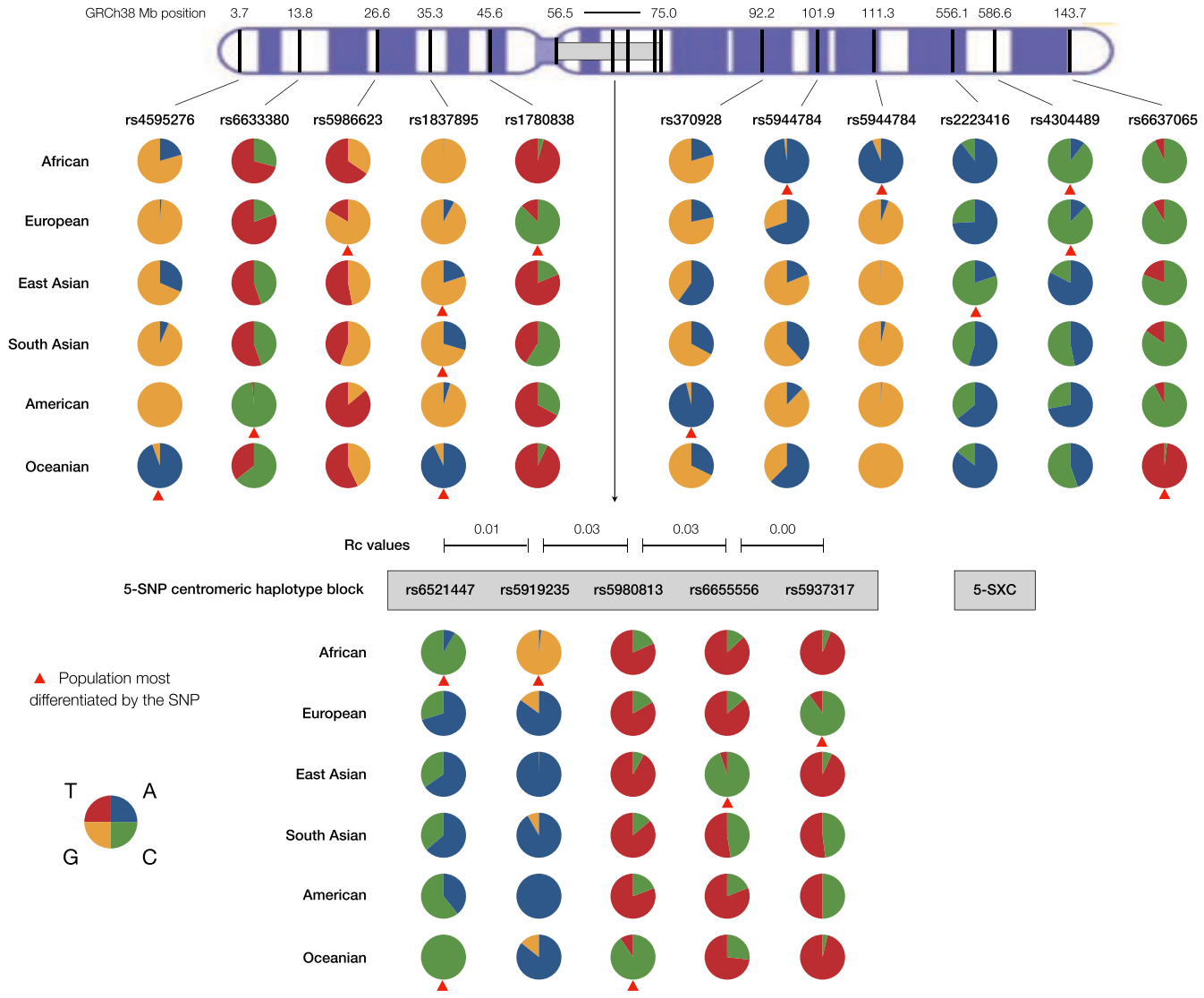


Fig. 2. X chromosome ideogram showing the positions of 16 ancestry-informative X-SNPs selected for ET. The grey bar on the 3′ side of the centromere defines the position of the 5-SNP X centromeric haplotype block (5-SXC) with estimated recombination fraction (Rc) values between the five component X-SNPs shown above each marker pair. Pie charts summarise the allele frequency distributions in six population groups, based on the detailed genotype data in Supplementary Table S1C. The population(s) most differentiated by each X-SNP are marked with red triangles.

100 times more likely a given population) and occupies a position in a PCA plot within a reference population cluster of points, the ancestry inference for the X-chromosome(s) of the individual can be considered to be reasonably secure. Analysis of all seven population groups targeted by ET indicated SAS and ME populations are not well differentiated using the 16 X-SNPs alone (data not shown), and PCA patterns lack clearly defined clusters for these two population groups that would aid easy interpretation. Therefore, we evaluated a five-population ancestry analysis of X-SNPs by testing a random selection of 20 SGDP samples, comprising four from each population group. Supplementary Fig. S3 shows the Bayes likelihood ratio values and PCA plots for this test and indicates all samples were well classified, although American and Oceanian PCA positions require scrutiny of the PC1 vs PC3 and PC2 vs PC3 plot patterns, and the 'Korean-1' likelihood, though correctly assigned as EAS, was below the '100 times more likely' threshold value.

Although this Bayes-PCA test illustrates the effectiveness of the X-SNPs compiled for ET and can be run independently of an autosomal SNP analysis, there is little reason to perform such a test if the sample gives no indications of admixture. When admixture is detected from the autosomal SNP patterns, then the 16 X-SNP data can be more informative, particularly when the sample is male, enabling X-SNP and Y-SNP genotypes to be directly compared. Furthermore, an atypical X chromosome can still occur and will be undetected if there are no apparent co-ancestry patterns amongst the autosomal SNP data. As an example of this phenomenon, the very evident pink reference sample point within the EUR cluster of the PCAs in Fig. 3 is a male PEL sample HG02265, which gave > 99% AMR genetic cluster membership proportions in both 1KG and our own analyses.

To evaluate how the full set of 16 X-SNP genotypes performed with admixed individuals, we used the 1KG admixed American genotypes listed in Supplementary Table S1C and applied a PCA test with reduced reference data comprising AFR (YRI), EUR (CEU) and AMR (CEPH-PEL) genotypes. Results from PCA tests for ACB and ASW males and females, tested separately, are shown in Fig. 3 A and 3B, alongside the genetic cluster analysis of 1000 Genomes for the 30 samples in each population with the lowest proportions of AFR co-ancestry (note 1000 Genomes used the ADMIXTURE genetic cluster algorithm, not STRUCTURE, but results are directly comparable). The overall percentage proportions of African, European, and American inferred X chromosome ancestries identified amongst male and female ACB, ASW and Study Brazilians

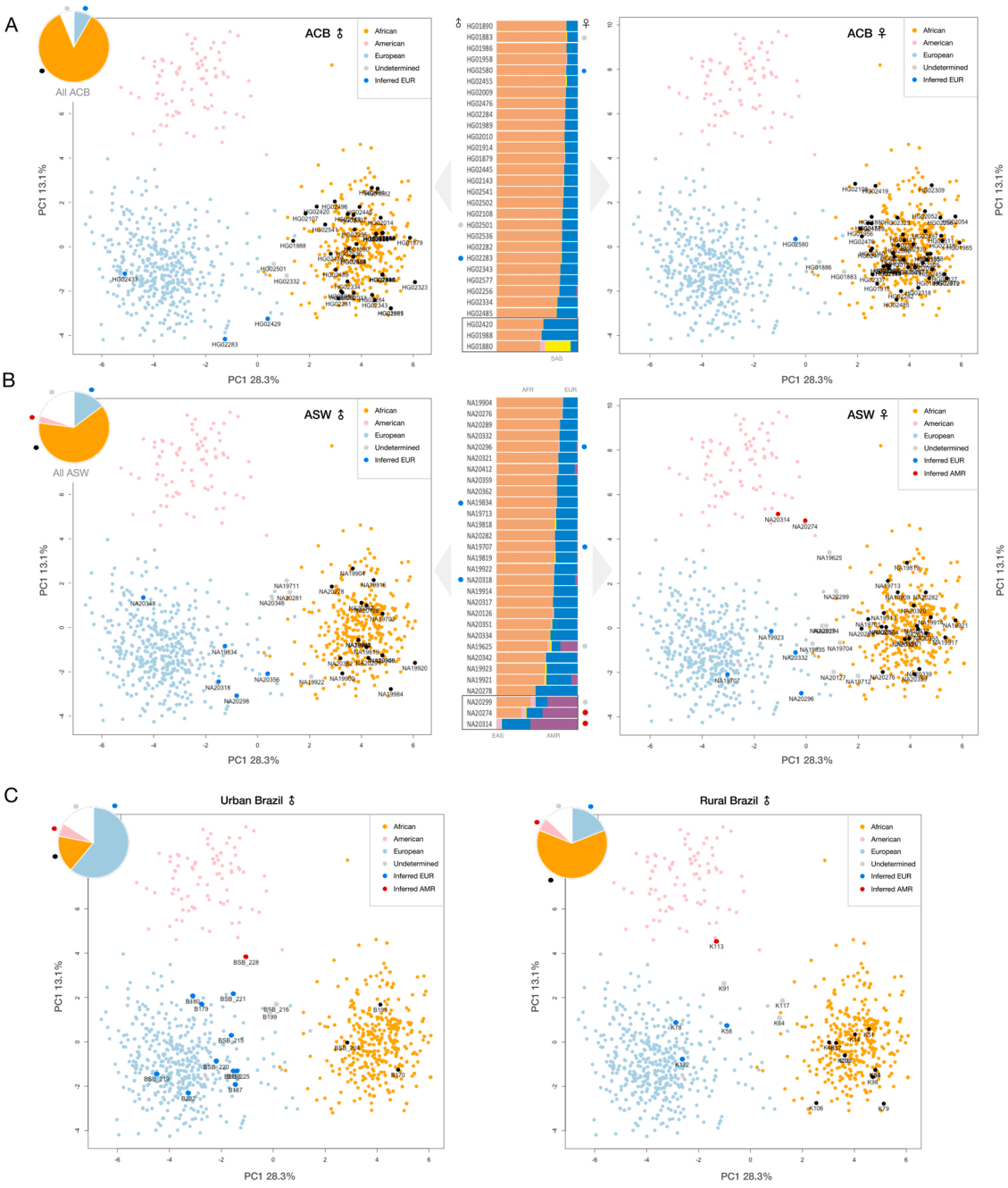


Fig. 3. PCA plots of AFR, EUR, AMR reference populations and admixed African populations from 1000 Genomes, using 16 X-SNP genotype data. **3 A**: African Caribbeans from Barbados (ACB); **3B**: African Americans from SW USA (ASW), each with male and female separately analysed; **3 C**: VISAGE Study population Urban Brazilian males and Rural Brazilian males. Pie charts combine data from both PCAs for ACB and ASW, with proportions of each identified X ancestry in each set of samples (Brazil populations with individual pie charts). Blank pie chart segments and grey dots represent undetermined PCA points occupying intermediate positions between the reference population PCA clusters. The genetic cluster plots in ACB and ASW show the thirty highest non-African co-ancestry proportions in each population taken from the 1000 Genomes own genetic cluster analyses using genome-wide SNP data (Supplementary Table S4A), with dots next to each sample denoting those with an identified non-African PCA position. Note that this compares the autosomal ancestry inference of the 1000 Genomes samples with that made here for the X chromosome, if not African.

using PCA tests, are summarised in Table 4.

Taking each sample set in turn, there are three ACB males inferred to have EUR X chromosomes, two that occupy PCA space between AFR and EUR clusters so are undetermined, and the rest (42) are inferred to have AFR X chromosomes. The individuals in intermediate PCA space or on the edge of the reference clusters could be interpreted to show some recombination, but this would only be discernible in males, since such positions in females can equally represent a heterologous X chromosome pair with different ancestries. This corresponds to proportions of ~6% of ACB males with a EUR X and ~90% with an AFR X, but only 4% of ACB males had X chromosome patterns that could not be discerned. There were five ASW males with an inferred EUR X (~19%), 17 with an inferred AFR X (~65%), and four undetermined.

Notably, amongst ACB females there were only two undetermined individuals with intermediate PCA positions and just one inferred EUR X chromosome pair. ASW females were the only sample set to show inferred AME X chromosome pairs, but both correlate well with the genetic cluster patterns from 1000 Genomes' analyses, particularly

Table 4
Proportions of African, European and American X chromosome ancestries (inferred using 16-SNP PCA tests) identified amongst African Caribbeans in Barbados (ACB); African Americans from southwest USA (ASW); and Study Brazilians from a rural region sample (Kalunga, Goiás State) and an urban sample (Brasilia). The increased levels of European co-ancestry in ASW compared to ACB; and the dominance of European co-ancestry in urban Brazilians compared to rural Brazilians are both evident in individuals whose X chromosome ancestries could be successfully inferred from their PCA positions.

Admixed population samples	Total no. of samples	Undetermined PCA position	X Ancestry Inference Success Rate	% AFR X chromosomes	% EUR X chromosomes	% AMR X chromosomes
ACB Males	47	2	96%	90%	6%	-
ACB Females	49	2	96%	94%	2%	-
ASW Males	26	4	85%	65%	20%	-
ASW Females	35	8	77%	60%	11%	6%
Urban Brazil Males*	16	2	87%	19%	62%	6%
Rural Brazil Males*	18	3	83%	61%	17%	6%

* Only one female in these population samples

sample NA20134, which has no detectable AFR ancestry with either test (although self-declared to have four African American grandparents). There were eight undetermined X ancestries in this set, although the PCA positions of NA19625 and NA20299 between AFR and AMR reference clusters match well with the genetic cluster proportions of both co-ancestries detected by 1000 Genomes. Although it is not possible to say whether these patterns are due to recombination amongst the 16 X-SNPs, or heterologous X chromosome ancestries. Four inferred EUR X chromosome pairs also reflect the higher levels of EUR co-ancestry in ASW compared to ACB, with a consequent smaller proportion of 23/35 inferred AFR X chromosome pairs (60%).

The Brazilian samples were assessed in the same way to evaluate how successfully a *de novo* sample set could be analysed using PCA cluster analysis of X-SNP data. This sample of 16 urban male Brazilians (Brasília) and 18 rural male Brazilians (Kalunga, Goiás State) included a single rural female Brazilian which was not analysed. Comparisons of the EUR, AFR and AMR X ancestries inferred from the PCA patterns showed a marked contrast between the urban and rural samples, with urban Brazilians having the highest proportion of EUR X chromosomes (62%) and lowest AFR (19%) amongst the four admixed populations studied here, whereas the rural Brazilians showed 17% EUR X chromosomes and 61% AFR - akin to the proportions seen in ASW. A single male from each Brazilian sample had an AMR X chromosome. Results from the PCA tests along with pie-chart summaries of the above ancestry portions of each population sample (i.e., males and females combined) are shown in Fig. 3. Similar analyses were made of male and female CLM, MXL, PUR and PEL, shown in Supplementary Figs S4A-S4H. However, the three-way co-ancestry patterns common to these population samples (i.e., AFR-EUR-AMR in varied co-ancestry ratios), make it difficult to infer X ancestries with the same level of confidence as ACB, ASW and Study Brazilians.

These results indicate that well separated clusters are obtained in PCA for 16 X-SNPs using reference data representing the three major contributor populations of admixed Americans. The system provides a simple way to compare the positions of individuals with unknown ancestry when they show signals of admixture in their autosomal SNP genetic cluster patterns. Our assessments suggest PCA provides a secure inference of the X ancestry in the majority of ACB and ASW test samples, and males have the advantage of Y-SNP genotype data from ET with which to compare the pattern of variation observed in the X-SNPs. When the levels of minority co-ancestry are low the X ancestry inference success rates are consequently higher, likely due to a smaller proportion of intermediate PCA positions (undetermined ancestry) caused by heterologous X pairs in females and reduced disruptive recombination in the X chromosomes of males.

All X-SNP data uploaded to the Snipper Bayes-PCA analysis portal are provided as a series of Excel worksheets (that can be made active individually for uploading to Snipper by placing in the leftmost 'worksheet 1′ position in the uploaded file) in Supplementary File S1.

DE SANTIAGO DE COMPOSTELA

3.2.2. The 5-SNP X centromere haplotype

Although recombination was not observed to be a frequent disruptor of the 16 X-SNP genotype combinations across the full chromosome length in ACB and ASW males, representing only 5–15% of intermediate, and therefore undetermined PCA positions, it is instructive to focus on the 5-SXC haplotypes with much lower levels of allelic assortment occurring compared to the whole chromosome. The estimated recombination rate [26] of the first 5′ p-arm ET X-SNP to the 5′-most 5-SXC SNP is 45.3%, and the 3′-most 5-SXC SNP to last 3′ q-arm X-SNP is 45%, compared with an estimated 6.8% across the 5-SXC haplotype span. The full details of the centimorgan values and Kosambi-adjusted Rc estimates [26] between each of the 16 X-SNPs of ET are listed in Supplementary Table S1C, rows 3–5. As a result of these estimated recombination rates, population-specific X-SNP patterns will become assorted in individuals with co-ancestry within a few generations, but the 5-SXC allelic combinations will stay intact across a much longer period of time (i.e., many more generations) after each admixture event. As such, the 5-SXC provides a more secure way to track the X ancestry of individuals with admixed backgrounds of unknown time-depth.

Analysis of 5-SXC haplotype frequency distributions in 1KG and CEPH populations are outlined in Fig. 4, with the underlying data for all populations studied listed in Supplementary Table S1C. It is evident from these haplotype frequency plots that African and East Asian 5-SXC haplotypes are highly specific, with four 'signature' haplotypes accounting for 90.3% of those observed in 1KG AFR (CGTTT; CGCTT; CGTCT; AGTTT, with CGTTT alone forming almost 60% of observed haplotypes), and 88.9% of 1KG EAS (AATCT; CATCT; AACCT; CACCT, with AATCT also near 60% of observed haplotypes). 1KG EUR have five specific haplotypes with a collective frequency of 75.5%, and CEPH AMR four haplotypes with a collective frequency of 60%, although it should be noted that the absence of the CGCTC haplotype is likely to be due to inaccurate phasing of these SNPs in the CEPH genome data, since CGCTC is found at an equivalent frequency to AACTT in 1KG admixed American populations (suggesting the collective AMR-specific haplotype frequency of 60% is likely to be an underestimate by ~8%). 1KG SAS almost exclusively show combinations of European and East Asian specific haplotypes and lack South Asian specific haplotypes, although AATTT accounts for 9.5% of total 5-SXC haplotypes in this population, and although present in EUR at 4% frequency, is not frequent in AFR or EAS. In contrast to SAS, CEPH OCE have a frequent, highly indicative CATTT haplotype accounting for 46.2% of all haplotypes observed in this sample, ten times more frequent than the 4.8% of CATTT haplotypes observed in 1KG SAS. The Study Fijian 5-SXC haplotype data is included (haplotypes in females were not phased but inferred) and show the CATTT and AATTT haplotypes indicate that Oceanian ancestries and South Asian admixture are observable characteristics of this population sample.

The 1KG admixed African American populations indicate little or no disruption of the 5-SXC haplotype combinations in those individuals with admixed backgrounds, and it is possible to infer a sex bias (AFR females-EUR males) in the admixture profiles of the ACB and ASW, since

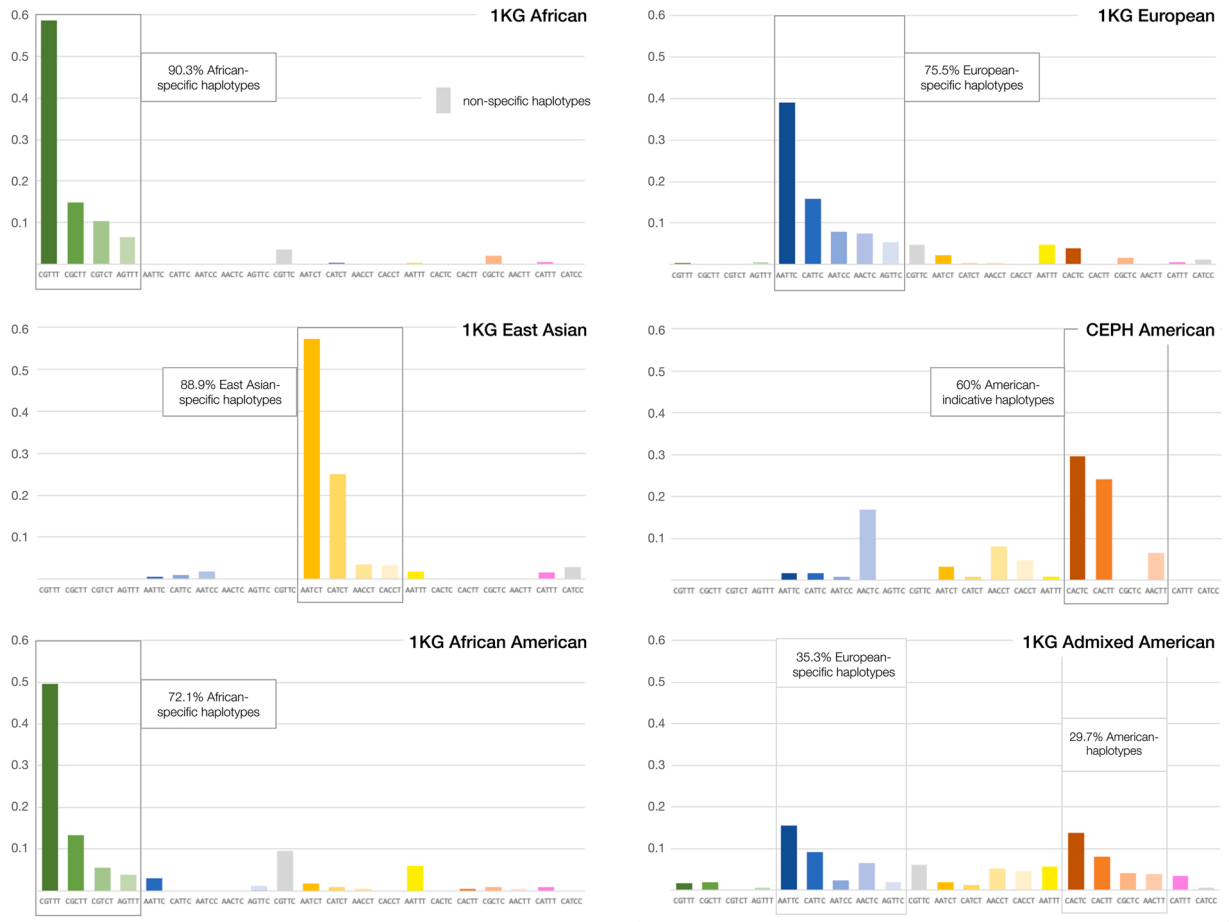


Fig. 4. Haplotype frequency estimates for the 5-SNP X centromeric haplotype block (5-SXC) in 1000 Genomes AFR, EUR, SAS, EAS, admixed African and admixed American populations, plus CEPH AMR (includes 18 1KG PEL), CEPH OCE and Study Fijians (to allow comparison with the relatively small sample size for CEPH OCE). All three main population groups have 4–5 specific haplotypes shown boxed and with a collective haplotype frequency. However, SAS mainly comprises an equal combination of EUR and EAS haplotypes, with AATTT the most differentiated haplotype compared to other populations. Almost half of OCE 5-SXC haplotypes are CATTT, making this haplotype highly informative, although it is found at 3–4% frequencies in other populations. Eleven haplotypes are not charted as they were observed at frequencies less than 3% of total variation in any one population.

very few AATTC EUR-specific haplotypes are observed in these samples, in line with the co-ancestry proportions estimated from PCA analyses shown in Table 4. Interestingly, the very similar total EUR-specific and AMR-specific haplotypes observed in the four admixed American populations of 1000 Genomes provides evidence of further sex bias (AMR females-EUR males), since EUR co-ancestry is the dominant component in CLM, and PUR, and about half of MXL autosomal SNP patterns (no formal Y-SNP analysis made). Although the 5-SXC haplotypes cannot be phased and therefore must be inferred, the real power of using a tightly linked combination of SNPs is their persistence as population-specific haplotypes in individuals that are likely to have had admixture events occurring in their family histories some time ago. Once again, males avoid phasing issues and allow comparison with patterns in all 16 X-SNPs, as well as the highly informative Y-SNP genotypes generated by ET.

An important point to highlight is the genotyping performance in the ET assay of rs5937317 - the key EUR-informative 3′ bounding SNP of the 5-SXC haplotype. The below-average sequence coverage observed for this SNP in MPS analysis means it has a much higher genotyping no-call rate than the other 15 X-SNPs, even when applying a more relaxed sequence coverage threshold of a minimum 20 reads. We observed a 26% no-call rate for this SNP when analysing the VISAGE study population samples, which is certain to be higher with forensic DNA. In sharp contrast, the other 15 X-SNPs gave a single no-call from 3600 genotypes successfully obtained. Therefore, to maintain the significant ancestry-informativeness value of the X-SNP panel in ET, detailed above, either *AmpliSeq* MPS primer design adjustments will be necessary, or closely sited substitute SNPs with comparable patterns of EUR-informative allele distributions should be identified.

3.3. Characteristics of Microhaplotypes selected for ET and mixed DNA sequence analysis

3.3.1. Patterns of variation in the 21 Microhaplotypes

Supplementary Fig. S5 provides scaled positional genomic maps outlining the component SNPs in the haplotype structure of the 21 MHs, which also include low frequency SNPs reported by the Ion S5 genotyping software, but not compiled when reconstructing the haplotypes of each locus. Our experience with developing MH loci for forensic MPS assays [11,27] has been that low frequency SNPs within a sequence segment that has much more polymorphic SNPs making up the common haplotypes, creates too much data complexity and these extra SNPs do not merit compilation since their variation minimally changes the distribution of haplotype variability. In terms of ancestry analysis, certain SNP alleles do vary considerably between populations and create the ancestry-informativeness in the selected MHs, even if there is little or no variation recorded in some populations. The MH maps in Supplementary Fig. S5 are placed above bar plots summarising the population

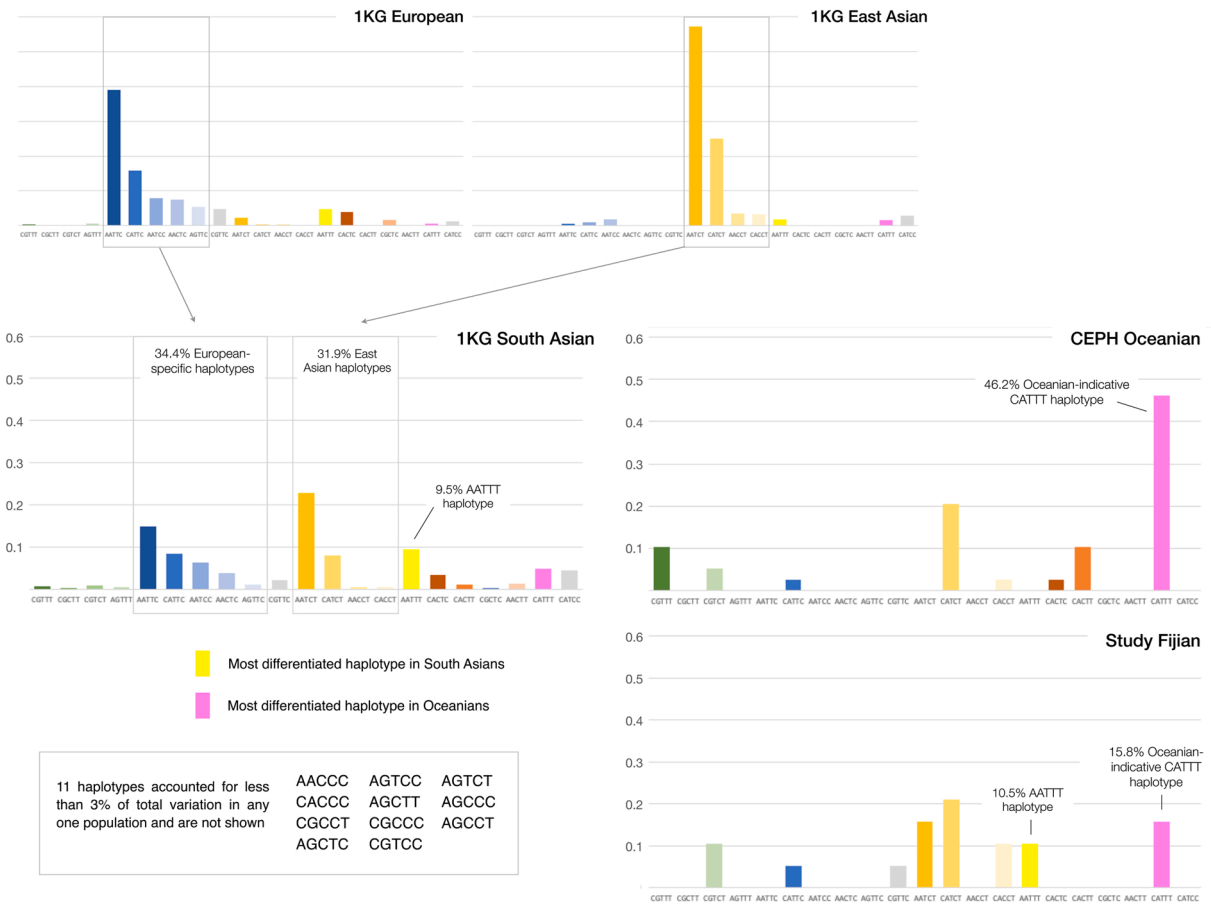


Fig. 4. (*continued*).

distribution of haplotype frequencies in the 21 MHs. Very rare haplotypes observed in one or two, or many populations, are marked in red or yellow to aid visibility. The underlying haplotype data in all populations including the VISAGE Study populations is compiled in Supplementary Table S2. We included both CEPH and Sanger Middle East population data in the bar plots for cross-comparison purposes, as the CEPH whole-genome-sequencing data was not phased so allelic combinations need to be inferred from the common haplotypes of each MH in other Eurasian population data. Therefore, some discrepancies occur, but these are mainly lower frequency haplotypes, and the Sanger ME haplotype frequencies should be considered the most reliable for reference purposes.

Although any review of MH haplotype frequency distributions is rather subjective, this is the first analysis of Middle East variation for these types of loci, so it is useful to discern the overall patterns of variation in the data. In line with previous observations [11,27], AFR haplotype frequencies show consistently higher levels of variation compared to other populations. The opposite characteristic applies to OCE levels of polymorphism in these loci, where fewer haplotypes are observed and one or two haplotypes predominate in many of the ET MHs, e.g., in 8pB the CATCA haplotype alone accounts for almost 85% of the total variation in Oceanians. The AME haplotype frequencies also represent lower levels of polymorphism, but only for certain MHs; (notably 1pD, 10pB, 15qD, and the TCT haplotype in MH03, at >75% frequency). In general, neither the CEPH nor Sanger ME haplotype distributions are very differentiated from EUR populations, and with the exception of 3qC, the same low level of differentiation applies to SAS vs EUR. The development of ancestry predictive systems that use single-site SNP genotypes in combination with haplotypes from a series of MHs has not been completed in either the VISAGE interpretative software or the research laboratories supporting the development of the VISAGE ET assay. Therefore, the MH data generated by ET has not been integrated in order to enhance the autosomal BGA SNP data forming the core of most forensic ancestry prediction systems. Despite the complications of using mixed marker datasets for Bayes analysis, MH data is suitable for inclusion in SNP-based analysis runs with STRUCTURE. However, we did not formally test the addition of the 21 MH loci to the 104 autosomal SNPs using STRUCTURE, although experience indicates no improvement is seen in the differentiation of the inferred genetic clusters from this algorithm when marker data is extended in this way.

3.3.2. *Pilot studies to evaluate ancestry-based deconvolution of simple mixtures using Microhaplotypes*

Fig. 5 summarises the findings of the pilot study of ancestry-based mixed DNA deconvolution made on the 2-way mixture constructed from EUR and AFR Coriell control DNAs. First, the STRUCTURE analysis of the standard reference populations of YRI, CEU and CHB indicates well differentiated genetic clusters for each population, despite being based on data from just 21 MH loci (Fig. 5A). Once it was confirmed that analysis of representative populations from the three main continental population groups was sufficiently informative, the cluster membership proportions for the two control DNAs were obtained and showed that STRUCTURE analysis of the haplotypes inferred from sequence read ratios would provide the means to infer the ancestry of each haplotype detected. Second, the voluntary scrutineers tasked with manually recognising the mixture component haplotypes produced inferences that were compiled as 21 consensus MH profiles for the 1:1, 3:1, 9:1 mixture ratios. When a scrutineer inferred an MH profile with more haplotypes than the others, the consensus profile defaulted to the least number of haplotypes and was therefore a conservative estimate. It was not

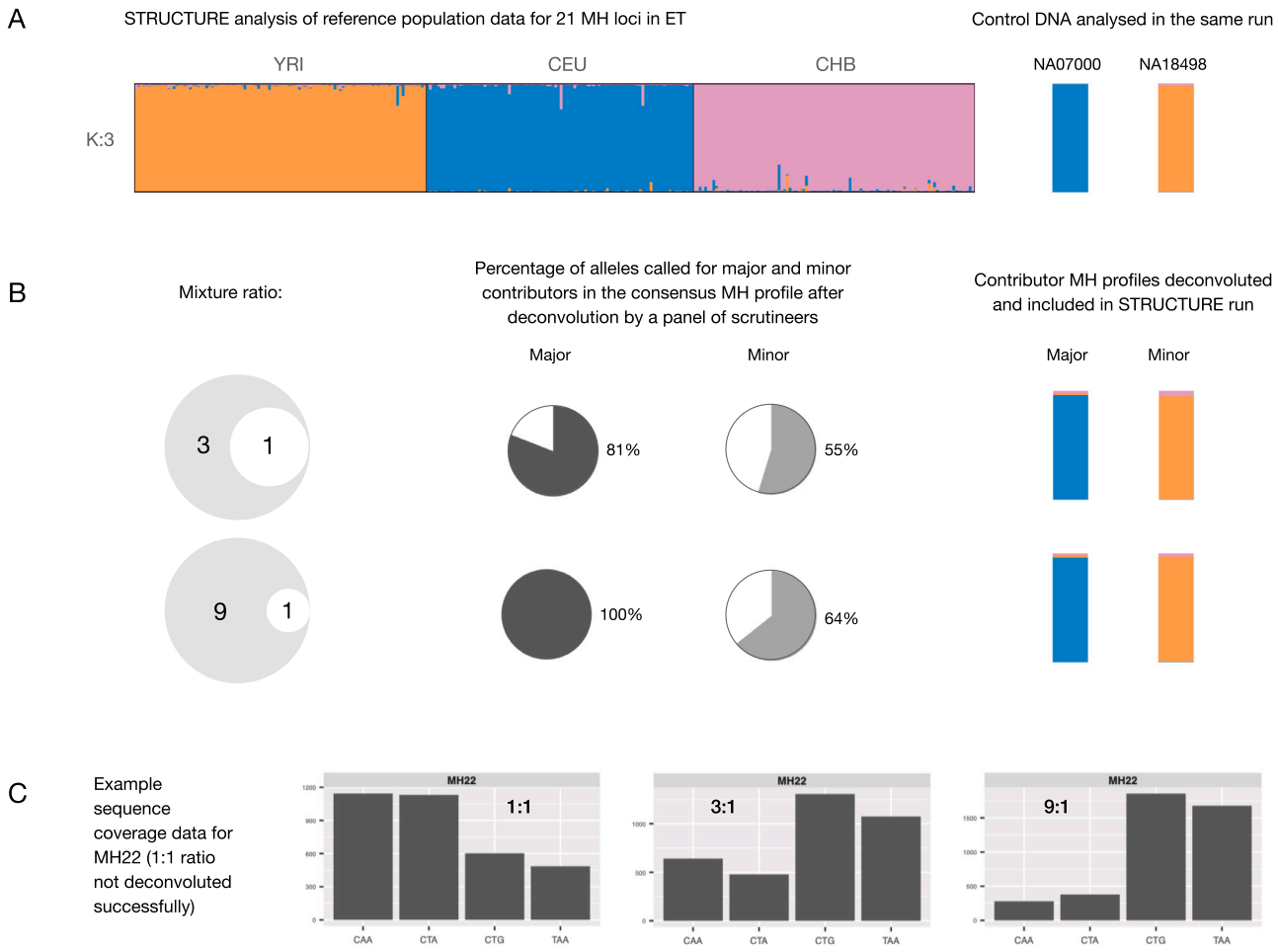


Fig. 5. Pilot study to evaluate the ability of the 21 Microhaplotypes of ET to detect mixed DNA and identify the ancestry of contributors in simple 2-way mixtures. **5 A:** STRUCTURE analysis of AFR, EUR, EAS reference populations indicates the 21 MHs differentiate these populations efficiently and MH profiles comprising haplotypes deconvoluted from a mixture can be included to infer their likely ancestry. **5B:** Percentage of component SNP alleles called for MH profiles identified by a panel of scrutineers of the MPS data for 3:1 and 9:1 mixture ratios (1:1 was not successfully deconvoluted and is not shown). STRUCTURE cluster plots for the deconvoluted MH profiles shown right. **5 C**: Example sequence coverage output for each identified haplotype in MH22. All three mixture ratios allow 'pairing' of two haplotypes per contributor, but this was not possible for the 1:1 mixture ratio in other MH loci or when less than four haplotypes are present.

possible to reliably de-convolute the 1:1 mixture, as sequence read ratios were mostly closely matched from both contributors. As the mixture ratio became more skewed, it was easier to distinguish the major and minor contributors as indicated by Fig. 5B, where 81% of major contributor haplotypes could be inferred and 55% of minor contributor haplotypes in the 3:1 mixture. These values improved to 100% and 64% respectively, in the 9:1 ratio; underlining the fact that more accentuated mixture ratios are easier to deconvolute in this way. There is a very slight indication of incomplete profile reconstruction in the STRUCTURE plots for the minor contributor of the 3:1 ratio, with an increase in the negligible EAS co-ancestry proportion, but the sequences detected from this mixture component would be unequivocally inferred to show AFR ancestry. Typical sequence coverage data (numbers of reads) are shown for example locus MH22 in Fig. 5 C, indicating the difference in read ratios can reflect the mixture ratio to a large extent, and it is clear that the CAA/CTA haplotypes belong to the minor contributor in both 3:1 and 9:1 mixture ratios. However, these sequence coverage readings also show that balanced mixtures do not necessarily lead to balanced reads between the component haplotypes detected. The full set of plots of sequence coverage of the identified haplotypes in the 1:1, 1:3 and 1:9 mixture ratios are given in Supplementary Fig. S6.

In the 1:1 mixture ratio, Supplementary Fig. S6 shows almost all MHs have more than 2 haplotypes (12 with 3 haplotypes, five with 4 haplotypes), and only MH18 and MH21 have the same level of sequence coverage for each of two haplotypes identified in these loci (MH locus 7pB has two haplotypes with a very skewed coverage ratio). These data emphasise the power of Microhaplotypes to detect mixtures even when full deconvolution is not feasible because individual haplotype combinations cannot be inferred from the sequence coverage ratios. Therefore, this pilot study using the 21 MH loci of ET suggests an efficient system for detecting mixed DNA can be applied independently of the single-site SNP data and can alert the user to the presence of a mixture. When two contributors are present in the mixed DNA at unequal proportions there is a good opportunity to identify individual haplotype pairs from each contributor, and if they have contrasting ancestries amongst African, European, or East Asian populations-of-origin, there is the ability to use STRUCTURE to identify these ancestries and assign them to both contributors.

Beyond MH haplotype analysis, bi-allelic SNPs have limited capabilities for mixture deconvolution from MPS sequence data [45] and given the power of multiple-haplotype MH loci to detect simple mixed DNA components, we would advocate discounting single-site SNP data when such mixtures are detected. In contrast, tri-allelic SNPs can detect simple mixtures more efficiently from the detection of three different alleles in the sequence read data for each nucleotide. Compared to the use of MHs to deconvolute mixtures as described above, tri-allelic SNPs have much more limited power, but can add detail to the observations based on the MH sequence data.

The parallel study of this paper, describing the development and inter-laboratory evaluation of the VISAGE ET MPS assay [46], examined the tri-allelic data from the same NA07000-NA18498 mixture series. We briefly summarise these findings below, which added evaluation of increased heterozygosity, and skews in the sequence coverage of the detected alleles of each tri-allelic SNP, in addition to recording the presence of three alleles. The level of tri-allelic SNP heterozygosity increased from ~25% for the contributors to more than 56% in the 1:1 and 1:3 mixture ratios, falling to lower levels in the 1:9 mixture ratio. The observed skews in the sequence coverage of the alleles of the tri-allelic SNPs were close to expectations in the 1:3 and 1:9 ratios, suggesting tri-allelic SNPs are informative markers for the analysis of mixed DNA from MPS data beyond the simple sequence coverage skews occurring with bi-allelic SNPs. Finally, three-allele patterns were found in 11 of the 26 tri-allelic SNPs of ET, which matches the expected number from examination of the contributor SNP genotypes.

3.4. Y-SNP genotypes in VISAGE Study populations

Although Y-SNP data was not compiled from the main human genome variant datasets because of incomplete data, all Y-SNP genotypes obtained for the VISAGE Study population males have been compiled and are listed in Supplementary Table S3. The 5-SXC haplotypes are also listed alongside the haplogroup manually inferred from the Y-SNP alleles in each sample and the description of the region where that haplogroup is most commonly observed.

Although a thorough analysis of the distribution of Y variability in the East African, Middle East and Fijian Study populations was not made, we decided the Brazilian samples would provide an informative pilot study for the comparison of X- and Y-SNP data in two populations likely to have different admixture histories. The Brazilian rural sample is of Kalungas who descended from escaped slaves and have lived in remote settlements in Goiás State for about 250 years. In contrast, the Brazilian urban sample is from Brasília, the capital of Brazil. Compilation of the X- and Y-SNP based matrilineal and patrilineal ancestries is summarised in Supplementary Fig. S7. This data revealed a noticeable European male – African female sex-biased admixture ratio in both population samples but was much more marked in the rural Kalungas. The urban Brazilian males had 6% African Y haplogroups and 45% African-specific 5-SXC haplotypes, while the rural Brazilian males had 39% African Y haplogroups and 92% African-specific 5-SXC haplotypes. European Y haplogroups were found in 94% of urban Brazilian males (19% were designated as 'ME-EUR' with a haplogroup distribution including East European, Caucasus and Middle East regions), and 18% of European-specific 5-SXC haplotypes, while the rural Kalungas had 67% European Y haplogroups and 8% European-specific 5-SXC haplotypes. An interesting point of comparison with the X-Y data is an independent study of the same Kalunga samples using 46 autosomal ancestry-informative Indels, by Carvalho Gontijo et al., in 2018 [47]. Carvalho Gontijo's study detected ~68% African and ~25% European co-ancestry proportions, with the other 7% American (plus a marginal East Asian proportion). These proportions are broadly positioned between the two contributor population ratios with a significant European male – African female sex bias, as we indicated with the X- and Y-SNP data for the Kalungas.

The X ancestry profile of the urban Brazilians showed equal 18% proportions of EUR, EAS and AMR-specific 5-SXC haplotypes. Therefore, a discernible European male sex bias exists in both samples, but is particularly strong in the isolated rural sample, where almost all the observed matrilineal X haplotypes are African-specific and two thirds of the patrilineal Y haplogroups have their most common distribution in Europe. Although this simply represents an initial exploration of data where X- and Y-SNP genotypes can be compared, it suggests a forensic ancestry test that combines each gonosomal marker set will have a degree of power to analyse patrilineal and matrilineal patterns in persons with admixed backgrounds. This is encouraging, given the early decision by VISAGE not to pursue mtDNA analysis as part of the ET assay.

3.5. STRUCTURE analysis of ET BGA SNP data

3.5.1. Worldwide population structure patterns inferred from the autosomal BGA SNPs of ET

We rely on STRUCTURE to analyse the ancestry of unknown donors in forensic DNA tests for several reasons: i. we have found it to be the most effective way to detect and analyse co-ancestry patterns in individuals with admixed backgrounds; ii. although not part of our studies here, STRUCTURE can combine and analyse variant data from different types of genomic marker, so MH loci, STRs, SNPs and Indels can be analysed together in a single run; iii. the availability of detailed genotype data from whole-genome-sequence variant datasets allows a wide range of reference populations to be compiled for almost any marker set, and nearly all SNPs identified in the human genome to date. Combining unknown forensic sample data marked as POPFLAG= 0, with size-adjusted reference data (approximately 100 samples per reference population) marked as POPFLAG= 1, provides an effective way to examine the likely ancestry of the unknown samples. A common problem with the use of STRUCTURE is overfitting the data to a number of inferred genetic clusters (K) greater than the actual clusters that can be properly discerned with the markers used. Since forensic BGA marker sets are limited in number to preserve assay sensitivity, the initial analysis of samples of unknown ancestry with STRUCTURE requires a cautious exploration of each K value, generally from K:2 to K:8. Our experience has indicated that data overfitting - when too many clusters are inferred and individual population groups begin to show irregular within-population cluster membership proportions - can occur after K:5, analysing a continentally-based reference population set of AFR, EUR, EAS, AMR, OCE that includes the less well differentiated population groups of ME and SAS. To counter these effects and to provide the optimum differentiation of genetic clusters, we have adopted a 'nested' approach to STRUCTURE analyses that runs a five-continent reference set with the unknown sample(s) set at K:5 expected clusters. Depending on the cluster membership patterns found in the POPFLAG= 0 samples, another K:5 run analyses the samples with a Eurasian sub-continental reference set of AFR, EUR, ME, SAS, EAS. We have found this improves the cluster patterns detected in admixed samples, which are predominantly from the Americas and therefore show co-ancestry proportions in varying degrees from AFR, EUR and/or AMR contributing populations. One problem can be the detection of SAS co-ancestry in the second Eurasian-centred STRUCTURE run, and in such cases the initial run's reference population data can be adjusted for K:5 expected clusters by swapping out the OCE populations. One example of when the exploratory STRUCTURE runs can require adjustment depending on the results of both analyses, is the 1KG sample HG01880 shown in Fig. 3, with ~30% SAS co-ancestry detected from 1000 Genomes' own genetic structure analyses [18]. Because we do not include SAS reference data in the first STRUCTURE run this would go undetected until the Eurasian sub-continental reference data run was completed, and a new run made with OCE reference genotypes swapped out for SAS.

Applying the Continental K:5 - Eurasian Sub-Continental K:5 nested approach described above to the full range of 1KG, CEPH, Sanger ME and VISAGE Study populations produced a generally robust identification of the majority cluster membership proportions in each sample. The minority cluster membership patterns in almost all samples produced a coherent pattern which matched the geographic location of the populations analysed, particularly those from the Middle East regions. When performing these STRUCTURE analyses, we consistently observed well differentiated genetic cluster patterns at K:6 in the Eurasian Sub-Continental runs when the CEPH Mozabite Algerian samples were included as a sixth population reference set marked as 'North African' (NAF) POPFLAG= 1. For this reason, we show the K:6 patterns generated using six reference populations which includes distinct NAF and ME reference datasets (ME comprising the three Israeli Arab populations of

Bedouin, Palestinian and Druze). Fig. 6 displays, in approximate geographic locations-of-sampling, the STRUCTURE cluster plot segments for the populations from each dataset that show detectable and varying degrees of co-ancestry. The cluster plots are generally arranged in descending order of major co-ancestry components have been expanded two-fold to show individual cluster patterns more clearly. The HGDP-CEPH Sardinian, Tuscan, Adygei EUR populations and the Pakistani SAS populations showed some co-ancestry patterns but are excluded for clarity. All other populations not shown in Fig. 6 had single cluster membership patterns matching those reported in numerous studies of the same samples using several forensic BGA SNP sets [1,4,13–18]. Therefore, we concentrated on results from admixed 1KG samples and the three VISAGE Study populations outside of Eurasia analysed with the initial K:5 STRUCTURE run; and the nine Sanger ME plus four VISAGE Study populations from North African, East African or Middle East regions, analysed with the Eurasian K:6 STRUCTURE run, which included a sixth NAF reference dataset. These runs analysed the 104 autosomal BGA SNPs in ET, comprising bi-allelic and tri-allelic loci. The average cluster membership proportions for the initial K:5 STRUCTURE run and the Eurasian K:6 STRUCTURE run for all 1KG, CEPH, Sanger ME and Study population samples included in each analysis, plus the corresponding segmented cluster plots from this data, are listed in full in Supplementary Tables S4A and S4B for Continental and Eurasian datasets, respectively.

Reviewing the Five-Continental K:5 reference and study population cluster plots first. The five admixed 1KG population cluster patterns shown top left in Fig. 6 plus the 67/85 admixed PEL, are discussed in the next section. In the reference population cluster plot the inability to match the numbers of Oceanian reference samples to those of the other populations is evident, so a degree of bias may have occurred in identifying and quantifying the OCE cluster membership proportions when detected as co-ancestry components in Study Fijians and CEPH Cambodians. Nevertheless, the large-scale reduction of OCE-informative SNPs from 23 in BT to 3 in ET has not affected the ability of the ET BGA SNPs to differentiate this population group. In fact, the first OCE sample set of Papua New Guinea is distinguishable from the second of Bougainvillean samples, with the detectable presence of EAS co-ancestry in the latter. One other cluster pattern to highlight is the 5th AMR sample set comprising CEPH Maya, which shows EUR co-ancestry at a higher level than the other CEPH AMR sample sets (set 2 =Karitiana; 3 =Surui; 4 =Colombians; 5 =Maya; 6 =Pima). This represents a close match to patterns obtained from the two landmark studies of the HGDP-

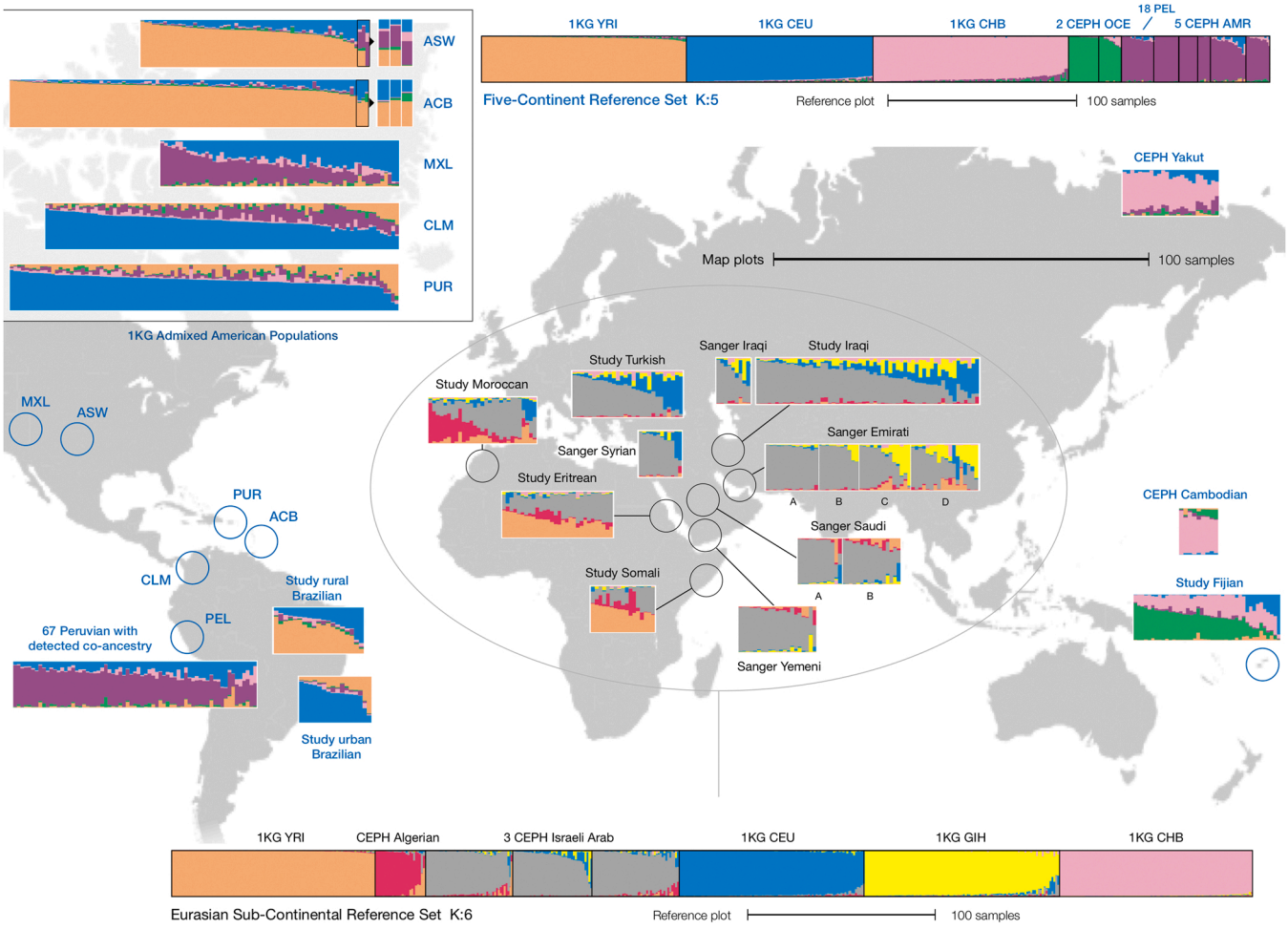


Fig. 6. Cluster plots of STRUCTURE analyses of selected 1KG, CEPH, Sanger ME and Study populations. Nested STRUCTURE analyses consisted of first stage K:5 runs using Five-Continental reference population datasets (POPFLAG=1) comprising 1KG AFR (YRI); EUR (CEU); EAS (CHB); 2 CEPH OCE populations; 5 CEPH AMR populations plus a subset of 1KG PEL with no non-AMR co-ancestry. Populations studied (POPFLAG=0) are shown left and right of central group of populations, comprising six 1KG admixed African and American populations; 67 PEL with detected non-AMR co-ancestry; Study Brazilian rural and urban populations; two CEPH East Asian populations with co-ancestry from other populations; Study Fijians. The central group of Middle East region populations was analysed with the second stage K:6 runs using Eurasian Sub-Continental reference population datasets, comprising 1KG YRI; CEPH Algerian Mozabite; 3 CEPH Israeli Arab populations; 1KG CEU, 1KG SAS (GIH); 1KG CHB. Populations tested were five VISAGE Study populations and nine Sanger ME populations (Emirati A-D and Saudi A-B are arranged separately but not located to a specific region). The three samples in ASW and ACB with highest levels of non-AFR co-ancestry shown on the right as expanded columns.

CEPH diversity panel, using larger marker sets (See Fig. 1 of [48], and Fig. 1 of [25]). The Study Fijian plot indicates most samples would be identified as having OCE origin but note five of the rightmost columns are self-declared Indo-Fijians likely to have SAS co-ancestry, which would be undetected with this reference population data absent from the Continental STRUCTURE run, but present in the Eurasian STRUCTURE run. This exemplifies the need to adjust reference data according to both STRUCTURE analyses (Fijian cluster plots from Eurasian STRUCTURE analysis runs not shown). Lastly, the two Study Brazilian sample cluster plots illustrate the contrast in admixture patterns between them. The rural Brazilian sample has predominant AFR co-ancestry (apart from the rightmost two individuals), contrasting with urban Brazilians, who show predominant EUR co-ancestry, apart from the two rightmost individuals. The two Brazilian samples inferred to have AMR X chromosomes (Fig. 3B) showed 3% (rural, K113) and 10% (urban, BSB228) AMR co-ancestry in this analysis.

The Eurasian sub-Continental K:6 reference and study cluster plots illustrate the successful differentiation of NAF and ME populations, although this was based on the single CEPH Algerian reference population, which could lead to biased analysis due to possible stratification of SNP variation in a population not necessarily representative of variability across a wider region. Therefore, the cluster patterns detected in the Study Moroccan sample are particularly relevant. Study Moroccans show a broad range of NAF co-ancestry proportions from 5–95% in two-thirds of samples, with the majority of Moroccans showing slightly higher proportions of ME co-ancestry than NAF, apart from two samples with AFR-EUR co-ancestry, and two with ME-EUR co-ancestry. AFR co-ancestry is detectable in 9 of the 27 Algerian reference samples, with majority AFR co-ancestry proportions in three. The other twelve Middle East region populations provide cluster patterns well matched to their locations. It is not possible to identify the Emirati A-D populations, but these appear to show a progression in SAS co-ancestry proportions in at least half of samples from C and D. The other Sanger ME populations show predominant ME cluster membership proportions in almost all samples, so would be distinguishable from a European individual apart from (rightmost) Turkish and Syrian samples, which retain a detectable ME co-ancestry. Considering the Middle East population sample set as a whole, a consistent geographic pattern is evident for the majority of samples in each population. This comprises i. a strong presence of the red NAF genetic cluster in half of samples from the Northwest corner of this region, which is shared with the grey ME cluster; ii. a detectable co-ancestry presence of the blue EUR cluster in about a third of samples in the North or Northeast corner, with a predominant ME co-ancestry in these samples from Turkey, Syria, and Iraq (plus minor SAS co-ancestry in most samples); iii. two East African sample sets with equal proportions of AFR and NAF-ME cluster memberships, in patterns which are generally distinct from the other ME populations; iv. a predominant ME cluster, mostly > 90%, in a majority of samples from populations around the Saudi Arabian Peninsula, comprising nearly all Yemeni, Saudi A and B, Emirati A and B, and half of the Iraqis and Syrians. Therefore, using a second STRUCTURE analysis with six reference populations, it is possible to identify ME co-ancestry in the majority of 'unknown' test samples in this study, with a NAF co-ancestry signal detected in half of Moroccans. As a rule of thumb, the presence of AFR and ME, and/or NAF joint cluster memberships suggests a pattern characteristic of East African ancestry. The presence of 15%− 25% ME co-ancestry membership proportions in 4/99 CEU reference samples, suggests a conservative approach would be to infer Middle East ancestry using a threshold of 20–25% or higher ME and/or NAF co-ancestry proportions. Note that this would identify most of the Study Turkish samples as having distinct patterns compared to Europeans. Even applying a stringent threshold of 25% minimum ME/NAF membership proportions to signify Middle East ancestry, rates of non-inference are low amongst these test populations. The two East African populations would have 2% non-inference; Emirati 12%; Moroccans 3%; Iraqis 10%; Turkish 18%, with secure ME inferences possible for all Syrian, Saudi Arabian and Yemeni samples.

3.5.2. *Analysing co-ancestry in admixed population samples with STRUCTURE*

In a criminal investigation, a forensic ancestry test that can reliably identify co-ancestry in a person with an admixed background would, in such cases, provide important information about the likely appearance of a suspect. When previously evaluating the ability of the VISAGE BT ancestry panel to detect admixture and estimate the co-ancestry proportions in such a sample, we made a formal comparison between the cluster membership patterns from analysing the same 504 1KG admixed samples with 572,000 Human Origins array SNPs vs the 115 BGA SNPs of BT. With the BGA SNPs of ET we did a similar comparison of the same samples but used the co-ancestry proportions estimated from genome-wide SNP data published by 1000 Genomes [18]. Supplementary Figs S8A-S8D shows the cluster plots from both analyses with the sample order dictated by the 1KG data arranged by descending majority co-ancestry membership proportions in each population. These plots show the complete 1KG sample set in Supplementary Fig. S8A, followed by expanded plots for admixed Africans ACB, ASW in Supplementary Fig. S8B, and admixed Americans CLM, PEL, PUR, MXL in Supplementary Fig. S8C. Supplementary Fig. S8D shows the correlation analyses and r^2 values used to gauge the levels of correlation between the co-ancestry proportion estimates made with each SNP set, combining AFR and AMR co-ancestry proportions into a single value and comparing EUR co-ancestry proportion estimates directly.

Several factors are evident from a review of the correlation values and STRUCTURE cluster plots produced by ET BGA SNP analyses. First, there is a good match between both SNP sets in the estimates of majority co-ancestry across all samples and populations, particularly when this is above 90%. Consequently, r^2 values are highest for comparisons of AFR co-ancestry proportion estimates in ACB and ASW, and those for EUR in CLM and MXL. Closely matched cluster plot patterns and correlation values are also seen in PEL, although combining AFR-AMR cluster proportion estimates to simplify analysis reduces these correlation values. Lastly, the three co-ancestry outliers (rightmost columns) in ACB and ASW, which are also highlighted in Fig. 4 and 8, present good cluster plot matches, with the SAS co-ancestry proportion recognised by the ET BGA SNPs when this population reference dataset is included as POPFLAG= 1 genotypes. The three ASW outliers indicate an over-estimation of AMR cluster proportions with ET BGA SNPs, and the same marginal but consistent effect is seen in PEL and MXL cluster plot patterns. The worst correlation and cluster plot matches are observed in the PUR comparisons. This appears to stem from a higher level of three-way admixture in this population, although CLM have similar admixture patterns, but produce much better correlation values of r^2= 0.758 for the combined AFR/AMR co-ancestry proportion estimates, compared to r^2= 0.334 in PUR. Much of the AMR co-ancestry estimation in PUR is eroded by many samples with EAS and SAS co-ancestry proportions, and it might be beneficial to consider a K:3 STRUCTURE analysis with AFR, EUR and AMR reference datasets, when three components of admixture are identified in unknown samples and two are either EUR and AMR, or EUR and AFR. A review of the cluster plot for PUR in Fig. 6 indicates this population is generally problematic to analyse for co-ancestry and a significant proportion of samples have low-level EAS and OCE co-ancestry proportions, when the 1KG data suggests these should be recognised as AMR co-ancestry. It is noteworthy that similar studies of the BGA SNPs in BT gave the lowest r^2 value for the PUR combined AFR/AMR co-ancestry proportion estimates of 0.446.

Overall, genetic cluster differentiations become less reliable in individuals with three different co-ancestry components, so STRUCTURE analyses of populations such as Brazil must be approached with caution. Three-way admixture continues to present a considerable challenge for STRUCTURE-based analysis of co-ancestry patterns when using ancestry tests on a much smaller scale than those used for population genetics studies. Therefore, a prudent measure is to explore a series of K:3 runs with different combinations of reference population datasets. Although we do not present further autosomal SNP analysis data for Fijians, this

population would be optimally analysed with EUR, SAS, EAS and OCE reference data in various combinations.

3.5.3. Comparisons of STRUCTURE analyses using 104 BGA SNPs vs combined 104 BGA plus 184 autosomal EVC SNPs

As three EVC-SNPs have been shared for pigmentation trait prediction and ancestry analysis purposes in both VISAGE SNP genotyping assays, it was considered worthwhile to formally evaluate the effect of combining all 104 autosomal BGA SNPs in ET with the 184 autosomal EVC-SNPs. The same Continental K:5 - Eurasian Sub-Continental K:6 nested analysis was made of five and six population reference datasets, respectively, as described in Section 3.5.1. Supplementary Fig. S9 shows the K:5 and K:6 cluster plots for both SNP sets, with accompanying Evanno charts of DeltaK and L(K) [36]. The overall quality of genetic clusters is noticeably reduced in the expanded 288 autosomal SNP dataset compared to the dedicated BGA SNP dataset, particularly for the less divergent populations of NAF, ME and SAS, where a significant number of mixed genetic cluster patterns are observed amongst these three populations, using all 288 SNPs.

4. Discussion

The studies described here have largely concentrated on the added benefit brought by including Y-SNP, X-SNP, and MH markers that all have strong population differentiation properties in the ET ancestry panel. While it is important to acknowledge that many of the autosomal SNPs originally part of the VISAGE BT ancestry panel were replaced with new markers for ET, most of these new BGA SNPs are already well established for forensic use. It was only necessary to adjust the balance of markers towards EUR, EAS and AMR differentiations, and reduce those for AFR and OCE. Expanding the set of ME-informative SNPs in ET has provided considerable benefits in terms of the successful identification of ME and NAF genetic clusters when STRUCTURE is run at K:6 with Eurasian-orientated reference populations (grey and red genetic clusters, respectively, in Fig. 6). Our analyses show in almost all central Middle East population samples from the Sanger ME variant datasets and VISAGE Study samples from these regions, there are majority membership proportions from one or both ME and NAF genetic clusters. Where samples are from regions on the periphery of the central Middle East area, the other genetic clusters the STRUCTURE analyses identified correspond well to their geographic position in this broadly-based region. Specifically, many Turkish show EUR co-ancestry; East Africans have predominant AFR co-ancestries; and the Sanger Emirati in the East (although we cannot place populations A-D in specific geographic positions) show SAS co-ancestry in many individuals. Although it is not appropriate or viable to use STRUCTURE to assign a sample to a specific population, the patterns we have generated with 'nested' Eurasian reference population STRUCTURE runs allow a sample with 'grey' and/or 'red' clusters proportions above 10% to be identified as coming from the Middle East, and exclude an origin from sub-Saharan Africa, Europe, South Asia, or East Asia. In many cases, individuals show a characteristic signature of North African or East African population origins as distinct from the central Middle Eastern regions shown in Fig. 6. Therefore, we consider the goal set by VISAGE of developing an ancestry panel that can efficiently differentiate Middle East population origins from the neighbouring population groups, was largely met and did not require a very large expansion of BGA SNP numbers in ET to accomplish this goal. The adaptation of STRUCTURE runs into a nested approach which analyses a reduced set of reference populations with a narrow range of possible K values, has helped to focus ancestry analyses on the most appropriate regions and as our analyses show, enables more detailed genetic cluster differentiations to be made for the Middle East.

The other expansion made for the ET ancestry panel - that of broadening the types of ancestry informative markers to include X-SNPs, Y-SNPs and MH loci have more specialised application in ancestry analyses used for forensic casework. With the distinctions that can be reliably made between AFR, EUR and AMR co-ancestries with the autosomal BGA SNPs of ET, admixed American individuals can be detected and then analysed efficiently. Consequently, more detail is obtained for male samples by adding the analysis of patterns of variation observed in X and Y chromosome markers. The level of detail we were able to achieve in the analysis of Brazilian samples, which are often too complex in their co-ancestry patterns to be easily studied with small-scale marker sets, highlights the power of combining marker sets with slightly contrasted genetic histories. Such histories often follow admixture events from up to three different contributing populations, and with the complicating effect of varied sex bias in different parts of the same geographic region. Nevertheless, we highlight the problems we encountered in reliably differentiating three-way co-ancestry cluster patterns in Puerto Ricans (PUR) and obtaining comparable data to those of the genome-wide SNP data from 1000 Genomes. Therefore, it is necessary to remain cautious when three different co-ancestries are detected in an individual, as small-scale autosomal BGA SNP panels may not reliably measure their relative proportions compared to genome-wide data.

A key characteristic favouring the use of Microhaplotypes in forensic DNA analysis has been their ability to analyse mixed DNA without the hindrance of non-allelic PCR stutter products complicating the patterns seen [49]. Previously, we developed an approach for analysing mixed DNA with MHs which specifically exploited MH loci with strongly contrasting haplotype frequencies in different population groups. Despite comprising a simple pilot study limited to a single mixed DNA at a few ratios, we have demonstrated the 21 MHs chosen for ET successfully assign ancestries to the components of 2-way mixed DNA, notably when there is imbalance in their ratios, making sequence comparisons easier to achieve. This approach is helped by the ease with which MH loci can be analysed with STRUCTURE and the differentiation they provide of Europe, Africa, and East Asia. Although such analyses are not amenable to a high-throughput MPS pipeline, since haplotypes must be reconstructed locus-by-locus and their sequence ratios estimated, the ability to detect the likely ancestry of contributors could potentially provide key extra information for investigators.

The adaptation of the BT ancestry panel comprising mainly established autosomal forensic BGA SNPs, into the much more broadly based set of BGA markers in ET represents a considerable enhancement of the scope and power of forensic ancestry analysis using MPS, as the chosen name for the VISAGE Enhanced Tool implies.

Acknowledgments

The study was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 740580 within the framework of the VISible Attributes through GEnomics (VISAGE) Project and Consortium. M.d.l.P. is 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 (ED481D-2021–008). J.R. is supported by the "Programa de axudas á etapa predoutoral" funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481A-2020/039). C.P., A.F.A., A.M.M., M.d.l.P., M.V.L. and the work to compile ancestry informative tri-allelic SNPs and microhaplotypes 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. The population studies by S.O. at University of Santiago de Compostela, were financed by the Fundação de Apoio a Pesquisa do Distrito Federal (FAPDF), Brazil.

The authors gratefully acknowledge the sharing of genetic cluster analysis information from the 1000 Genomes Phase III SNP data, kindly provided by Adam Auton, Department of Genetics, Albert Einstein College of Medicine, Bronx, NYC, USA. The authors thank Luciana Maia

Escher dos Santos and Sabrina Guimarães Paiva for their dedicated work in the collection of samples from rural and urban Brazil used in this study. All STRUCTURE analyses were performed by the FinisTerrae II supercomputer at the Centro de Supercomputación de Galicia, Santiago de Compostela (CESGA), Spain.

Appendix A

Centres and investigators of the VISible Attributes through GEnomics (VISAGE) Consortium, Website: http://www.visage-h2020.eu/ (accessed 1st February 2023).

- Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands: Manfred Kayser, Vivian Kalamara, Arwin Ralf, Athina Vidaki.
- Jagiellonian University, Krakow, Poland: Wojciech Branicki, Ewelina Pośpiech, Aleksandra Pisarek.
- Universidade de Santiago de Compostela, Santiago de Compostela, Spain: Ángel Carracedo, Maria Victoria Lareu, Christopher Phillips, Ana Freire-Aradas, Ana Mosquera-Miguel, María de la Puente.
- Medizinische Universität Innsbruck, Innsbruck, Austria: Walther Parson, Catarina Xavier, Antonia Heidegger, Harald Niederstätter.
- Universität zu Köln, Cologne, Germany: Michael Nothnagel, Maria-Alexandra Katsara, Tarek Khellaf.
- King's College London, London, UK: Barbara Prainsack, Gabrielle Samuel.
- Klinikum der Universität zu Köln, Cologne, Germany: Peter M. Schneider, Theresa E. Gross, Jan Fleckhaus, Elaine Cheung.
- Bundeskriminalamt, Wiesbaden, Germany: Ingo Bastisch, Nathalie Schury, Jens Teodoridis, Martina Unterländer.
- Institut National de Police Scientifique, Lyon, France: François-Xavier Laurent, Caroline Bouakaze, Yann Chantrel, Anna Delest, Clémence Hollard, Ayhan Ulus, Julien Vannier.
- Netherlands Forensic Institute, The Hague, the Netherlands: Titia Sijen, Kris van der Gaag, Marina Ventayol-Garcia.
- National Forensic Centre, Swedish Police Authority, Linköping, Sweden: Johannes Hedman, Klara Junker, Maja Sidstedt.
- Metropolitan Police Service, London, United Kingdom: Shazia Khan, Carole E. Ames, Andrew Revoir.
- Centralne Laboratorium Kryminalistyczne Policji, Warsaw, Poland: Magdalena Spólnicka, Ewa Kartasinska, Anna Woźniak.

Appendix B. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2023.102853.

References

[1] C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. 18 (2015) 49–65.
[2] M. Kayser, Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int. Genet. 18 (2015) 33–48.
[3] A. Freire-Aradas, C. Phillips, M.V. Lareu, Forensic individual age estimation with DNA: from initial approaches to methylation tests, Forensic Sci. Rev. 29 (2017) 121–144.
[4] M. de la Puente, J. Ruiz-Ramírez, A. Ambroa-Conde, C. Xavier, J. Pardo-Seco, J. Álvarez-Dios, A. Freire-Aradas, A. Mosquera-Miguel, T.E. Gross, E.Y.Y. Cheung, et al., Development and evaluation of the ancestry informative marker panel of the VISAGE basic tool, Genes 12 (2021) 1284.
[5] C. Xavier, M. de la Puente, A. Mosquera-Miguel, A. Freire-Aradas, V. Kalamara, A. Vidaki, T.E. Gross, A. Revoir, E. Pośpiech, E. Kartasińska, et al., Development and validation of the VISAGE AmpliSeq basic tool to predict appearance and ancestry from DNA, Forensic Sci. Int. Genet. 48 (2020), 102336.
[6] L. Palencia-Madrid, C. Xavier, M. de la Puente, C. Hohoff, C. Phillips, M. Kayser, W. Parson, VISAGE consortium, evaluation of the VISAGE basic tool for appearance and ancestry prediction using PowerSeq chemistry on the MiSeq FGx system, Genes 11 (2020) 708.
[7] A. Heidegger, C. Xavier, H. Niederstätter, M. de la Puente, E. Pośpiech, A. Pisarek, M. Kayser, W. Branicki, W. Parson, VISAGE consortium, development and optimization of the VISAGE basic prototype tool for forensic age estimation, Forensic Sci. Int. Genet. 48 (2020), 102322.
[8] A. Woźniak, A. Heidegger, D. Piniewska-Róg, E. Pośpiech, C. Xavier, A. Pisarek, E. Kartasińska, M. Boroń, A. Freire-Aradas, M. Wojtas, et al., Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones, Aging 13 (2021) 6459–6484.
[9] A. Pisarek, E. Pośpiech, A. Heidegger, C. Xavier, A. Papież, D. Piniewska-Róg, V. Kalamara, R. Potabattula, M. Bochenek, M. Sikora-Polaczek, et al., Epigenetic age prediction in semen - marker selection and model development, Aging 13 (2021) 19145–19164.
[10] A. Heidegger, A. Pisarek, M. de la Puente, H. Niederstätter, E. Pośpiech, A. Woźniak, N. Schury, M. Unterländer, M. Sidstedt, K. Junker, et al., Development and inter-laboratory validation of the VISAGE enhanced tool for age estimation from semen using quantitative DNA methylation analysis, Forensic Sci. Int. Genet. 56 (2020), 102596.
[11] M. de la Puente, M.J. Ruiz-Ramírez, A. Ambroa-Conde, C. Xavier, J. Amigo, M. A. Casares de Cal, A. Gómez-Tato, A. Carracedo, W. Parson, C. Phillips, M.V. Lareu, Broadening the applicability of a custom multi-platform panel of Microhaplotypes: Bio-geographical ancestry inference and expanded reference data, Front. Genet. 11 (2020), 581041.
[12] V. Pereira, A. Freire-Aradas, D. Ballard, C. Børsting, V. Diez, P. Pruszkowska-Przybylska, J. Ribeiro, N.M. Achakzai, A. Aliferi, O. Bulbul, et al., Development and validation of the EUROFORGEN NAME (North African and Middle Eastern) ancestry panel, Forensic Sci. Int. Genet. 42 (2019) 260–267.
[13] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Børsting, P. Johansen, M. Fondevila, et al., Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set, Forensic Sci. Int. Genet. 11 (2014) 13–25.
[14] J.M. Galanter, J.C. Fernandez-Lopez, C.R. Gignoux, J. Barnholtz-Sloan, C. Fernandez-Rozadilla, M. Via, A. Hidalgo-Miranda, A.V. Contreras, L.U. Figueroa, P. Raska, et al., Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas, PLoS Genet 8 (2012), e1002554.
[15] C. Phillips, A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M.D. Perez Carceles, A. Carracedo, P.M. Schneider, M.V. Lareu, Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int. Genet. 7 (2013) 359–366.
[16] C. Santos, C. Phillips, M. Fondevila, R. Daniel, R.A.H. van Oorschot, E.G. Burchard, M.S. Schanfield, L.J. Souto, J. Uacyisrael, M. Via, et al., Pacifiplex: An ancestry-informative SNP panel centred on Australia and the Pacific region, Forensic Sci. Int. Genet. 20 (2016) 71–80.
[17] C. Carvalho Gontijo, L.G. Porras-Hurtado, A. Freire-Aradas, M. Fondevila, C. Santos, A. Salas, J. Henao, C. Isaza, L. Beltrán, V. Nogueira Silbiger, et al., PIMA: A population informative multiplex for the Americas, Forensic Sci. Int. Genet. 44 (2020), 102200.
[18] A. The 1000 Genomes Project Consortium, L.D. Auton, R.M. Brooks, E.P. Durbin, H. M. Garrison, J.O. Kang, J.L. Korbel, S. Marchini, G.A. McCarthy, McVean, et al., A global reference for human genetic variation, Nature 526 (2015) 68–74.
[19] J. Amigo, C. Phillips, M. Lareu, Á. Carracedo, The SNP*for*ID browser: an online tool for query and display of frequency data from the SNPforID project, Int. J. Leg. Med 122 (2008) 435–440.
[20] A. Bergström, S.A. McCarthy, R. Hui, M.A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, et al., Insights into human genetic variation and population history from 929 diverse genomes, Science 367 (2020) 1339–1349.
[21] M. Byrska-Bishop, U.S. Evani, X. Zhao, A.O. Basile, H.J. Abel, A.A. Regier, A. Corvelo, W.E. Clarke, R. Musunuri, K. Nagulapalli, et al., High coverage whole-genome-sequencing of the expanded 1000 Genomes Project cohort including 602 trios, Cell 185 (2022) 3426–3440. VCF data available online: https://www.internationalgenome.org/dataportal/data-collection/30x-grch38 and, ⟨http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/⟩.
[22] M.A. Almarri, M. Haber, R.A. Lootah, P. Hallast, S. Al Turki, H.C. Martin, Y. Xue, C. Tyler-Smith, The genomic history of the Middle East, Cell 184 (2021) 4612–4625.
[23] C. Phillips, J. Amigo, A.O. Tillmar, M.A. Peck, M. de la Puente, J. Ruiz-Ramírez, F. Bittner, Š. Idrizbegović, Y. Wang, T.J. Parsons, et al., A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel, Forensic Sci. Int. Genet. 46 (2020), 102232.
[24] A. Ralf, M. van Oven, D. Montiel González, P. de Knijff, K. van der Beek, S. Wootton, R. Lagacé, M. Kayser, Forensic Y-SNP analysis beyond SNaPshot: High-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing, Forensic Sci. Int. Genet. 41 (2019) 93–106.
[25] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H. M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, R.M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100–1104.
[26] C. Phillips, D. Ballard, P. Gill, D.S. Court, A. Carracedo, M.V. Lareu, The recombination landscape around forensic STRs: accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data, Forensic Sci. Int. Genet. 6 (2012) 345–365.
[27] C. Phillips, D. McNevin, K.K. Kidd, R. Lagacé, S. Wootton, M. de la Puente, A. Freire-Aradas, A. Mosquera-Miguel, M. Eduardoff, T.E. Gross, et al., MAPlex-A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations, Forensic Sci. Int. Genet. 42 (2019) 213–226.

[28] E.Y.Y. Cheung, C. Phillips, M. Eduardoff, M.V. Lareu, D. McNevin, Performance of ancestry-informative SNP and microhaplotype markers, Forensic Sci. Int. Genet. 43 (2019), 102141.
[29] K.K. Kidd, W.C. Speed, A.J. Pakstis, D.S. Podini, R. Lagacé, J. Chang, S. Wootton, E. Haigh, U. Soundararajan, Evaluating 130 microhaplotypes across a global set of 83 populations, Forensic Sci. Int. Genet. 6 (2017) 29–37.
[30] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, et al., The simons genome diversity project: 300 genomes from 142 diverse populations, Nature 538 (2016) 201–206.
[31] L. Pagani, D.J. Lawson, E. Jagoda, A. Mörseburg, A. Eriksson, M. Mitt, F. Clemente, G. Hudjashov, M. DeGiorgio, L. Saag, et al., Genomic analyses inform on migration events during the peopling of Eurasia, Nature 538 (2016) 238–242.
[32] C. Phillips, J. Amigo, D. McNevin, M. de la Puente, E.Y.Y. Cheung, M.V. Lareu, Online population data resources for forensic SNP analysis with Massively Parallel Sequencing: An overview of online population data for forensic purposes, in: E. Pilli, A. Berti (Eds.), In Forensic DNA Analysis: Technological Development and Innovative Applications, CRC Press, Boca Raton, FL, USA, 2021.
[33] Available online: http://mathgene.usc.es/Snipper/ Multiple profiles classifier at: ⟨http://mathgene.usc.es/snipper/analysismultipleprofiles.html⟩ (both accessed 1st February 2023).
[34] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.
[35] N.M. Kopelman, J. Mayzel, M. Jakobsson, N.A. Rosenberg, I. Mayrose, Clumpak: a program for identifying clustering modes and packaging population structure inferences across K, Mol. Ecol. Resour. 15 (2015) 1179–1191.
[36] G. Evanno, S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, Mol. Ecol. 14 (2005) 2611–2620.
[37] C. Santos, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, A. Carracedo, M.V. Lareu, Inference of ancestry in forensic analysis II: analysis of genetic data, Methods Mol. Biol. 1420 (2016) 255–285.
[38] M. de la Puente, C. Phillips, C. Xavier, J. Amigo, A. Carracedo, W. Parson, M. V. Lareu, 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. 48 (2020), 102213.
[39] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760.
[40] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[41] N. Thomas, R Package - Microhaplot, (2019) ⟨https://github.com/ngthomas/microhaplot⟩. (Accessed 1st February 2023).
[42] C. Phillips, J. Amigo, A. Carracedo, M.V. Lareu, Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data, Forensic Sci. Int. Genet. 19 (2015) 100–106.
[43] M. Lek, K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J.S. Ware, J.A.J. Hill, B.B. Cummings, et al., Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (2016) 285–291.
[44] ⟨http://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=6:60527829–60528829;v=rs3857620;vdb=variation;vf=169483878⟩, (Accessed 1st February 2023).
[45] Ø. Bleka, M. Eduardoff, C. Santos, C. Phillips, W. Parson, P. Gill, Open source software EuroForMix can be used to analyse complex SNP mixtures, Forensic Sci. Int. Genet. 31 (2017) 105–110.
[46] C. Xavier, M. de la Puente, M. Mosquera-Miguel, A. Freire-Aradas, V. Kalamara, A. Revoir, T.E. Gross, P.M. Schneider, C. Ames, C. Hohoff, et al., Development and inter-laboratory evaluation of the VISAGE Enhanced Tool for appearance and ancestry inference from DNA, Forensic Sci. Int. Genet. 61 (2022), 102779.
[47] C. Carvalho Gontijo, F. Macêdo Mendes, C.A. Santos, M. de, N. Klautau-Guimarães, M.V. Lareu, A. Carracedo, C. Phillips, S.F. Oliveira, Ancestry analysis in rural Brazilian populations of African descent, Forensic Sci. Int. Genet. 36 (2018) 160–166.
[48] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L. A. Zhivotovsky, M.W. Feldman, Genetic structure of human populations, Science 298 (2002) 2381–2385.
[49] L. Bennett, F. Oldoni, K. Long, S. Cisana, K. Madella, S. Wootton, J. Chang, R. Hasegawa, R. Lagacé, K.K. Kidd, D. Podini, Mixture deconvolution by massively parallel sequencing of microhaplotypes, Int. J. Leg. Med. 133 (2019) 719–729.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Article 3. Broadening the Applicability of a Custom Multi-Platform Panel of Microhaplotypes: Bio-Geographical Ancestry Inference and Expanded Reference Data

Reference: Frontiers in genetics, 11, 581041

ISSN: 1664-8021

DOI: 10.3389/fgene.2020.581041

Web link:
https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2020.581041/full

Authors: de la Puente M[1,2], Ruiz-Ramírez J[1], Ambroa-Conde[1], Xavier C[2], Amigo J[3], Casares de Cal MA[4], Gómez-Tato A[4], Carracedo A[1,3], Parson W[2,5], Phillips C[1], Lareu MV[1]

Author information:
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.

Quality metrics (JCR, 2020):
- Journal Impact factor: 4.599
- Quartile: Q2 (genetics & heredity)
- Rank by JCR: 48/176 (genetics & heredity)

Contribution: methodology, MPS sequencing, collection of population data, data curation and formal analysis.

Authorization:

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 October 2020
doi: 10.3389/fgene.2020.581041

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-Ramírez[1], Adrián Ambroa-Conde[1], Catarina Xavier[2], Jorge Amigo[3], María Ángeles Casares de Cal[4], Antonio Gómez-Tato[4], Ángel Carracedo[1,3], Walther Parson[2,5], Christopher Phillips[1]* and María Victoria Lareu[1]*

[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

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.delapuente.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

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 (Sobrino 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

USC DE COMPOSTELA

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 (I*n*) (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), I*n* 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 ± 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 ± 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

USC DE COMPOSTELA

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 discordancies were explored further in IGV in order to clarify possible causes. Most discordancies (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 errouneously in a very limited number of cases, which could be due to the fact that WGS reads do not neccesarily 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

USC DE COMPOSTELA

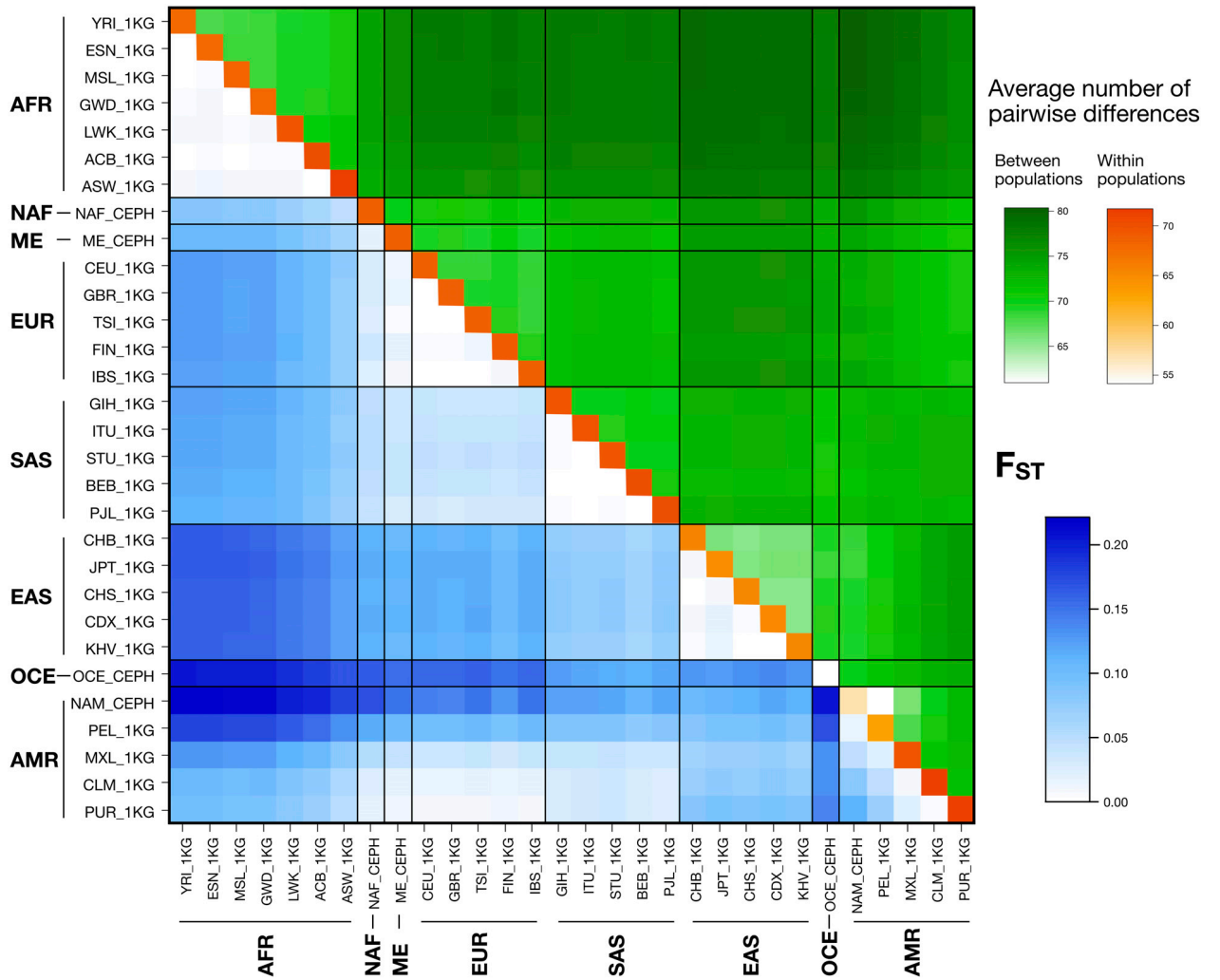


FIGURE 1 | Pairwise F_{ST} (blue) and number of pairwise genotype differences between (green) and within (orange) populations for the autosomal MHs. Populations are named and grouped into eight major populations according to **Supplementary Table S5**.

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

USC DE COMPOSTELA

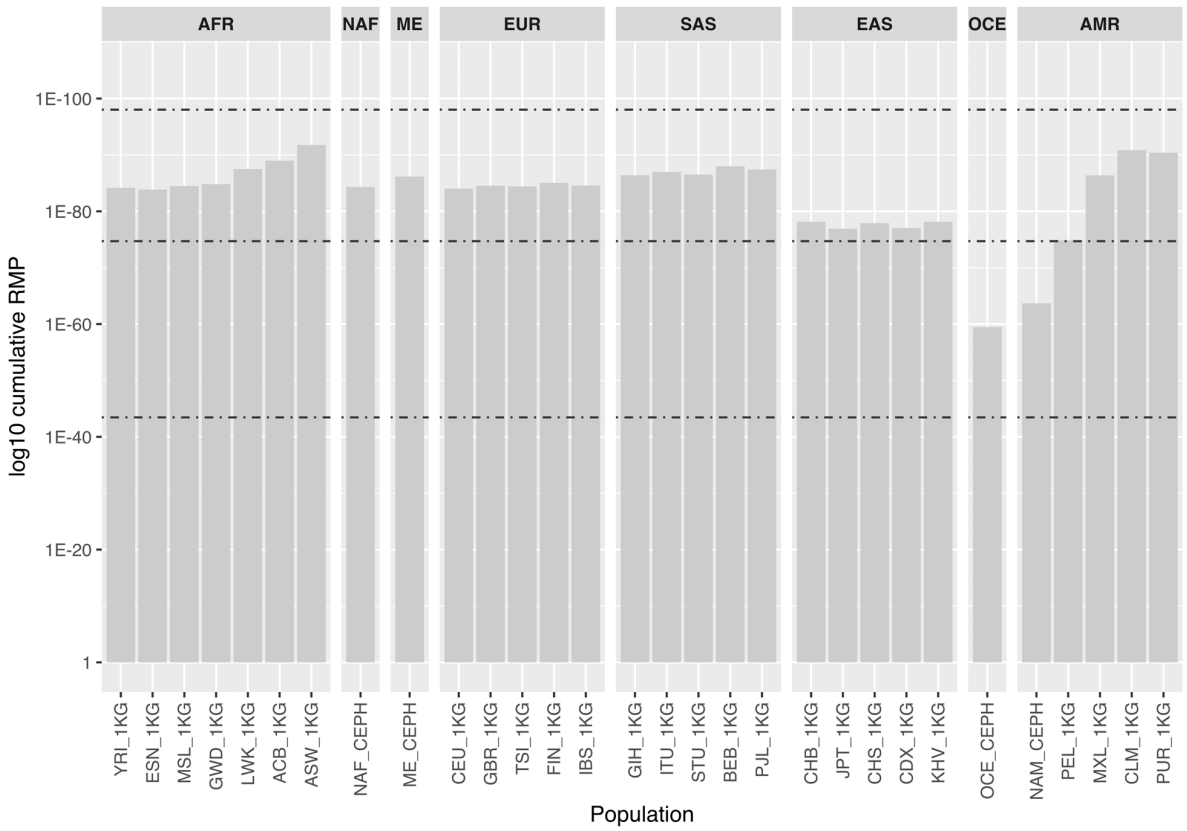


FIGURE 2 | Bar chart represents log10 cumulative random match probability values (i.e., the probability that two individuals share the same profile) for the 30 populations considered, based on the autosomal MH data only. Populations are named and grouped into eight major populations according to **Supplementary 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

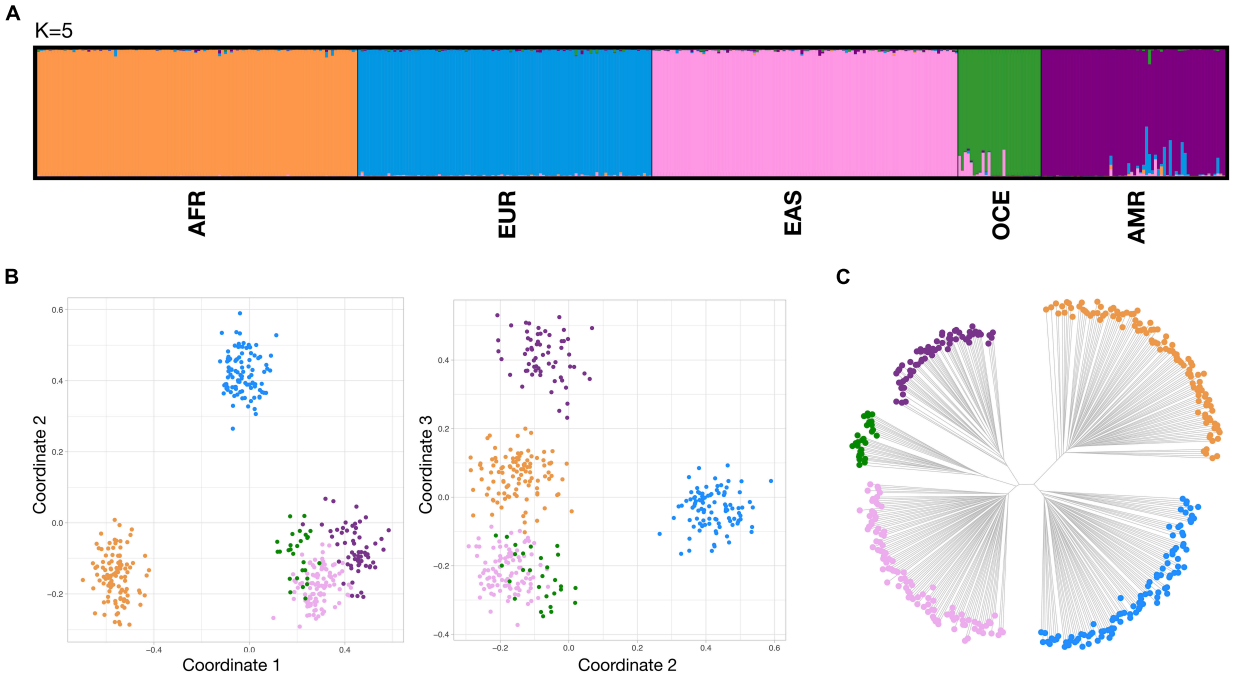


FIGURE 3 | Bio-geographical ancestry analysis of the five continental reference populations. **(A)** STRUCTURE results of ancestry proportions at $K = 5$. Each bar represents an individual and is colored in segments whose lengths correspond to their genetic cluster membership coefficients in up to five inferred population groups. **(B)** Three dimensional MDS analysis showing coordinates 1 and 2 (left) and 2 and 3 (right). **(C)** Neighbor Joining (NJ) tree analysis. For the MDS and NJ-tree plots, populations are colored according to the five different clusters which correspond to the five major populations identified in the STRUCTURE plot.

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**

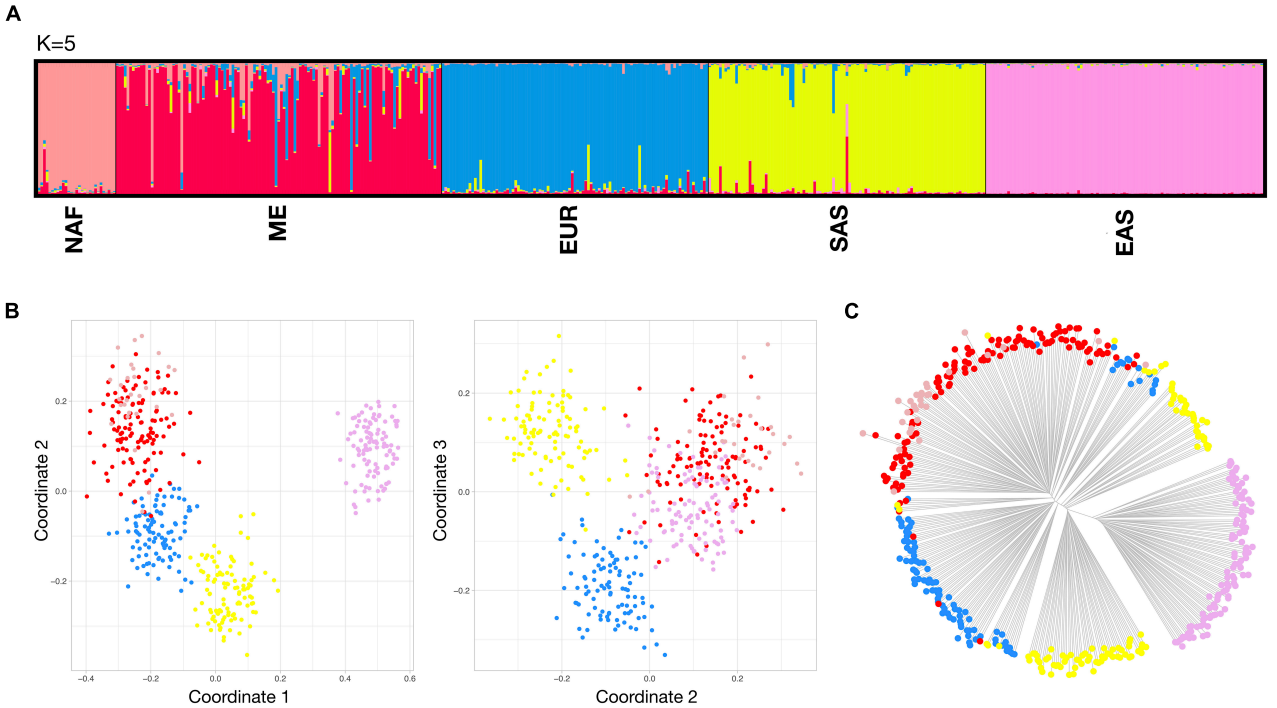


FIGURE 4 | Bio-geographical ancestry analysis of the five NAF-Eurasia reference population sets. **(A)** STRUCTURE results of ancestry proportions at $K = 5$. Each bar represents an individual and is colored in segments whose lengths correspond to their genetic cluster membership coefficients in up to five inferred population groups. **(B)** Three dimensional MDS analysis showing coordinates 1 and 2 (left) and 2 and 3 (right). **(C)** Neighbor Joining (NJ) tree analysis. For the MDS and NJ-tree, populations are colored according to the five different clusters which correspond to the five major populations identified in the STRUCTURE plot.

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.

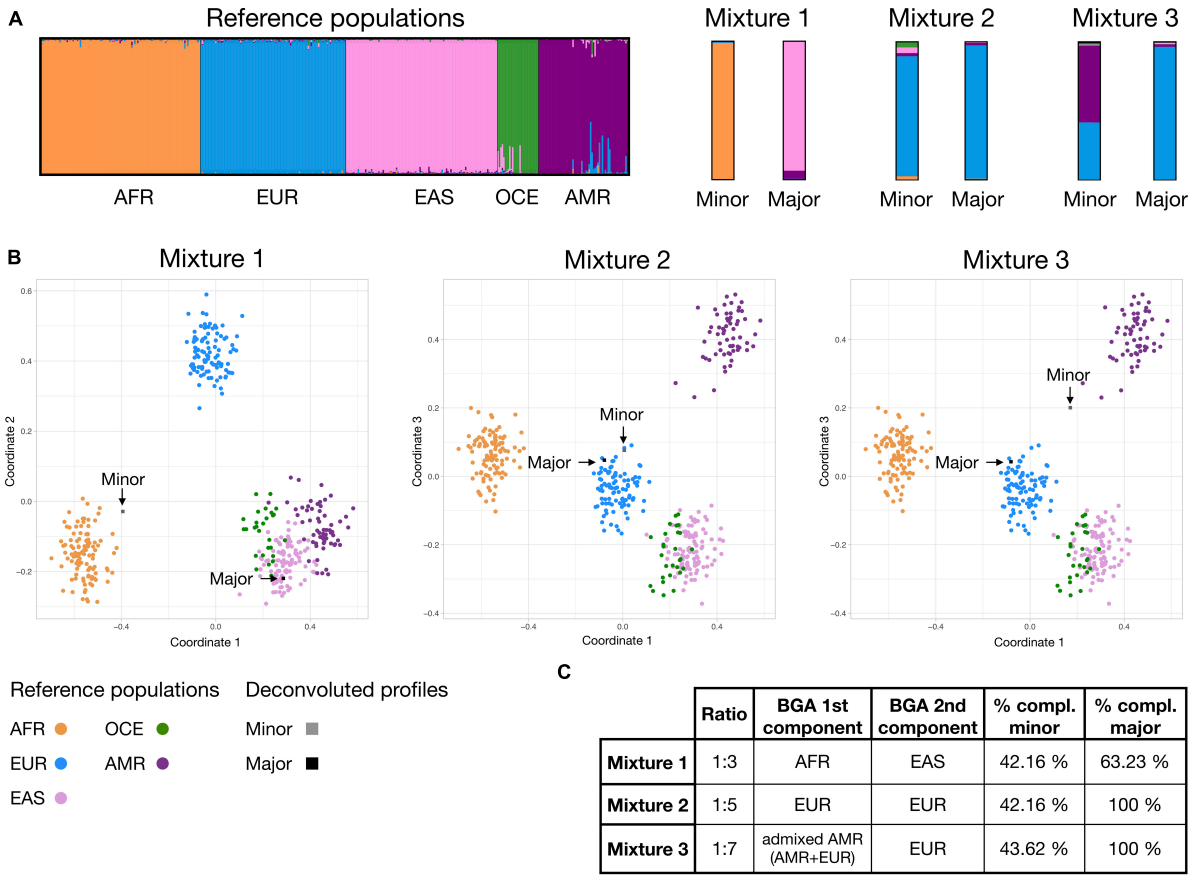
For mixture 1, with the most balanced ratio of 1:3, both the major and minor components resulted in partial profiles after deconvolution, reaching profile completeness percentages of 42.16 and 63.23% respectively. For mixtures 2 and 3, the higher imbalance of the components at ratios 1:5 and 1:7 allowed 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

USC DE COMPOSTELA



C

	Ratio	BGA 1st component	BGA 2nd component	% compl. minor	% compl. major
Mixture 1	1:3	AFR	EAS	42.16 %	63.23 %
Mixture 2	1:5	EUR	EUR	42.16 %	100 %
Mixture 3	1:7	admixed AMR (AMR+EUR)	EUR	43.62 %	100 %

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 costumization 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 sentitivity 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,

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

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 mantaining 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 differentation (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; Daca-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 abscense 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, ÁC, 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, MÁCdC, 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).

USC DE COMPOSTELA

ACKNOWLEDGMENTS

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

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

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., Danecek, 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 McNevin, 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, Á, 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., Kriegel, 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., Kernytsky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

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.fsigss.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., 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. *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

Porras-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

USC DE SANTIAGO DE COMPOSTELA

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

Sobrino, 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.

Article 4. *Eurasiaplex*-2: Shifting the focus to SNPs with high population specificity increases the power of forensic ancestry marker sets

Reference: Forensic Science International. Genetics, 61, 102780

ISSN: 1878-0326

DOI: 10.1016/j.fsigen.2022.102780

Web link:

Authors: Phillips C[1], de la Puente M[1], Ruiz-Ramirez J[1], Staniewska A[1,2], Ambroa-Conde A[1], Freire-Aradas A[1], Mosquera-Miguel A[1], Rodriguez [A], Lareu MV[1]

Author information:
1. Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
2. Institute of Anthropology and Ethnology, Adam Mickiewicz University in Poznań, Poland

Quality metrics (JCR, 2022):
- Journal Impact factor: 3.1
- Quartile: Q1 (legal medicine) and Q2 (genetics & heredity)
- Rank by JCR: 3/17 (legal medicine) and 80/171 (genetics & heredity)

Contribution: methodology, collection of population data

Authorization:



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

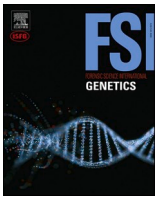
Forensic Science International: Genetics 61 (2022) 102780



Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen





Eurasiaplex-2: Shifting the focus to SNPs with high population specificity increases the power of forensic ancestry marker sets

C. Phillips [a,*], M. de la Puente [a], J. Ruiz-Ramirez [a], A. Staniewska [a,b], A. Ambroa-Conde [a], A. Freire-Aradas [a], A. Mosquera-Miguel [a], A. Rodriguez [a], M.V. Lareu [a]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
[b] Institute of Anthropology and Ethnology, Adam Mickiewicz University in Poznań, Poland

ARTICLE INFO

Keywords:
SNPs
South Asia
Forensic ancestry analysis
Population-specific alleles
1000 Genomes
HGDP-CEPH

ABSTRACT

To compile a new South Asian-informative panel of forensic ancestry SNPs, we changed the strategy for selecting the most powerful markers for this purpose by targeting polymorphisms with near absolute specificity – when the South Asian-informative allele identified is absent from all other populations or present at frequencies below 0.001 (one in a thousand). More than 120 candidate SNPs were identified from 1000 Genomes datasets satisfying an allele frequency screen of ≥ 0.1 (10 % or more) allele frequency in South Asians, and ≤ 0.001 (0.1 % or less) in African, East Asian, and European populations. From the candidate pool of markers, a final panel of 36 SNPs, widely distributed across most autosomes, were selected that had allele frequencies in the five 1000 Genomes South Asian populations ranging from 0.4 to 0.15. Slightly lower average allele frequencies, but consistent patterns of informativeness were observed in gnomAD South Asian datasets used to validate the 1000 Genomes variant annotations. We named the panel of 36 South Asian-specific SNPs *Eurasiaplex-2*, and the informativeness of the panel was evaluated by compiling worldwide population data from 4097 samples in four genome variation databases that largely complement the global sampling of 1000 Genomes. Consistent patterns of allele frequency distribution, which were specific to South Asia, were observed in all populations in, or closely sited to, the Indian sub-continent. Pakistani populations from the HGDP-CEPH panel had markedly lower allele frequencies, highlighting the need to develop a statistical system to evaluate the ancestry inference value of counting the number of population-specific alleles present in an individual.

1. Introduction

We developed the original *Eurasiaplex* forensic single nucleotide polymorphism (SNP) ancestry panel in 2013 [1] specifically to enhance the distinction of European and South Asian ancestries. These population groups are more closely positioned geographically and lack physical barriers to migration, so consequently are genetically less well differentiated than other continentally defined population groups. For a sizeable proportion of their genomic variation, populations of the Indian sub-continent show allele frequencies with variability positioned in the middle of an allele frequency cline running between Europe and East Asia. Key additional patterns of variation are defined by differing ratios of variability from the inferred founding populations of Ancestral North Indians and Ancestral South Indians [2,3]. Such variation therefore tends to have limited informativeness for distinguishing South Asian individuals from Europeans or East Asians. Nevertheless, the *Eurasiaplex* SNPs have proved to be a useful set for supplementing other forensic ancestry panels that have a stronger emphasis on differentiating the five continentally based population groups of Africa, Europe, East Asia, America, and Oceania [4–6]. When considering the development of a new set of South Asian-informative SNPs to improve the power of massively parallel sequencing (MPS) ancestry tests for forensic use, we decided to revise the strategy for selecting optimum AIMs. This was accomplished by a change of focus away from allele frequency contrasts between South Asia and Europe and/or East Asia, towards selecting SNPs with near-absolute population specificity, defined in this respect as variants with zero or extremely low allele frequencies in population groups outside the targeted region. Therefore, despite the new variants identified having allele frequencies as low as 0.1 in South Asian populations, in all other regions the allele frequency of the specific allele is generally considerably lower, with values of 0.001 (1-in-1000) or less. Although an allele frequency of 0.1 is seemingly uncommon variation,

* Corresponding author.
E-mail address: c.phillips@mac.com (C. Phillips).

https://doi.org/10.1016/j.fsigen.2022.102780
Received 3 June 2022; Received in revised form 16 September 2022; Accepted 18 September 2022
Available online 20 September 2022
1872-4973/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

18 % of genotypes will be heterozygotes with the specific allele, and when specific allele frequencies are as high as 0.4, 64 % of genotypes in total carry that allele. These frequencies of a specific allele at any one locus contrast with those in individuals from other regions, including areas neighbouring South Asia, of less than 1 in 1000. As a result, when panels of thirty or more markers with near-absolute specificity are compiled, between 12 and 18 specific-allele genotypes are detected in individuals from the target population group, compared to a maximum of one, two or, much more rarely, three genotypes in individuals from elsewhere.

To develop a second-generation forensic AIMs panel for the analysis of South Asian ancestry, which we called *Eurasiaplex-2*, approximately 120 candidate SNPs were compiled with absolute or near-absolute specificity to South Asia, a region extending from the Indian subcontinent into Afghanistan in the northwest, and Myanmar along with closely sited parts of the SE Asian archipelago in the east. We used the five 1000 Genomes populations with South Asian origins to detect SNPs with specific alleles that had frequencies ranging from 10 % to 40 % in these populations but contrasting with zero frequencies or in the range of 1-in-200 to 1-in-1000 in the project's populations from East Asia, Europe, and Africa. Once identified, the best markers were compiled into a smaller panel of 36 AIMs suitable for inclusion in future globally applicable ancestry panels for SNP genotyping using MPS. The 36 SNPs selected for the *Eurasiaplex-2* panel were cross-checked for consistent South Asian-specific allele frequency patterns amongst the widely dispersed population sampling of five other whole-genome-sequence human diversity projects.

2. Methods and materials

2.1. Changing the concept of population informativeness when selecting forensic ancestry SNPs

Fig. 1 details three ancestry SNPs of potential interest for inclusion in a panel of South Asian-informative markers. SNP rs10008492 was chosen for the original *Eurasiaplex* panel as there is evident differentiation in the rs10008492-C allele frequencies (blue segment) between South Asian and European populations, although it is also evident that this allele only very weakly differentiates East Asian populations. This is reflected in the contrasting pairwise I_n values of 0.16 and 0.02, respectively, shown for each population comparison. The I_n divergence metric is widely used to gauge population differentiations [7] and was the basis for the first *Eurasiaplex* SNP selection process (Fig. 1 of [1]). Another simple metric often used and indicative of ancestry inference power, is the allele frequency differential (δ); the absolute difference of one population's allele frequency from another - in this case, the European-South Asian δ is 0.5, a comparatively high value. SNP rs6053171 provides very good differentiation between Europeans and East Asians, and South Asians have intermediate frequencies for both alleles, consequently giving high I_n and δ, comparing South Asians to both other populations. This SNP was chosen in a study by Pfaffelhuber et al. in 2020 [8] as part of a set of twelve loci to compile an optimum ancestry SNP set for distinguishing African, European, East Asian and South Asian populations, selected using a supervised feature selection system [8]. However, given a South Asian heterozygosity of 48 %, over half of genotypes are homozygous TT or GG and therefore these individuals are not distinguished from Europeans (69 % of TT homozygotes) or from East Asians (31 % of GG). The third SNP rs371763923 has the lowest I_n and δ values of all three loci, so would not be amongst the top selections as an ancestry marker. However, the power of this type of SNP's allele frequency distribution lies in the zero frequency for the rs371763923-G allele (yellow segment) in both Europeans and East Asians. In the 33 % of South Asians where the G allele is detected as a heterozygote, or the 4 % as a homozygote, these genotypes strongly signal origins from this population group as they are not observed elsewhere. Supplementary Fig. S1 details all the individual population allele frequencies in each SNP, indicating there are also zero rs371763923-G allele frequencies in African, Native American, Oceanian and Middle East populations, making this SNP universally specific for South Asia. Only one population, 1000 Genomes KHV (Kinh in Ho Chi Minh City, Vietnam) has two individuals with rs371763923-G alleles, representing < 1 % overall frequency. We selected multiple, well-spaced SNPs on each chromosome that display this kind of highly specific frequency distribution, by applying the strict criterium of the lowest possible specific allele frequency across all populations outside of South Asia. In all cases, the South Asian-specific allele was the Reference Sequence (*RefSeq*) alternative allele, not the reference allele (herein, Alt and Ref alleles, respectively).

2.2. Marker selection

BCFtools was used to make selections of suitable candidate SNPs from publicly available 1000 Genomes Phase III variant catalogues [9]. The chromosome based VCF data from the 1000 Genomes FTP site was searched using the simple allele frequency intersect of: '> 0.1 in South Asian, < 0.01 in African, East Asian, European and admixed American population sample' frequency cut-offs. Multiple-allele variants were excluded, and X-/Y-chromosome variants were identified but not

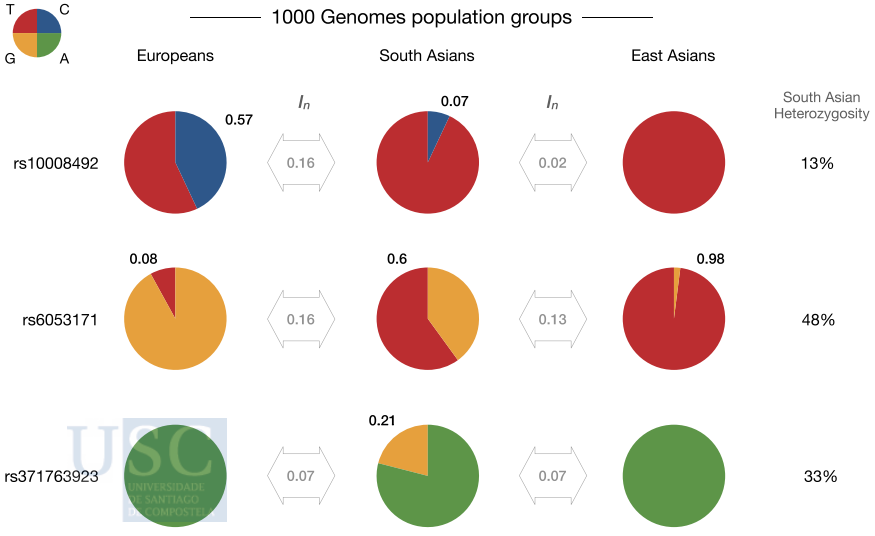


Fig. 1. Three different types of ancestry SNP with potentially South Asian-informative allele frequency distributions. The well-established I_n Divergence measurement of population diversity are shown for each pairwise comparison: European-South Asian left, and South Asian-East Asian right. SNP rs10008492 was part of the original *Eurasiaplex* panel and is informative for the differentiation of Europeans, but not for East Asians. SNP rs6053171 has high Divergence values, but South Asian homozygotes are uninformative for either Europeans or East Asians. Only SNP rs371763923 is equally informative for both comparisons and despite relatively low Divergence values, over 33 % of genotypes would be highly informative heterozygotes or GG homozygotes that are not found in the other populations.

compiled into the final candidate lists for each chromosome. Lists of candidate SNPs were assembled in Excel per chromosome, for further allele frequency comparisons of allele frequencies in African, East Asian, European, and American population samples to maximise the level of South Asian specificity. It should be emphasised that we routinely compile 1000 Genomes variant data from the high sequence coverage datasets generated by the NYGC sequencing of the project's sample set, which has greatly improved the quality of SNP genotype calls across the human genome [10]. The SNP rs3857620 is illustrative of the problem of performing searches for ancestry markers using variant catalogues based on low coverage sequencing data, and which may continue to harbour incorrectly called genotypes. This SNP was identified by Zhao et al., in 2019 [11] as a South Asian-informative marker suitable for a small-scale forensic ancestry panel proposed to consist of 36 SNPs in total. SNP rs3857620 was also adopted for the VISAGE Enhanced Tool for Ancestry and Appearance in a much larger set of SNPs genotyped with MPS [12]. However, the currently listed genotypes in the 1000 Genomes Ensembl portal [13] indicating a 46 % frequency of the A-allele in South Asian populations vs. 0–13 % in the other population groups, contradicts an A-allele frequency estimate across all 1000 Genomes populations sequenced at high coverage, of less than 1 %. For this reason, we cross-checked all current human genome variant databases to ensure each SNP of interest showed consistent patterns across all datasets.

Once compiled, single SNPs were selected from each cluster of markers on any one chromosome segment showing near-identical allele frequency patterns, to optimise the genomic distribution of the final marker set. An exception was made to this rule for chromosome 16, which had a very large extended haplotype of markers with high South Asian specificity at 16p11.2-q11.2. In this case, multiple SNPs were chosen at widely dispersed positions which were clearly part of a large-scale chromosome segment where haplotypes of specific SNP variants had been preserved at very high frequencies in many South Asian populations.

2.3. Human genome variant databases accessed

The SNP variant catalogue of 1000 Genomes Phase III was interrogated with the South Asian-targeted allele frequency intersect described in Section 2.2. All genotypes were cross-checked for accuracy with the NYGC high sequence coverage dataset for each candidate SNP in turn. To further check accuracy of allele frequency estimates made from each set of 1000 Genomes genotypes, the gnomAD (Genome Aggregation Database [14]) v.3.1.2 dataset was checked for all *Eurasiaplex-2* candidate SNPs. The gnomAD database is the largest publicly available collection of population variation compiled from large-scale human genome sequencing projects. It only reports allele frequencies per population but compiles the most up-to-date data from 1000 Genomes (i.e., the NYGC high sequence coverage variant data); whole-genome sequence data for the widely used HGDP-CEPH diversity panel, as well as more than 152,000 samples from large-scale African, European (Finnish and non-Finnish compiled separately), East Asian, South Asian, Middle Eastern, Latino (admixed American), Ashkenazi Jewish and Amish population samples. Studies of patterns of human variation on this scale provide very accurate allele frequency estimates, and gnomAD data is particularly sensitive to very rare variation such as a tri-allelic polymorphism where a second alternative allele (Alt-2) at a SNP site is present in one population at a very low frequency. Genotypes from the HGDP-CEPH panel were obtained from this project's FTP site [15].

Human variant data from Simons Foundation human genome diversity project (herein, SGDP [16]) offers information on geographic areas outside of those covered by 1000 Genomes or the HGDP-CEPH sampling regimes; in particular, Northeast Asia (broadly, Siberia east towards the Bering Straits); Southeast Island Asia; and Central South Asia (broadly, the Caucasus, east towards the central Asian Steppe immediately north of the Hindu Kush, Kashmir, and Tibet). There are 278 samples with genome-wide SNP datasets, of which 22 overlap with 1000 Genomes and 133 overlap with the HGDP-CEPH panel samples, leaving 123 unique samples, from 67 populations, but each a very limited number of 1, 2 or 3 samples per population. The Estonian Biocentre genome diversity panel (herein, EGDP [17]) largely mirrors the sampling regimes of SGDP by being mainly samples of 1, 2 or up to 6 individuals per population or region. EGDP has 402 individual samples from 121 populations that almost all complement the SGDP samples by providing extensive geographic coverage of Siberian, Northeast Asian, Eastern European, and Central South Asian regions. We compiled the 123 SGDP-unique genotype datasets and 402 EGDP genotype datasets for the candidate *Eurasiaplex-2* SNPs.

Lastly, we collected genotypes for the *Eurasiaplex-2* final selection of 36 SNPs from the Singapore Genome Variation Project (herein SGVP), a whole-genome sequencing study from 2014 which included 36 individuals of Indian descent residing in Singapore, plus 96 Malays in Singapore [18,19].

2.4. Statistical considerations

To emphasise the power of specific variation to signal a particular population, Bayes analysis which generates a likelihood ratio (LR) between two possible populations-of-origin starts to build very high cumulative probabilities in SNPs with specific alleles at even moderate frequencies. The highest likelihood ratio for an rs10008492-TT or rs6053171-GG homozygote comparing South Asian and European allele frequencies (shown in Fig. 1) produces a probability of ~4.7 times more likely South Asian. The same likelihood test for an rs371763923-AG heterozygote, and applying a conservative 'global' G allele frequency of 1 % for all non-South Asian populations, produces a probability of 16.75 times more likely South Asian. Given most populations outside of South Asia have a zero frequency for the specific allele of most of the SNPs chosen for *Eurasiaplex-2*, Snipper avoids zero-value numerators in LR calculations by applying the default value of $1/n + 1$; where n is the sample size for that population.

We explored the application of Bayes analysis using the Snipper multiple profiles SNP classifier [20], which accepts multiple SNP profiles and generates principal component analysis (PCA) plots in the same test. Allowance was made for non-independence of linked SNPs when applying the chromosome 16 haplotype SNPs by choosing the 'Hardy--Weinberg principle need not apply' option in Snipper, which adjusts for association of allele frequencies amongst closely sited SNPs on the same chromosome. Additionally, estimations were made of likely recombination rates amongst the haplotype component SNPs by measuring the recombination fraction values between these markers using the HapMap genetic map for this chromosome, as previously described [21].

An alternative to Bayes LR tests and PCA is to run a genetic cluster algorithm such as STRUCTURE [22]. Since we were only compiling markers specific to one population group, there is limited information that can be obtained from STRUCTURE analyses seeking two genetic clusters (i.e., setting analysis runs for a K value of 2). Nevertheless, we performed a simple comparison of STRUCTURE analyses of 1000 genomes populations using *Eurasiaplex* and *Eurasiaplex-2* SNP sets to explore the extra power potentially gained from alleles with absolute specificity to a single population group.

A much simpler and potentially informative alternative to all three of the established population analysis systems, is to simply count the number of specific alleles found in any one individual and assess if these match the patterns observed across the whole region of interest. This was done in the current study on a large scale with the five South Asian 1000 Genomes sample sets and included HGDP-CEPH samples plus those on a much smaller scale, but geographically dispersed and covering a wide range of different populations from the Indian subcontinent samples of SGDP and EGDP.

3. Results

3.1. Screening South Asian-specific candidate SNPs

A total of 123 candidate South Asian-specific SNPs were compiled by applying the allele frequency intersect to the 1000 genomes Phase III variant catalogue. The full candidate SNP set is listed with summary genomic details and complete genotype data (including 1000 Genomes NYGC high coverage, HGDP-CEPH, SGDP and EGDP genotypes) in Supplementary Tables S1A and S1B, respectively. Genotype concordance was checked by comparing the currently published 1000 Genomes Phase III data used for the allele frequency screening, with the high coverage genotype calls from the NYGC re-sequenced 1000 Genomes samples. The 2,505 sample-by-sample comparisons for each of the 123 candidate SNPs are listed in Supplementary Table S1C.

An initial filter set was applied to exclude SNPs that had one of three characteristics: i. genotype discordancy rates higher than 20 incompatibilities (approximately 1 % or more differences in genotype calls); ii. SNPs that lost the expected specific allele pattern when the NYGC genotypes of South Asian samples were compared to those of the other populations; iii. SNP pairs which were physically well separated on the same chromosome segment, but which showed identical allele frequencies indicating they were in linkage disequilibrium - in such cases, one SNP was chosen. Note that the third screening rule was not applied to SNPs in the chromosome-16 extended haplotype. Fig. 2A shows the ten SNPs with more than 20 discordant genotypes, excluded from further consideration. No suitable SNPs were identified on chromosome 21, and the single SNP on chromosome 22, rs113693449, was excluded due to a high level of genotyping discordancy. Fig. 2B provides summary allele frequency charts for the SNPs that were expected to have South Asian-specific alleles in searches of the 1000 Genomes Phase III data, which were not detected in the high coverage NYGC data: rs3857620, rs199671447 and rs113693449. The NYGC data for SNP rs11103281 maintained the allele expected in South Asian samples, but it was also detected in the other population groups so lacked specificity. All four of the above SNPs had zero South Asian-specific allele frequencies in the Gujarati in Houston US (GIH), which suggested discrepancies in the way these SNPs had been genotyped - notably that the GIH population samples were originally studied by HapMap, and this data may simply have been merged with the South Asian populations added to 1000 Genomes Phase III studies. SNP rs9915709 had a much higher frequency in African samples than South Asians, therefore although this SNP was listed, in practice it would be less informative than other SNPs with zero, or near-zero allele frequencies across all non-South Asian populations. Finally, SNP rs371441513 had the highest level of genotype discordancy, which suggests it has sequencing issues, meaning it is unlikely to be reliably genotyped with any assay developed for *Eurasiaplex-2* SNPs.

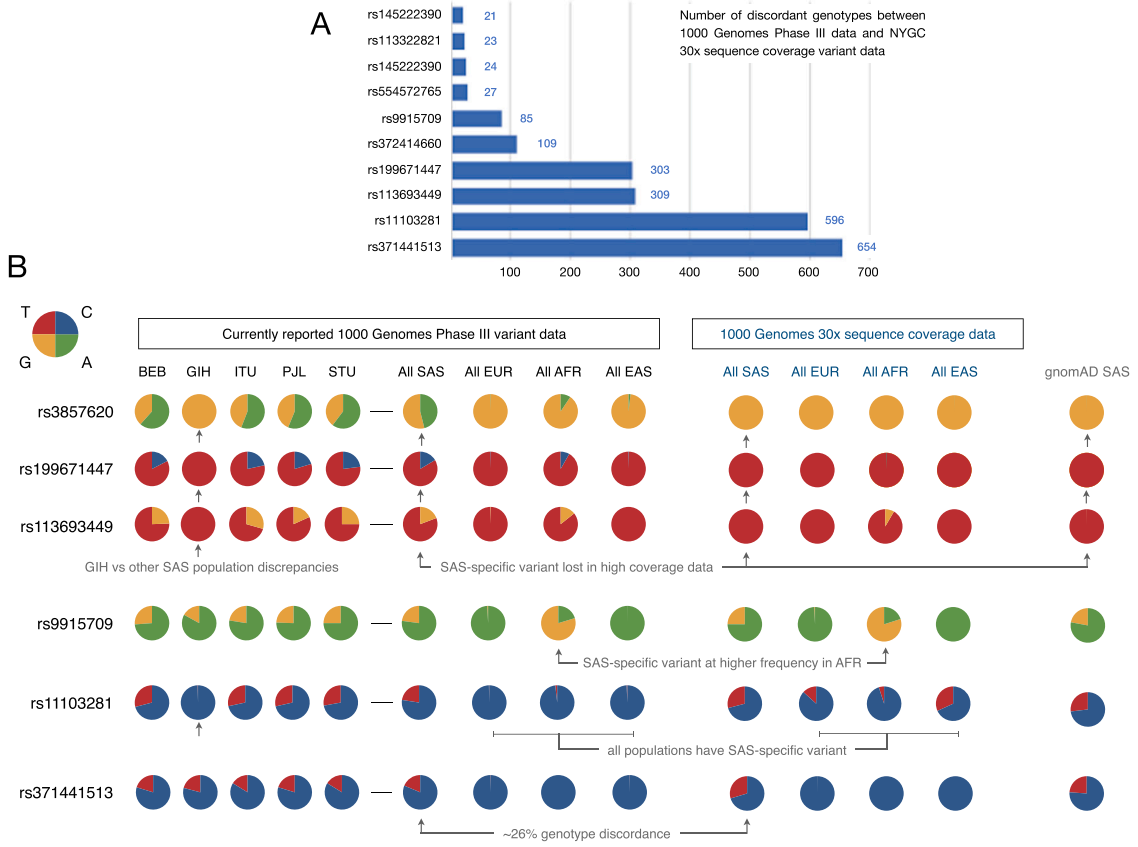


Fig. 2. (A) Ten *Eurasiaplex-2* candidate SNPs with discordant genotypes between the current 1000 genomes Phase III data (2–3x sequence coverage) and the high coverage (30x) NYGC re-sequenced samples. Although the top four have relatively low numbers of discordant genotypes, the other six indicate sequence alignment problems or complex sites (e.g., with closely positioned Indels) and all were rejected to minimise the risk of genotyping problems using MPS. (B) Three *Eurasiaplex-2* candidate SNPs with misleading allele frequency information in the current 1000 Genomes Phase III variant dataset (upper left-hand pie charts for South Asian populations and population group summaries), contrasting with the allele frequency information from the higher coverage sequence analysis data for the same samples, plus gnomAD South Asian data, where frequency estimates are based on ~4800 samples (upper right-hand charts). Lower pie charts show additional problems of insufficient specificity in rs9915709 (South Asian-specific allele at a higher frequency in African populations) and rs11103281 (all population groups have the South-Asian specific allele at >0.05 frequencies), plus SNP rs371441513 with the highest recorded genotype discordancy rate indicating a potentially complex variant site.

3.2. Selecting a core set of South Asian-specific SNPs for Eurasiaplex-2

3.2.1. Patterns of South Asian-specific genotype distributions

Table 1 outlines genomic details and summary allele frequencies of the 36 SNPs selected for *Eurasiaplex-2*. It is noteworthy that only the two SNPs rs77510889 and rs17158407 were previously identified in the 1000 Genomes Phase I variant catalogue (accessible in the SPSmart ENGINES genome browser [23]). This is mainly due to an absence of South Asian population samples in this first 1000 Genomes SNP genotyping project phase, so a SNP with the Alt allele only present in South Asian populations would consist entirely of Ref allele homozygotes in all Phase I populations and thus not be identified as a variant. SNPs were given internal codes based on their chromosome and *RefSeq* 5′ to 3′ locations, from 1A to 20 (i.e., one SNP on this chromosome). Generally, gnomAD South Asian-specific allele frequencies were slightly lower than most or all of those in 1000 Genomes. With the exception of rs374908464, the CEPH Pakistani allele frequencies averaged across the eight populations, are substantially lower than either 1000 Genomes or gnomAD South Asian samples. The overall average South Asian-specific allele frequency of 0.085 in CEPH Pakistani samples is two- to three-times lower, and for 27 of the 36 SNPs would not meet the selection criteria of > 0.1 allele frequencies.

The full list of genotypes and summary allele frequency estimates for the 36 SNPs in each population group are listed in Supplementary Table S1D. For each of the 4097 samples listed in this table, numbers of South Asian-specific genotypes and alleles were counted. The individual South Asian-specific genotype counts are plotted as blue bars in Fig. 3A (the four 1000 Genomes population groups), and Fig. 3B (all other samples from HGDP-CEPH, SGDP, EGDP and SGVP SNP databases), with the small CEPH Uyghur and EGDP Roma population samples highlighted as red bars for clarity. Aligned above each set of bar plots in Fig. 3A and B are graphic representations of South Asian-specific allele homozygotes (red lines) and heterozygotes (orange lines), with non-specific allele homozygotes in grey. All 36 of these non-specific alleles are the *RefSeq* reference allele. These graphics highlight the large number of informative genotypes recorded in the 1000 Genomes South Asians, with only rs77510889 (internal code 'SNP 10′) indicating a higher-than-average number of South Asian-specific genotypes in 1000 Genomes African and CEPH African populations. Note that there are four SNPs not genotyped in EGDP, five in SGVP Singapore Indian, and two in Singapore Malay genome datasets. In each of the other databases, South Asian samples are clearly indicated by blue bars with prominent heights and the corresponding dense patterns of orange and red lines. The average number of informative genotypes per individual is given above each South Asian population box (values for the individual CEPH Pakistani populations below the bar plots). The contrast is evident between the CEPH Pakistani populations and 1000 Genomes South Asian populations, with an average of ~14 specific genotypes per individual in BEB, ITU, GIH, STU populations, dropping to less than 12 in PJL, Punjabi from Lahore, Pakistan. The CEPH Pakistani populations have much lower average numbers of informative genotypes per individual, which range from ~8 in the Sindhi to ~2 in the Hazara samples. In the other datasets, the SGDP and EGDP South Asian, plus SGVP Singapore Indian samples have close to an average of 11 informative genotypes per individual (adjusted for the missing SNPs), which can be taken to represent an overall median number of informative genotypes for this target population group in the 36 SNPs selected. The other 1000 Genomes population groups have the expected very low average informative genotype value of less than 0.2, allowing a degree of differentiation to be made between East Asians and the South Asian-related samples of CEPH Uyghur, occupying regions to the northeast of the Indian sub-continent; and EGDP Roma, a trans-national cultural isolate suggested to have originated from a proto-Romani population living in northwest India [24]. The Uyghur have relatively low average informative genotype levels of 1.8, but the Roma, although only five individuals and based on 32/36 SNPs, show a high average value of ~6.7 informative genotypes.

3.2.2. South Asian-specific allele frequency estimates from Eurasiaplex-2 SNP genotypes

Although genotype frequencies provide indications of the population informativeness of the SNPs selected for *Eurasiaplex-2*, examination of allele frequencies allows a clearer overview of how the SNP variability is distributed across the Indian sub-continent. When population-wide average South Asian-specific allele frequencies are calculated for each of the 36 SNPs from 1000 Genomes combined South Asian populations, gnomAD samples (approximately 5000 South Asians, with no specific geographic data provided), and CEPH Pakistani datasets, the contrast between Pakistan and the rest of South Asia is further underlined. Fig. 4 gives bar plots for the South Asian-specific average allele frequencies per SNP in each dataset, ranking the markers from most informative to least, left to right. The Fig. 4 plots suggest a close match in frequency estimates between 1000 Genomes and gnomAD data, but much lower specific allele frequency estimates for Pakistanis that are between 10 % and 25 % of the other dataset values. The reciprocal bar plots in Fig. 4 use a 100-fold smaller scale and show most South Asian-specific allele frequencies in other parts of the world rarely exceed 0.001–0.003, or a maximum of 1 in 300. The boxed values for rs77510889 indicate high frequencies for the G-allele outside of South Asia, which exclude population data where this allele was at a particularly high frequency - notably the HGDP-CEPH hunter-gatherer populations of San, Mbuti Pygmies and Biaka Pygmies. South Asian-specific alleles in the rest of the world for rs186371551 and rs184748067 also had higher-than-average frequencies but were more widely dispersed.

Although there are very different sample sizes between the populations sampled by 1000 Genomes, CEPH, SGDP and EGDP (approximately 100, 25, 2–3, and 2–6, respectively), it is instructive to map the distribution of South Asian-specific allele frequencies in all the populations studied which are located in, or near, the Indian sub-continent. Fig. 5 provides pie charts of the average percentage of South Asian-specific alleles in each population sample in 1000 Genomes, CEPH, SGDP and EGDP (percentage values adjusted for four missing SNPs in EGDP) from in or near the Indian sub-continent, mapped to their approximate geographic locations. The UK-resident 1000 Genomes Indian Telugu (ITU) and Sri Lankan Tamil (STU); and US-resident Gujarati (GIH) populations are placed in their approximate locations of origin, and the EGDP Roma have no sampling location described. The CEPH Cambodian and EGDP Aeta from the Philippines are also too distant from the centre of South Asia to be easily placed on this map. Note that the average percentage of South Asian-specific alleles in all other population samples not shown in Fig. 5 was less than 0.5 %, except the two SGVP samples: Indians in Singapore = 20 %; Malays in Singapore = 0.76 %. Patterns of average allele frequency distributions in the 31 population samples shown in Fig. 5 indicate a high percentage value in most populations within India and Bangladesh, but a sharp drop in these values in populations from regions northwest and east of this country. A notable exception is the 1000 Genomes Punjabi from Lahore, Pakistan, with this population group occupying the most easterly part of Pakistan at the northwest corner of India. The smallest recorded average percentage of South Asian-specific alleles were observed in the most geographically distant Iranians in the West (1.38 %), Uyghur in the north (2.5 %) and Cambodians in the east (2 %).

3.3. Analysis of the six Eurasiaplex-2 SNPs on chromosome 16

Six SNPs chosen from chromosome 16 (C16) had the highest levels of South Asian specificity and three were clustered around the centromere, potentially meaning they could show full or high levels of allelic association due to reduced recombination, precluding their use as independent loci. Table 2 summarises the recombination rates between the six SNPs using the HapMap genetic map database to estimate map distances in Centimorgans (cM) and Kosambi-adjusted recombination fractions (Rc). The Rc estimates in Table 2 indicate the three 5′ SNPs rs368479296-rs376893831-rs370130302 show minimal linkage with

Table 1
Genomic details and South Asian-specific allele frequencies from 1000 Genomes, HGDP-CEPH and gnomAD databases of an optimum set of 36 SNPs.

Genomic details								Frequency of the alternative (South Asian-specific) allele in 1000 Genomes groups/South Asian populations								HGDP-CEPH	gnomAD
No.	Code	Chr.	GRCh37	GRCh38	rs-number	Ref.	Alt.	Gene	African	European	East Asian	BEB	GIH	ITU	PJL	STU	Pakistani	South Asian
1	1A	1	27588988	27262497	rs191008849	T	C	WDTC1	0	0.0010	0.0010	0.2267	0.2500	0.2353	0.1927	0.2353	0.2171	0.1866
2	1B	1	207023473	206850128	rs370300597	C	G	–	0	0	0	0.2093	0.1977	0.1961	0.1302	0.2549	0.2072	0.1459
3	2A	2	18702265	18520999	rs373262633	A	G	LOC105373454	0	0	0	0.1802	0.1977	0.2451	0.1354	0.1765	0.1836	0.1336
4	2B	2	98816440	98199977	rs183145214	A	G	VWA3B	0	0	0.0050	0.2151	0.1744	0.1520	0.2031	0.2157	0.1873	0.1861
5	3A	3	44250057	44208565	rs578118259	T	G	–	0	0	0.0010	0.2558	0.2791	0.2843	0.1875	0.2353	0.2444	0.1817
6	3B	3	159603038	159885249	rs369609492	C	A	SCHIP1	0	0	0	0.2500	0.2209	0.2157	0.0938	0.1912	0.1836	0.1572
7	3C	3	167730032	168012244	rs375081853	A	G	GOLIM4	0	0	0.0010	0.1570	0.1512	0.2059	0.1615	0.1912	0.1873	0.1588
8	4A	4	115827968	114906812	rs182767282	T	C	NDST4	0	0.0010	0.0010	0.1453	0.1919	0.2157	0.2031	0.2598	0.2270	0.1703
9	4B	4	152007237	151086085	rs146398591	A	G	–	0	0.0010	0.0030	0.2733	0.2500	0.3235	0.1875	0.2794	0.2457	0.2192
10	4C	4	167677008	166755857	rs554572765	A	C	SPOCK3	0	0	0	0.1860	0.1919	0.1912	0.1458	0.2010	0.1811	0.1068
11	5	5	33704125	33704020	rs375710694	T	C	ADAMTS12	0	0	0.0020	0.1512	0.1628	0.2206	0.1615	0.2059	0.1923	0.1468
12	6A	6	117206569	116885406	rs186371551	G	A	RFX6	0	0.0050	0.0050	0.1279	0.1686	0.2304	0.1823	0.1716	0.2084	0.1703
13	6B	6	130116672	129795527	rs368650154	C	T	–	0	0	0.0010	0.1686	0.1395	0.2255	0.1615	0.2402	0.1935	0.1578
14	6C	6	154459950	154138815	rs368661757	C	A	OPRM1	0	0	0.0010	0.2384	0.2500	0.2451	0.2135	0.2549	0.2444	0.1946
15	7	7	50238881	50199285	rs368444091	C	T	–	0	0	0	0.1686	0.1860	0.1814	0.1667	0.1912	0.1774	0.1396
16	9	9	97560517	94798235	rs187619767	C	T	AOPEP	0	0	0.0020	0.2442	0.2849	0.1618	0.1458	0.2255	0.1849	0.1740
17	10	10	122095086	120335574	rs77510889*	A	G	–	0.0714	0	0.0020	0.2267	0.2093	0.2304	0.2031	0.2206	0.2171	0.1593
18	11A	11	29262859	29241312	rs370097977	T	C	–	0	0	0.0050	0.1628	0.1686	0.1912	0.2031	0.2451	0.2258	0.1529
19	11B	11	59462759	59695286	rs375766368	G	A	–	0	0	0.0010	0.1453	0.1512	0.2157	0.1771	0.1520	0.1824	0.1222
20	11C	11	72175159	72464115	rs377589165	G	A	–	0	0	0	0.1512	0.1860	0.2206	0.1615	0.1863	0.1873	0.1514
21	12A	12	4268703	4159537	rs376263717	C	T	–	0	0.0020	0.0010	0.2267	0.2035	0.2598	0.1510	0.2010	0.1911	0.1655
22	12B	12	22570119	22417185	rs371763923	A	G	–	0	0	0.0020	0.2500	0.2442	0.2059	0.1510	0.2304	0.2022	0.1636
23	12C	12	50428379	50034596	rs368764180	A	G	RACGAP1	0	0	0.0020	0.1570	0.1395	0.1961	0.0885	0.1863	0.1787	0.1388
24	13	13	56057499	55483364	rs184748067	G	A	–	0	0.0060	0.0069	0.1744	0.1802	0.2353	0.1615	0.2304	0.1998	0.1600
25	14	14	65712298	65245580	rs189013802	G	A	–	0	0	0.0079	0.1744	0.1744	0.2402	0.1823	0.2059	0.1873	0.1372
26	15	15	83236825	82568075	rs17158407*	C	T	CPEB1	0	0.0010	0.0020	0.2558	0.2558	0.2990	0.3125	0.3333	0.2990	0.2681
27	16A	16	3178971	3128970	rs368479296	C	T	ZNF213-AS1	0	0	0.0020	0.1802	0.1802	0.1422	0.1667	0.2108	0.1836	0.1645
28	16B	16	23053815	23042494	rs376893831	G	T	–	0	0	0.0010	0.2849	0.2674	0.1569	0.1823	0.1961	0.2146	0.1991
29	16C	16	28588059	28576738	rs370130302	C	G	SGF29	0	0	0.0010	0.2907	0.2674	0.1961	0.1719	0.2206	0.2134	0.2024
30	16D	16	33921593	34119126	rs368738705	C	T	–	0	0	0	0.4186	0.4302	0.4853	0.3385	0.4412	0.4007	0.3329
31	16E	16	46499858	46465946	rs368538881	C	T	–	0	0	0	0.4128	0.4070	0.4951	0.3021	0.4608	0.4020	0.3549
32	16F	16	48327788	48293877	rs377323011	A	G	LONP2	0	0	0.0010	0.3314	0.2965	0.4314	0.2760	0.3775	0.3362	0.2742
33	17A	17	43964966	45887600	rs369091847	A	T	MAPT-AS1	0	0	0	0.1337	0.1512	0.2402	0.1250	0.2451	0.2035	0.1604
34	17B	17	80660204	82702328	rs376153825	G	C	LOC105376791	0	0	0.0010	0.1279	0.1279	0.1765	0.1771	0.1716	0.1762	0.1273
35	19	19	8371240	8306356	rs374908464	A	G	CD320	0	0	0.0020	0.1802	0.2093	0.2500	0.1719	0.1078	0.1985	0.1740
36	20	20	4987550	5006904	rs186201674	C	T	SLC23A2	0.001	0.0040	0.0050	0.2093	0.2209	0.2255	0.1875	0.2451	0.2109	0.1545
					* SNP also identified in 1000 Genomes Phase I			Average:	0.002	0.001	0.002	0.214	0.216	0.240	0.182	0.233	0.219	0.178

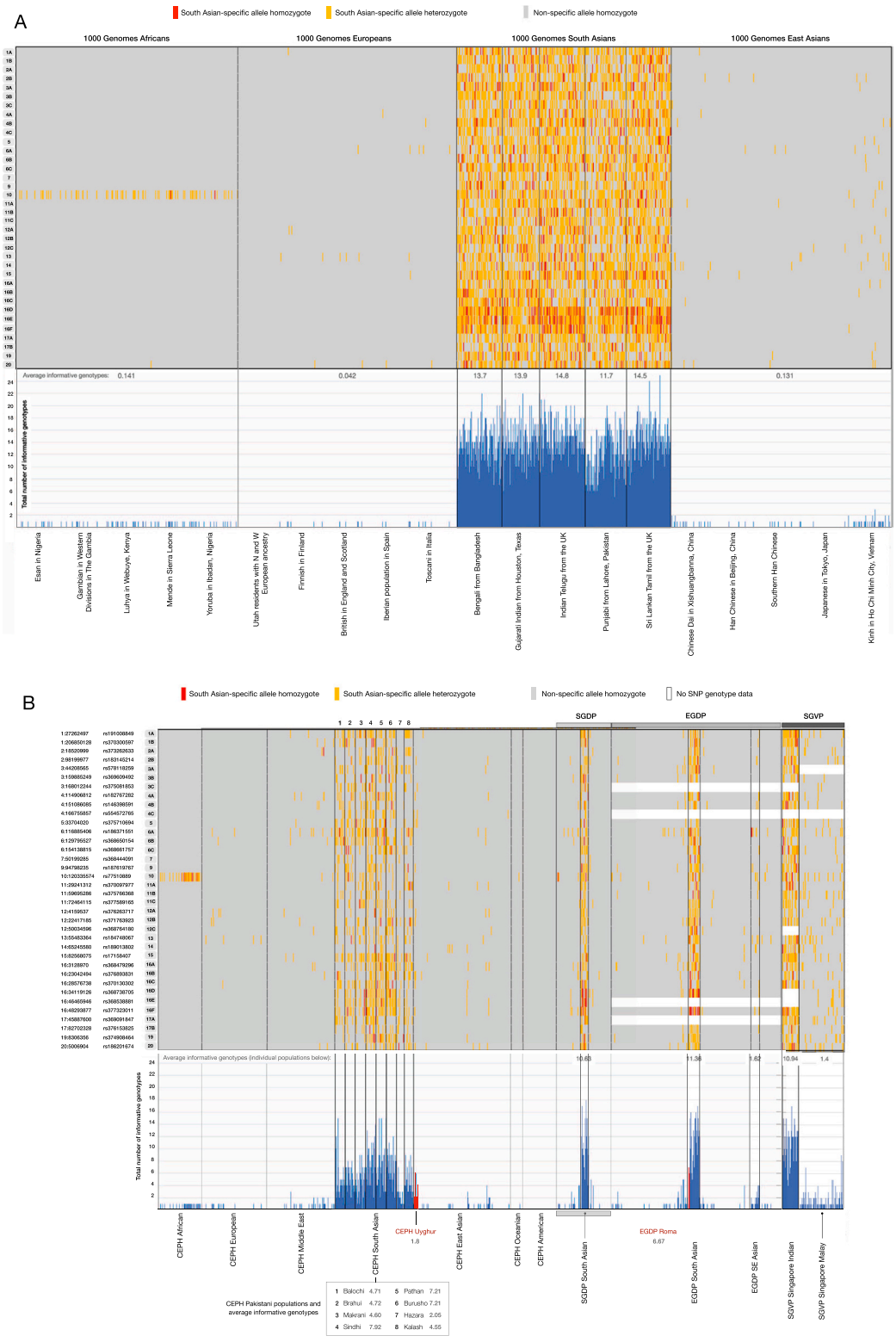


Fig. 3. (A) South Asian-specific genotype distributions and total number of informative genotypes in each 1000 Genomes sample; specific allele homozygotes marked in red and heterozygotes in orange (both genotypes counted singly) of 36 *Eurasiaplex-2* SNPs in twenty 1000 Genomes populations. The average number of informative genotypes are shown as group-wide values for Africans, Europeans, and East Asians, and for individual populations for South Asians. Internal codes are used to label each SNP, which are detailed in B. (B) South Asian-specific genotype distributions and total number of informative genotypes in four whole-genome-sequencing human diversity projects, additional to 1000 Genomes: HGDP-CEPH diversity panel; Simons Foundation genome diversity project (SGDP); Estonian Biocentre diversity project (EGDP); Singapore genome variation project. HGDP-CEPH Uyghur and EGDP Roma geographic outlier samples are marked in red. The average number of informative genotypes are shown individually for eight HGDP-CEPH Pakistani populations and HGDP-CEPH Uyghur; SGDP South Asian samples; EGDP South Asian, SE Asian and Roma samples; SGVP Singapore Indian and Singapore Malay samples. Note EGDP lacks data for four SNPs, SGVP lacks data for five in Singapore Indian and two in Singapore Malay samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

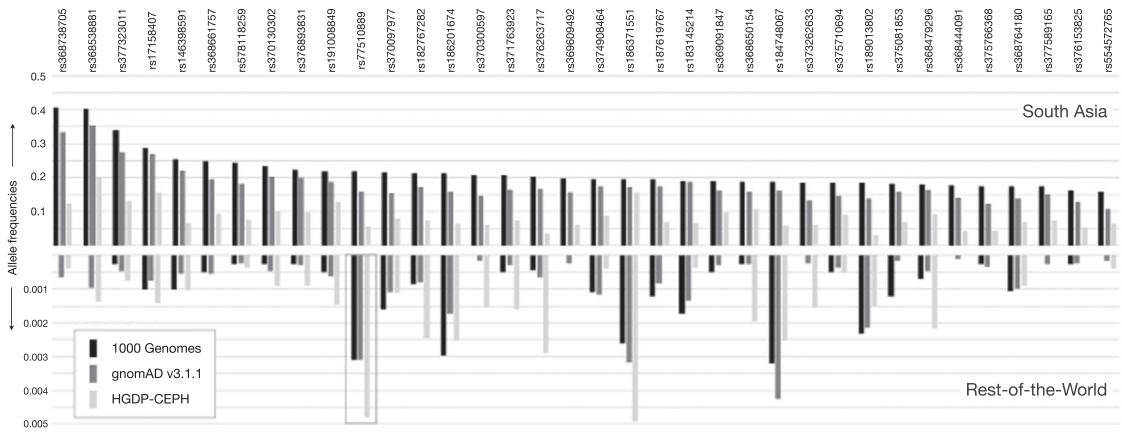


Fig. 4. Allele frequency spectra of the 36 *Eurasiaplex-2* SNPs. Markers are arranged in descending 1000 Genomes South Asian-specific average allele frequency and show bars for 1000 Genomes South Asians, gnomAD South Asians and CEPH Pakistanis. Rest-of-the-World average frequencies are compiled from 1000 Genomes African, European, and East Asian average frequencies, plus gnomAD and HGDP-CEPH average non-South Asian population data. The Rest-of-the-World axis shows 1/100th allele frequency values compared to those of South Asian populations. The boxed values of rs77510889 exclude high frequencies for the South Asian-specific allele amongst HGDP-CEPH African Khoisan, Biaka Pygmies and Mbuti Pygmies (rs77510889-G allele frequency of 0.23 in HGDP-CEPH African hunter-gatherer populations vs. 0.06 in HGDP-CEPH Pakistani populations).

32 % and 11 % recombination fractions, but the other three 3′ centromeric SNPs rs368738705-rs368538881-rs377323011 have a much lower level of recombination of ~0.35 % across the 3-SNP span. The top half of Fig. 6 outlines the structural landscape of these three C16 centromeric SNPs and the common genotype combinations they form in South Asian vs. other populations.

To explore the possibility of association between the rs368479296-rs376893831-rs370130302 SNPs, we decided to treat them as a haplotype and gauge haplotype diversity in the 1000 Genomes samples (excluding admixed samples). Although the Rc values between these SNPs are very low, the physical distances in megabases (Mb) are much bigger, with the 12.35 Mb span between rs368738705 and rs368538881 alone representing nearly 14 % of the total C16 length. Therefore, it would not be possible for 1000 Genomes to accomplish accurate phasing of alleles for this series of SNPs over such distances. We decided to convert any localised phasing (i.e., a SNP allele's phase with reference to its immediate neighbours) made by 1000 Genomes of rs368479296-rs376893831-rs370130302 heterozygotes, into alphabetic order, respectively: TC>CT, TC>CT, GA>AG. This created the root haplotype of CCA – universally present in all 1000 Genomes non-South Asian samples, plus the South Asian-specific TTG haplotype, exclusively confined to these populations in 1000 Genomes. In this way, any 3-SNP genotypes which are not CC-CC-AA; TT-TT-GG; or, CT-CT-AG, represent disruptions to the South Asian-specific haplotypes which signify reduced allelic association. We counted the derived haplotypes amongst South Asian individuals as homozygous genotypes in each SNP, specifically (in bold): **CC**-CT-AG and **TT**-CT-AG in rs368479296; CT-**CC**-AG and CT-**TT**-AG in rs376893831; CT-CT-**AA** and CT-CT-**GG** in rs370130302. These patterns can be interpreted to indicate disrupting recombination between rs368479296-rs376893831, double recombination between rs368479296-rs376893831 and rs376893831-rs370130302, or recombination between rs376893831-rs370130302, respectively. Although true phasing cannot be achieved, the extent to which the above six derived haplotypes occur in 1000 Genomes populations will indicate the level of disruption of allelic association amongst the three SNPs. The rs368479296-rs376893831-rs370130302 inferred haplotypes for all population samples are listed in Supplementary File S1E and the numbers of each haplotype in 1000 Genomes populations are summarised in the lower half of Fig. 6. Except for a singleton CCG haplotype (present as CC-CC-AG genotypes in a Vietnamese KHV sample), all 3,021 non-South Asian samples from 1000 Genomes had the CCA root haplotype. Amongst the South Asian samples, 206 were inferred to have specific TTG haplotypes and 464 non-specific CCA haplotypes, but a significant number of South Asian specific haplotypes, a total of 308, were derived, i.e., inferred to be different combinations of alleles to either CCA or TTG. This would suggest there is very little association between the centromeric C16 SNPs. Furthermore, levels of recombination between these three SNPs are likely to be higher, given the South Asian individuals with CT-CT-AG genotypes (88 of 489, 18 %) cannot be reliably phased.

3.4. Statistical analyses

3.4.1. Conventional Bayes analysis of South Asian population variability

The results of the Bayes analysis likelihood assessments and PCA patterns generated by Snipper, are summarised in Supplementary File S1. First, evaluations were made of pairwise cumulative Divergence ($\boldsymbol{I_n}$) values calculated for South Asian vs. European, and vs. East Asian populations, using 1000 Genomes data and comparing the 23 SNPs of the original *Eurasiaplex* with the 36 of *Eurasiaplex-2* panel. Supplementary File S1.1 shows the cumulative $\boldsymbol{I_n}$ for South Asians vs. Europeans at the point of 23 SNPs have similar values in *Eurasiaplex* (1.87) compared to *Eurasiaplex-2* (2.01), but for South Asian vs. East Asian variation *Eurasiaplex* only reaches 0.77, compared to 1.99 from 23 *Eurasiaplex-2* SNPs. This highlights how frequencies for the specific alleles of zero, or near zero outside of the targeted population, produce almost identical $\boldsymbol{I_n}$ values in all populations comparisons and for each SNP added. This is expressed in the cumulative $\boldsymbol{I_n}$ chart in Supplementary File S1 as two diverging and flattening curves in *Eurasiaplex*, compared with the two straight lines with identical trajectories in *Eurasiaplex-2*. Therefore, when the final cumulative values are calculated for 36 *Eurasiaplex-2* SNPs, each population comparison has almost identical values of 2.84 (vs. Europe) and 2.79 (vs. East Asia). When population specific SNPs are combined in the future to differentiate all the main population groups, it will be straightforward to balance the $\boldsymbol{I_n}$ values for each comparison as this will just entail adjustment of the number of SNPs needed to reach a final cumulative value that can be matched across all populations.

Second, the distribution of likelihood ratios (LR) from the comparison of 1000 Genomes South Asian and African likelihoods (Africans produced the second highest likelihoods in five population comparisons) using Bayes analyses in Snipper, was compiled in a chart of ranked LRs shown in Supplementary File S1.2. These values are generally much higher than those observed with alternative ancestry SNPs [4,6], with the bulk of samples producing LR values in excess of 1E+ 12 or '1 in a

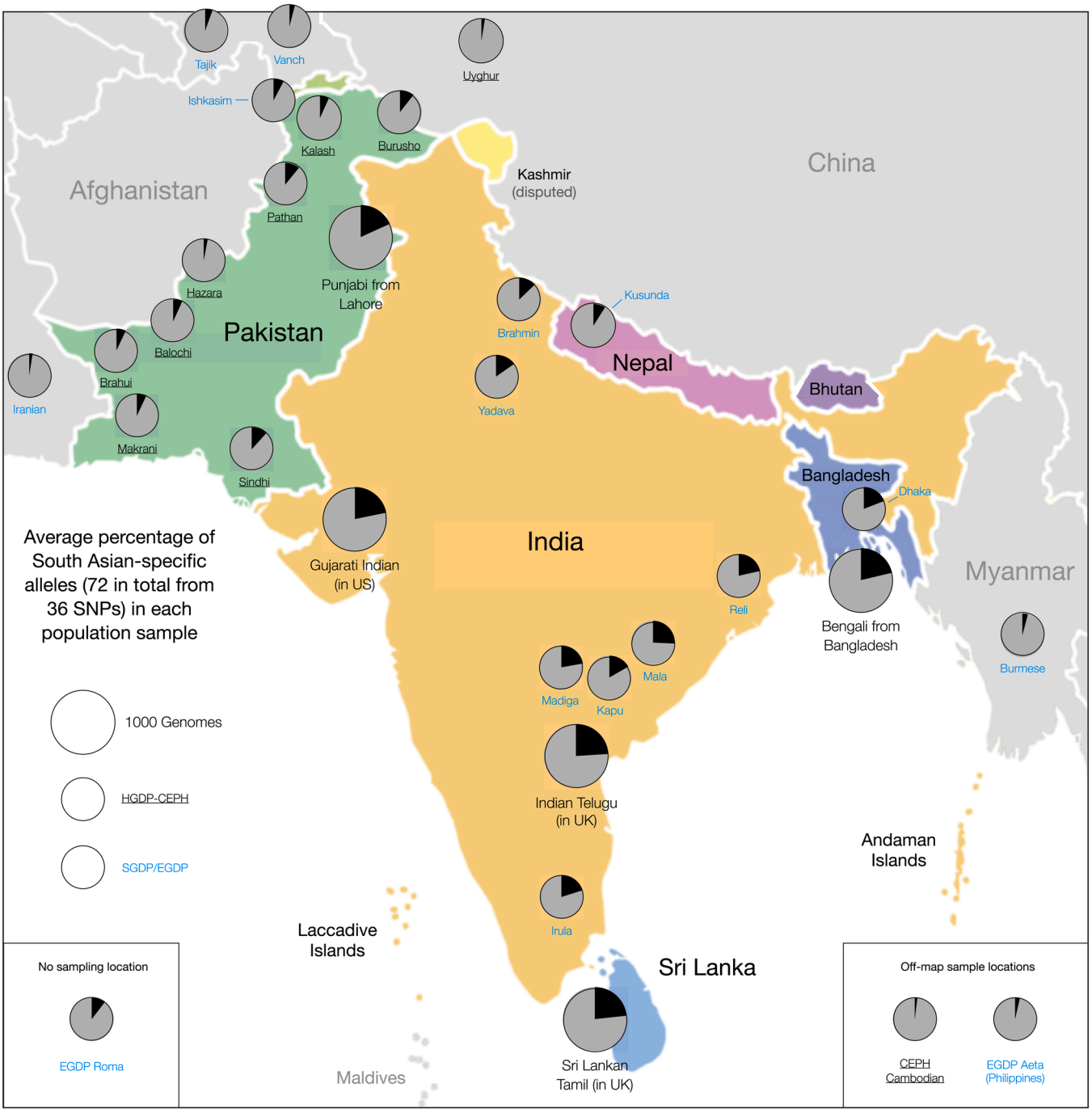


Fig. 5. The average percentage of South Asian-specific alleles (from a total of 72) in each population sample with a recorded value higher than 0.5 %. All such populations are located within or closely sited to the Indian sub-continent apart from the CEPH Cambodian and EGDP Aeta of the Philippines. Populations to the west, east and north of India tend to show much lower average percent specific alleles compared to the maximum value of 24 % seen in the 1000 Genomes STU and ITU population samples. These population samples based in the UK, and the GIH in the US, are positioned in their approximate geographic location. EGDP Roma samples have no stated sampling location.

trillion times more likely to be from South Asia than Africa'. Nevertheless, five individuals had LR values below '2000 times more likely', that all corresponded to samples with the lowest number (6) of South Asian-specific alleles. Third, two PCA plots are shown in Supplementary File S1.3 from analyses in Snipper of 1000 Genomes YRI, CEU, CHB, and CEPH Native American, Oceanian populations compared with CEPH Pakistanis (Plot 1), and compared with 1000 Genomes GIH (plot 2). These plots illustrate the very tightly distributed set of PCA points in populations outside of South Asia obtained with *Eurasiaplex-2* SNPs. The target population PCA point distributions are different between Pakistanis and GIH, with Pakistanis equally diffuse in distribution, but overlapping with the much smaller area of the 2D plot occupied by populations outside of South Asia. While PCA itself would not be a system of choice for assigning ancestry and this represents analysis with a single set of population-specific markers, it is interesting that zero or near-zero allele frequencies in most populations causes samples, even in large numbers, to occupy a very small area of the PCA plot.

Table 2
HapMap genetic map analysis of the six chromosome 16 South Asian-specific SNPs in *Eurasiaplex-2*. The map distance was estimated in Centimorgans (cM) and the recombination fraction (Rc) calculated from the cM values using Kosambi adjusted data. The 5′ SNPs rs368479296-rs376893831-rs370130302 show very little linkage with 32 % and 11 % recombination fractions, but the other three 3′ SNPs rs368738705-rs368538881-rs377323011 (in bold) have a much lower level of recombination of less than 0.3 %.

Internal ID	SNP	GRCh37 position	GRCh38 position	cM inter-SNP distance	Kosambi-adjusted Rc
16A	rs368479296	3178971	3128970		
16B	rs376893831	23053815	23042494	38.681	0.324515
16C	rs370130302	28588059	28576738	11.3898	0.111968
16D	**rs368738705**	33921593	34119126	1.793	0.017922
16E	**rs368538881**	46499858	46465946	0.0422	0.000422
16F	**rs377323011**	48327788	48293877	0.3118	0.003118

3.4.2. *Genetic cluster analysis with STRUCTURE comparing Eurasiaplex and Eurasiaplex-2 SNPs*

Supplementary File S1.4 shows the cluster plots from separate STRUCTURE analysis of 1000 Genomes populations using the original 23 *Eurasiaplex* SNPs and the 36 *Eurasiaplex-2* SNPs. Patterns show that the *Eurasiaplex-2* SNP set clearly distinguishes South Asians from all other 1000 Genomes populations at K:2, with a clean set of columns and some minor mixed cluster proportions in the PJL. As there is almost no genetic variation present in non-South Asian samples for *Eurasiaplex-2* SNPs, no other genetic clusters are identified at K:3, K:4, or higher K values (data not shown). It is noteworthy that the South Asian-specific rs77510889-G allele detected in African populations did not produce an African cluster for any higher K values analysed. In contrast, the 23 *Eurasiaplex* SNPs distinguish Europeans as the first major genetic cluster at K:2, then Africans at K:3, with South Asians only emerging as a differentiated population group at K:4. To some degree, these patterns are likely to reflect the selection of the original *Eurasiaplex* SNPs that had strongly contrasting allele frequencies between Europeans and other population groups, including small, but above-average allele frequency differences between Europeans and South Asians.

3.4.3. *Exploration of a simple South Asian-specific allele counting system*

Arguably, a much more straightforward system of population assignment than Bayes analysis can be achieved by simply counting the number of South Asian-specific *Eurasiaplex-2* alleles in an individual. Fig. 3 illustrating the worldwide distribution of genotype counts, and Fig. 5, those of allele counts greater than 0.5 %, show they are both almost completely confined to the Indian sub-continent and adjoining areas. These patterns show that highly contrasted allele counts in individuals from South Asia vs. individuals from other regions will give strong indications of origins from the regions targeted by *Eurasiaplex-2,* provided there is no overlap between minimum and maximum counts from each set of populations. Fig. 7 plots the distribution of South Asian-

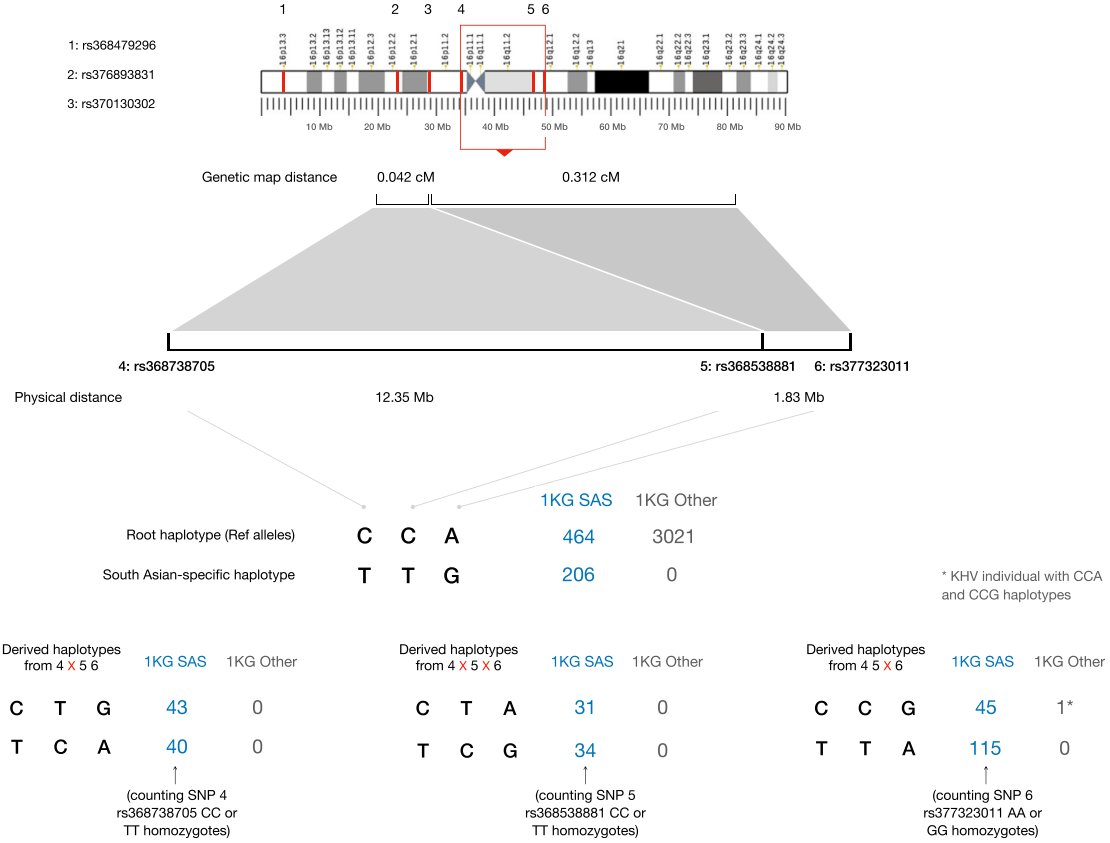


Fig. 6. The six SNPs of Chromosome 16 (red bars) and the haplotype landscape around SNPs 4–6 [rs368738705-rs368538881-rs377323011] closest to the centromere, where Centimorgan (cM) genetic map distances are particularly small (red box). As all three SNPs have alphabetic ordering for the Ref and Alt (South Asian-specific) alleles - i.e., CT, CT, AG, respectively, all 1000 Genomes heterozygous genotypes were alphabetised, and homozygote genotypes counted in order to record recombination between the root haplotype CCA and the South Asian-specific haplotype TTG. In this way rs368738705 CC or TT homozygote genotypes indicated recombination between rs368738705-rs368538881 (‘4 × 5 × 6′); rs377323011 AA or GG indicated recombination between rs368538881-rs377323011 (‘4 5 × 6′); and double sequential recombination events (‘4 × 5 × 6′) were recorded as rs368538881 AA or CC homozygotes. Counts are given for the 1000 Genomes South Asians (1KG SAS) vs. all other 1000 Genomes populations (1KG Other, excluding admixed individuals), indicating widespread disruption of CCA and TTG haplotypes and likely minor levels of association amongst the South Asian-specific alleles of these three SNPs. The single CCG haplotype recorded in a non-South Asian individual was inferred from CC-CC-AG genotypes. Mb: megabase; KHV: Kinh in Ho Chi Minh City, Vietnam.

specific allele counts in the four 1000 Genomes population groups plus CEPH Pakistanis. In the 1000 Genomes Africans, Europeans and East Asians, the majority of samples, some 80–90 %, have no specific alleles present in any genotypes. Africans have 16 % of genotypes with a single specific allele, mainly due to the rs77510889-G allele in these populations, but that represents the upper limit of specific allele counts in this population group. Only East Asians have a few individuals with more than one specific allele, with five samples having two alleles and a singleton with a maximum of three. The lower limit of specific alleles in South Asians is 5, with a singleton sample with this number, and then ten with 6 alleles, meaning there is no overlap in counts between the South Asians of 1000 Genomes and all individuals from the other population groups. Although a count of three, four or five specific alleles might be considered ambiguous, such a small number of individuals could stay unassigned. Therefore, there would be a very low probability of incorrect assignment of individuals using a lower limit of six specific alleles to signal South Asian origins.

The CEPH Pakistanis show a distribution with a greater degree of overlap with populations outside of South Asia, which is not unexpected, but strict adherence to a minimum six specific alleles means just over half of Pakistanis (88 of 168 with five or less South Asian-Specific alleles) are not assigned as South Asian. We intend to develop a statistical system for the handling of population-specific allele counts, based on hypothesis testing where the null hypothesis represents origins in the specific-allele target population. This will be more easily accomplished when a globally applicable set of population-specific SNPs has been compiled for all population groups.

4. Discussion

By constructing a SNP panel composed of a completely new type of ancestry marker focused on variation that is specific to the single targeted South Asia population group, rather than using SNPs with highly contrasting but shared variation across multiple populations, we have identified a characteristic signature of South Asian origin in almost all individuals from the Indian sub-continent. This specific-allele signature, illustrated by the dense pattern of orange and red bars in Fig. 3, is clearly observed across a large, broadly-based collection of South Asian samples taken from across the world. Apart from rs77510889, that showed a relatively high frequency of the South Asian-specific G-allele in Africans, all other SNP alleles chosen to be specific to South Asia had frequencies below 0.005 (0.5 %) in populations outside this region. Applying the same allele frequency intersect to other 1000 Genomes population groups plus Oceanian, American, and Middle East populations represented in the HGDP-CEPH panel, will allow the compilation of a global population-specific ancestry marker set, with marker numbers appropriate for MPS-scale SNP multiplexes. The process of building a large MPS multiplex requires some adjustments of targeted SNPs with poor context sequence or flanking region variation that interferes with sequence alignments or primer binding. For this reason, we chose to report a full list of suitable South Asian-specific candidate loci, rather than build a small-scale multiplex of 30–40 markers, as we have done for many forensic SNP panels previously [1,25,26]. The near-equal cumulative Divergence values between South Asians vs. East Asians and vs. Europeans shown by the 36 SNPs in *Eurasiaplex-2*, will make the process of balancing SNPs specific for each population group relatively straightforward, as the number of markers can be adjusted to produce a comparable average number of population-specific alleles per individual from that population.

The variant data and its statistical treatment that we have briefly explored in this study, requires a rethink of how best to adapt highly population-specific SNP allele patterns of variation into a forensic ancestry prediction framework. We do not feel that Bayes analysis or PCA will provide the necessary detailed assessments of the number of alleles specific to a given population that are detected in an individual. We expect STRUCTURE to be more sensitive to specific-allele patterns as well as being able to efficiently analyse co-ancestry in individuals with admixed backgrounds [6,12]. Until a global panel of population-specific SNPs is constructed this will need to be explored *in-silico* once enough candidate SNPs for each population group have been compiled. It is also useful to begin to develop a statistical framework that centres on formal testing of hypotheses for H1: originating from the target population vs. H2: originating from another.

An ancestry SNP selection process that enriches for markers with near-absolute specificity is more likely to detect variation present in the targeted population due to less common evolutionary genetic processes than those that underlie traditionally compiled ancestry sets (i.e., natural selection and genetic drift). These processes might include recent

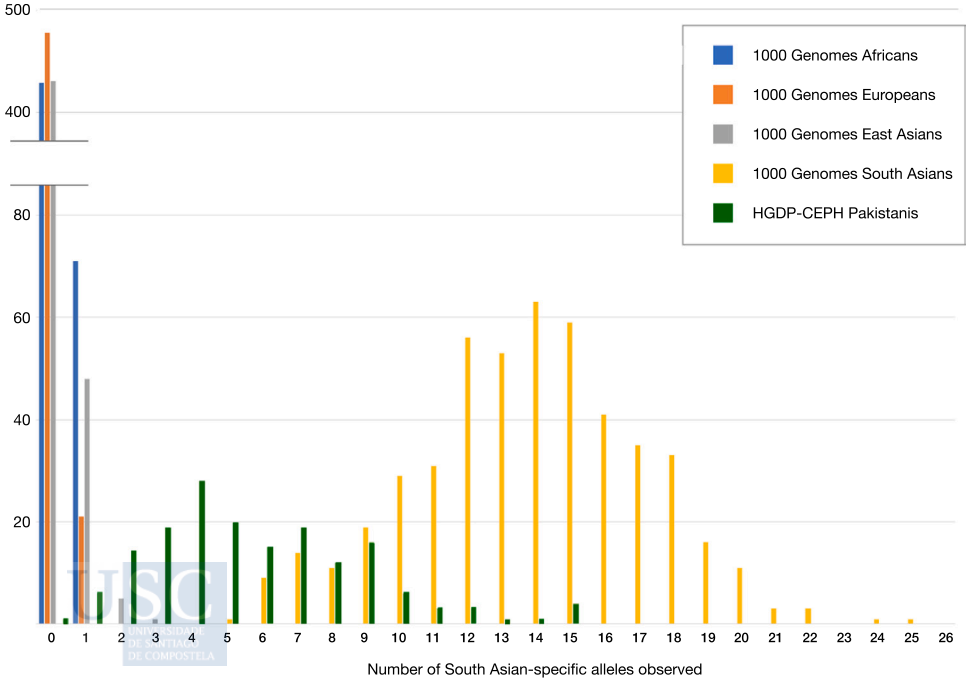


Fig. 7. Distribution of South Asian-specific allele counts in the four 1000 Genomes population groups, showing a bell-shaped distribution of specific alleles in the South Asian populations, and the lowest value represented by a single South Asian individual with five specific alleles. There is no overlap with the distribution of specific allele counts in the other 1000 Genomes populations with a single East Asian individual with a maximum three specific alleles. CEPH Pakistani samples show a degree of overlap with a single individual having no specific alleles, but more than half of Pakistani samples having less than a nominal six specific alleles lower limit to signify South Asian origins.

mutation events creating new SNP variants confined to a specific geographic region [27]; gene flow from Hominin introgression taking place in a particular locality [28,29]; localised selective sweeps, which might favour certain low frequency variants which then become region-specific [30]. Any of these processes could have occurred after the South Asian root populations of Ancestral North Indians and Ancestral South Indians [2,3] separated from other Eurasian groups of populations [31]. There is also the additional complexity of the caste system in Indian populations creating highly stratified distributions of variability across the sub-continent (e.g., the Brahmin caste has higher Iranian ancestry than other Indian castes and this differentiation would be maintained by reduced outbreeding across castes [32]). The silk roads provided a strong driver of East-West gene flow across the central parts of Eurasia, but these were largely routed to the north of the Himalayas which acted as a lengthy barrier to mass movements into the Indian sub-continent. Overall, the South Asian SNP variation we have identified and compiled represents only a very small proportion of total genomic variability, but the markers have maintained their high levels of specificity by consistently showing zero, or near zero allele frequencies in every region outside of the Indian sub-continent studied so far.

Acknowledgements

M.d.l.P. is supported by a post-doctorate grant funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481D-2021-008). J.R. is supported by the “Programa de axudas á etapa predoutoral” funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481A-2020-039).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2022.102780.

References

[1] C. Phillips, A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M.D. Perez Carceles, Á. Carracedo, P.M. Schneider, M.V. Lareu, *Eurasiaplex*: a forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int. Genet. 7 (2013) 359–366.
[2] D. Reich, K. Thangaraj, N. Patterson, A.L. Price, L. Singh, Reconstructing Indian population history, Nature 461 (2009) 489–494.
[3] P.P. Majumder, The human genetic history of South Asia, Curr. Biol. 20 (2010) R184–187.
[4] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Børsting, P. Johansen, M. Fondevila, et al., Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set, Forensic Sci. Int. Genet. 11 (2014) 13–25.
[5] J.R. Kidd, F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd, Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples, Invest. Genet. 2 (2011) 1.
[6] M. de la Puente, J. Ruiz-Ramírez, A. Ambroa-Conde, C. Xavier, J. Pardo-Seco, J. Álvarez-Dios, A. Freire-Aradas, A. Mosquera-Miguel, T.E. Gross, E.Y.Y. Cheung, et al., Development and evaluation of the ancestry informative marker panel of the VISAGE basic tool, Genes 12 (2021) 1284.
[7] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet. 73 (2003) 1402–1422.
[8] P. Pfaffelhuber, F. Grundner-Culemann, V. Lipphardt, F. Baumdicker, How to choose sets of ancestry informative markers: a supervised feature selection approach, Forensic Sci. Int. Genet. 46 (2020), 102259.
[9] The 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E. P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, et al., A global reference for human genetic variation, Nature 526 (2015) 68–74.
[10] M. Byrska-Bishop, U.S. Evani, X. Zhao, A.O. Basile, H.J. Abel, A.A. Regier, A. André Corvelo, W.E. Clarke, R. Musunuri, K. Nagulapalli, et al., High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios, bioRxiv preprint, posted February 7 2021 doi: ⟨https://doi.org/10.1101/2021.02.06.430068⟩.
[11] S. Zhao, C.-M. Shi, L. Ma, Q. Liu, Y. Liu, F. Wu, L. Chi, H. Chen, AIM-SNPtag: a computationally efficient approach for developing ancestry-informative SNP panels, Forensic Sci. Int. Genet. 38 (2019) 245–253.
[12] J. Ruiz-Ramírez, M. de la Puente, C. Xavier, A. Ambroa-Conde, J. Álvarez-Dios, A. Freire-Aradas, A. Mosquera-Miguel, A. Ralf, C. Amory, M.A. Katsara, et al., Development and evaluations of the ancestry informative markers of the VISAGE enhanced tool for appearance and ancestry, Forensic Sc. Int. Genet. (2022).
[13] ⟨http://www.ensembl.org/Homo_sapiens/Variation/Explore?r=6:60527829–60528829;v=rs3857620;vdb=variation;vf=169483878⟩ (Accessed June 2022).
[14] M. Lek, K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J.S. Ware, A.J. Hill, B.B. Cummings, et al., Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (2016) 285–291.
[15] A. Bergström, S.A. McCarthy, R. Hui, M.A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, J, et al., Insights into human genetic variation and population history from 929 diverse genomes, Science 367 (2020) 1339–1349.
[16] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, et al., The Simons Genome Diversity Project: 300 genomes from 142 diverse populations, Nature 538 (2016) 201–206.
[17] L. Pagani, D.J. Lawson, E. Jagoda, A. Mörseburg, A. Eriksson, M. Mitt, F. Clemente, G. Hudjashov, M. DeGiorgio, L. Saag, et al., Genomic analyses inform on migration events during the peopling of Eurasia, Nature 538 (2016) 238–242.
[18] L.-P. Wong, R.T. Ong, W.T. Poh, X. Liu, P. Chen, R. Li, K. Koi-Yau Lam, N. Esakimuthu Pillai, K.-S. Sim, H. Xu, et al., Deep whole-genome sequencing of 100 southeast Asian Malays, Am. J. Hum. Genet. 92 (2013) 52–66.
[19] L.-P. Wong, J. Kuan-Han Lai, W.-Y. Saw, R.T. Ong, A. Youzhi Cheng, N. Esakimuthu Pillai, X. Liu, W. Xu, P. Chen, J.-N. Foo, et al., Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing, PLoS Genet. 10 (2014), e1004377.
[20] ⟨http://mathgene.usc.es/snipper/analysismultipleprofiles.html⟩.
[21] C. Phillips, D. Ballard, P. Gill, D. Syndercombe Court, A. Carracedo, M.V. Lareu, The recombination landscape around forensic STRs: accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data, Forensic Sci. Int. Genet. 6 (2012) 354–365.
[22] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo, M.V. Lareu, An overview of STRUCTURE: applications, parameter settings, and supporting software, Front. Genet. 4 (2013) 98.
[23] J. Amigo, A. Salas, C. Phillips, ENGINES: exploring single nucleotide variation in entire human genomes, BMC Bioinf. 12 (2011) 105.
[24] A. Gómez-Carballa, J. Pardo-Seco, L. Fachal, A. Vega, M. Cebey, N. Martinón-Torres, F. Martinón-Torres, A. Salas, Indian signatures in the westernmost edge of the European Romani diaspora: New insight from mitogenomes, PLoS One 8 (2013), e75397.
[25] M. de la Puente, C. Santos, M. Fondevila, L. Manzo, EUROFORGEN-NoE Consortium, A. Carracedo, M.V. Lareu, C. Phillips, The Global AIMs Nano set: a 31-plex SNaPshot assay of ancestry-informative SNPs, Forensic Sci. Int. Genet. 22 (2016) 81–88.
[26] C. Phillips, L. Manzo, M. de la Puente, M. Fondevila, M.V. Lareu, The MASTiFF panel - a versatile multiple-allele SNP test for forensics, Int. J. Leg. Med. 134 (2020) 441–450.
[27] L.M. Williams, M.F. Oleksiak, Ecologically and evolutionarily important SNPs identified in natural populations, Mol. Biol. Evol. 28 (2011) 1817–1826.
[28] S. Sankararaman, N. Patterson, H. Li, S. Pääbo, D. Reich, The date of interbreeding between Neandertals and modern humans, PLoS Genet. 8 (2012), e1002947.
[29] E. Huerta-Sánchez, F.P. Casey, Archaic inheritance: supporting high-altitude life in Tibet, J. Appl. Physiol. 119 (1985) 1129–1134.
[30] H. Chen, N. Patterson, D. Reich, Population differentiation as a test for selective sweeps, Genome Res. 20 (2010) 393–402.
[31] V.M. Narasimhan, N. Patterson, P. Moorjani, N. Rohland, R. Bernardos, S. Mallick, I. Lazaridis, N. Nakatsuka, I. Olalde, M. Lipson, et al., The formation of human populations in South and Central Asia, Science 365 (2019) eaat7487.
[32] G. Debortoli, C. Abbatangelo, F. Ceballos, C. Fortes-Lima, H.L. Norton, S. Ozarkar, E.J. Parra, M. Jonnalagadda, Novel insights on demographic history of tribal and caste groups from West Maharashtra (India) using genome-wide data, Sci. Rep. 10 (2020) 10075.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Article 5. A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel

Reference: Forensic Science International. Genetics, 46, 102232

ISSN: 1878-0326

DOI: 10.1016/j.fsigen.2020.102232

Web link: https://www.sciencedirect.com/science/article/pii/S187249732030003X

Authors: Phillips C[1], Amigo J[2], Tillmar AO[3,4], Peck MA[5], de la Puente M[1], Ruiz-Ramírez[1], F. Bittner[5], Idrizbegović Š[5], Wang Y[6], Parsons TJ[5], Lareu MV[1]

Author information:
1. Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain.
2. Fundación Pública Galega de Medicina Xenómica SERGAS, Grupo de Medicina Xenómica USC, IDIS, Santiago de Compostela, Spain.
3. Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden.
4. Department of Clinical and Experimental Medicine, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden.
5. International Commission on Missing Persons, Koninginnegracht 12a, The Hague, Netherlands.
6. Qiagen, 6951 Executive Way, Frederick, MD 21703, US.

Author information:
1. Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
2. Institute of Anthropology and Ethnology, Adam Mickiewicz University in Poznań, Poland

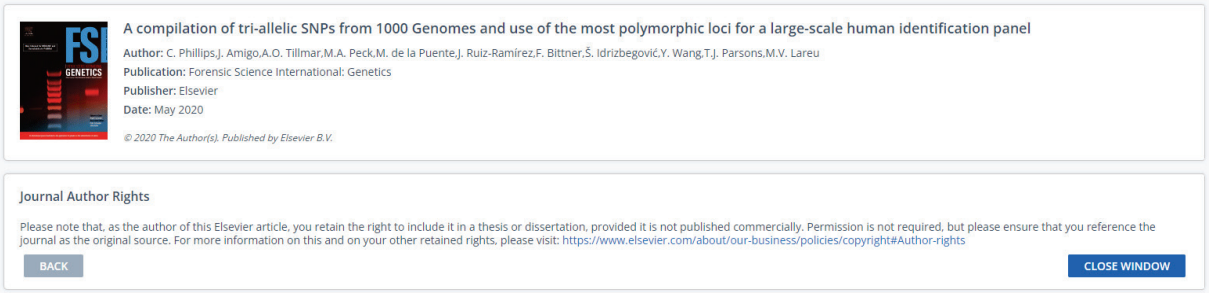
USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Quality metrics (JCR, 2020):

- Journal Impact factor: 4.882
- Quartile: Q1 (legal medicine) and Q1 (genetics & heredity)
- Rank by JCR: 1/17 (legal medicine) and 43/176 (genetics & heredity)

Contribution: Research

Authorization:



A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel

Author: C. Phillips,J. Amigo,A.O. Tillmar,M.A. Peck,M. de la Puente,J. Ruiz-Ramírez,F. Bittner,Š. Idrizbegović,Y. Wang,T.J. Parsons,M.V. Lareu

Publication: Forensic Science International: Genetics

Publisher: Elsevier

Date: May 2020

© 2020 The Author(s). Published by Elsevier B.V.

Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

BACK CLOSE WINDOW

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Forensic Science International: Genetics 46 (2020) 102232

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel

C. Phillips[a,*], J. Amigo[b], A.O. Tillmar[c,d], M.A. Peck[e], M. de la Puente[a], J. Ruiz-Ramírez[a], F. Bittner[e], Š. Idrizbegović[e], Y. Wang[f], T.J. Parsons[e], M.V. Lareu[a]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain
[b] Fundación Pública Galega de Medicina Xenómica SERGAS, Grupo de Medicina Xenómica USC, IDIS, Santiago de Compostela, Spain
[c] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
[d] Department of Clinical and Experimental Medicine, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden
[e] International Commission on Missing Persons, Koninginnegracht 12a, The Hague, Netherlands
[f] Qiagen, 6951 Executive Way, Frederick, MD 21703, US

ARTICLE INFO

Keywords:
Tri-allelic SNPs
1000 Genomes
Missing persons identification
Massively parallel sequencing

ABSTRACT

In a directed search of 1000 Genomes Phase III variation data, 271,934 tri-allelic single nucleotide polymorphisms (SNPs) were identified amongst the genotypes of 2,504 individuals from 26 populations. The majority of tri-allelic SNPs have three nucleotide substitution-based alleles at the same position, while a much smaller proportion, which we did not compile, have a nucleotide insertion/deletion plus substitution alleles. SNPs with three alleles have higher discrimination power than binary loci but keep the same characteristic of optimum amplification of the fragmented DNA found in highly degraded forensic samples. Although most of the tri-allelic SNPs identified had one or two alleles at low frequencies, often single observations, we present a full compilation of the genome positions, rs-numbers and genotypes of all tri-allelic SNPs detected by the 1000 Genomes project from the more detailed analyses it applied to Phase III sequence data. A total of 8,705 tri-allelic SNPs had overall heterozygosities (averaged across all 1000 Genomes populations) higher than the binary SNP maximum value of 0.5. Of these, 1,637 displayed the highest average heterozygosity values of 0.6-0.666. The most informative tri-allelic SNPs we identified were used to construct a large-scale human identification panel for massively parallel sequencing, designed for the identification of missing persons. The large-scale MPS identification panel comprised: 1,241 autosomal tri-allelic SNPs and 29 X tri-allelic SNPs (plus 46 microhaplotypes adapted for genotyping from reduced length sequences). Allele frequency estimates are detailed for African, European, South Asian and East Asian population groups plus the Peruvian population sampled by 1000 Genomes for the 1,270 tri-allelic SNPs of the final MPS panel. We describe the selection criteria, kinship simulation experiments and genomic analyses used to select the tri-allelic SNP components of the panel. Approximately 5 % of the tri-allelic SNPs selected for the large-scale MPS identification panel gave three-genotype patterns in single individual samples or discordant genotypes for genomic control DNAs. A likely explanation for some of these unreliably genotyped loci is that they map to multiple sites in the genome - highlighting the need for caution and detailed scrutiny of multiple-allele variant data when designing future forensic SNP panels, as such patterns can arise from common structural variation in the genome, such as segmental duplications.

1. Introduction

The primary aim of the 1000 Genomes Project was to construct a comprehensive catalog of human variation using fully evolved high-throughput sequencing technologies [1]. Single nucleotide polymorphisms (SNPs) are by far the most common variant type in the human genome, amounting to more than 96 % of polymorphic variation. Amongst the 88 million variants recorded by the project's final data release, SNPs comprised 84.7 million variants with minor allele frequencies > 1 % [2]. The 1000 Genomes project reported the compiled variant catalogs in two stages named Phase I [3] and Phase III [2]; producing an initial genome variation map in Phase I based on 1,092

* Corresponding author.
E-mail address: c.phillips@mac.com (C. Phillips).

https://doi.org/10.1016/j.fsigen.2020.102232
Received 3 December 2019; Received in revised form 29 December 2019; Accepted 2 January 2020
Available online 17 January 2020
1872-4973/ © 2020 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

individuals (from 14 worldwide populations), which then progressed to the more extensive Phase III variant map from the sequencing of 2,504 individuals (26 populations). We originally developed a simple search interface for the Phase I data called ENGINES [4] accepting multiple SNP queries (using rs-number lists), or searches for all the Phase I SNPs identified within a user-defined chromosome segment or gene. When ENGINES is queried with SNPs known to have three nucleotide alleles, commonly termed tri-allelic SNPs [5–12], such as the well-established forensic markers rs4540055 and rs5030240 [5–9], the Phase I allelic data returned is binary with just two alleles shown. We interpreted this variant data to indicate that the sequence analysis used to identify the 1000 genomes Phase I variants took a cautious approach when finding atypical patterns of nucleotide variation and reserved description of such loci for more detailed re-sequencing. Consequently, when the Phase III catalog was released, tri-allelic SNPs had been re-instated at all previously identified sites, such as rs4540055. These SNP's polymorphisms had now been fully characterized and matched their known patterns of variation in the main population groups. Multiple allele SNPs (i.e. including tetra-allelic SNPs [11], as well as tri-allelic SNPs described here) now made up a significant proportion of the total variation observed - with approximately 0.32 % of all listed SNPs showing three alleles in Phase III. Although this is a higher proportion than might be expected given two separate substitutions must become fixed in a population to establish such polymorphisms, there is growing evidence that the substitutions in tri-allelic SNPs are not independent events and the mutations that create these SNPs occur at a higher frequency than predicted from the distribution of binary SNPs in the human genome [13,14].

The six genotypes possible in tri-allelic SNPs clearly raises the overall level of polymorphism per marker compared to binary SNPs, making tri-allelic variants compelling markers for forensic use, since a multiplex of such SNPs achieves higher discrimination power, but can still be based on short amplified fragment sizes not possible with most mainstream STRs. This study compiled all the detected tri-allelic SNPs in the 1000 Genomes Phase III catalog and identified the most informative of them. A massively parallel sequencing (MPS) test was designed which combined the most polymorphic tri-allelic SNPs to make a single large-scale MPS multiplex for the identification of missing persons (herein, the MPS Panel). As so many highly polymorphic tri-allelic SNPs were listed by 1000 Genomes, it was viable to carefully adjust the chromosomal distribution of the loci to minimize linkage in closely clustering markers – that could potentially interfere with probability calculations assuming marker independence in related individuals [15]. Although an extensive list of candidate identification markers allows the selection of a p-arm and q-arm location per chromosome, as previously made with forensic ID-SNP panels [16], we sought to develop a multiplex for MPS on a much larger scale than ~50 component loci. Simulations based on 1000 Genomes allele frequency estimates were used to gauge the informativeness of the final MPS multiplex design when applied to the most difficult kinship testing scenarios of pairwise comparisons of distant relationships in a deficient pedigree. This paper reports the genomic data of all tri-allelic SNPs now listed by 1000 Genomes; the population variation in the most polymorphic of these SNPs selected for the MPS Panel; details of apparent tri-allelic SNPs that were removed from the final panel due to atypical patterns of allelic variation; and the performance of the SNP set's genotype data in simulated kinship testing scenarios typically encountered in the identification of missing persons. The development, optimization and operational application of this large-scale MPS multiplex for the identification of missing persons, as well as the adaptation of a sub-set of microhaplotypes and addition of a small set of binary SNPs in the same panel, are the subject of a separate publication.

2. Materials and methods

2.1. Directed searches of the 1000 Genomes Phase III variant database to discover tri-allelic SNPs

1000 Genomes Phase III variant data were downloaded from the project's FTP site [17] and SNP sites tagged as "MULTI_ALLELIC" were extracted with a combination of *bcftools*, *sed* and *grep* scripting. The 1000 Genomes VCF data for multiple allele SNPs comprises the reference allele (i.e. RefSeq) with label "0", and alternative alleles listed alphabetically are assigned "1" and "2"; plus "3" when all four nucleotide substitution alleles are observed. The VCF header line for each variant combines the populations of five continental regions into groups with averaged allele frequency estimates (AMR: admixed Americans; AFR: Africans; EUR: Europeans; SAS: South Asians; EAS: East Asians). This frequency data was used to gauge the level of polymorphism in each population group by simple heterozygosity value estimation. Because of high levels of admixture in American continent population samples, only the Peruvians from Lima, Peru (PEL) data was used to assess Native American variation in the SNPs selected for the MPS Panel (see the predominantly red AME population structure of PEL in Fig. 2a of [2]). Note that since the 1000 Genomes project closed, ongoing human variant data curation has been made by IGSR, the International Genomes Sample Resource [18] and this organisation also compiles data from TOPmed (Trans-Omics for Precision Medicine, 9,000 genomes [19]) and gnomAD (Genome Aggregation Database, 4,368 genomes [20]). Although TOPmed and gnomAD only report allele frequencies, they benefit from larger sample sizes than 1000 Genomes and therefore greater sensitivity to detect very low frequency "2 and 3 alternative alleles". In this study, all tetra-allelic sites detected were listed alongside tri-allelic SNPs but excluded from analyses, as they had been the subject of a separate detailed compilation to identify informative forensic markers with four alleles [11].

There is not always a direct relationship between the binary SNP allele data in Phase I and the characteristics listed for the same loci in Phase III when they are identified to be multi-allelic. Table 1 lists genomic details of the six tri-allelic SNPs adopted in an MPS forensic

Table 1
Genomic details of six well established tri-allelic SNPs and the contrasting allele frequency estimates from 1000 Genomes Phase I and Phase III (for population CEU: CEPH Europeans from Utah). Allele frequency estimates from gnomAD (Genome Aggregation Database) are for non-Finnish Europeans (NFE). Chr. chromosome; Ref. reference (i.e. RefSeq); Alt. alternative.

	Locus details			1000 Genomes Phase I					1000 Genomes Phase III							gnomAD		
SNP ID	Chr.	GRCh37 position	Gene	Ref.	Alt.	Ancestral	Ref.	Alt-1	Ref.	Alt-1	Alt-2	Ancestral	Ref.	Alt-1	Alt-2	Ref.	Alt-1	Alt-2
rs17287498	10	54530788	MBL2	C	A	A	0.77	0.23	C	A	T	A	(no data listed in 1000 genomes)			0.5661	0.2055	0.2284
rs2069945	20	33761837	PROCR	C	G	C	0.517	0.483	C	A	G	C	0.48	0.101	0.419	0.4348	0.1006	0.4646
rs2184030	1	206667441	–	G	A	C	0.46	0.54	G	A	C	C	0.465	0.515	0.02	0.52611	0.4596	0.0143
rs433342	8	17747876	FGL1	A	G	A	0.259	0.741	A	C	G	A	0.232	0.04	0.727	0.259	0.028	0.713
rs4540055	4	38803255	TLR1	A	C	A	0.989	0.011	A	C	T	T	0.793	0.01	0.197	0.7761	0.0256	0.1983
rs5030240	11	32424389	WT1	C	A	–	0.948	0.052	C	A	G	A	0.712	0.056	0.232	0.6987	0.0927	0.2086

ancestry analysis panel [9], which illustrates some of the Phase I vs Phase III disparities that can arise when comparing both variant catalogs. First, the alternative allele identified in Phase I is not necessarily very frequent; alleles rs4540055-C and rs5030240-A are less frequent than the rs4540055-T and rs5030240-G alt-2 alleles discovered later in Phase III. Second, alphabetic sorting of 2 or 3 alternative alleles often leads to re-assignment of the VCF allele numbers, as occurs for rs2069945-G and rs433342-G. Lastly, the identification of the ancestral nucleotide can also be affected by the discovery of multiple alleles at a SNP site, as shown by the shift in rs4540055 from an A to a T ancestral allele.

Table 1 also shows that rs17287498 has no Phase III variant data, with just gnomAD allele frequencies listed by 1000 Genomes. Therefore, this study's reliance on Phase III data sources led to an extensive but slightly incomplete catalog of human multi-allelic SNPs. We also queried the gnomAD v2.1.1 genome and v2.1.1 exome databases using in-house scripts (available on request), although no formal comparisons were made in this study between both databases to assess the degree of allele frequency concordance of the multi-allelic SNPs identified in 1000 Genomes. Finally, for the subset of SNPs selected for the MPS Panel, population variation beyond the main 1000 Genomes continental population groups listed above, was compiled from Simons Foundation Genome Diversity Project (SGDP) sequence analysis of 263 individual genomes [21], from which we took 44 Native American and 21 Oceanian samples. Although limited in size, these SGDP samples enabled an initial assessment of any collective population divergence present in the component SNPs of the MPS Panel.

2.2. Selection of tri-allelic SNPs for the MPS Panel and rules used to minimize linkage between SNPs

All human tri-allelic SNPs identified were ranked by overall heterozygosity (i.e. based on the average 1000 Genomes allele frequencies of 2,504 samples given in the VCF header) in lists per chromosome. All loci with overall heterozygosity values greater than 0.5 were brought together into a single candidate set of 3,426 autosomal SNPs and 264 X-Chromosome SNPs. Candidate SNPs tended to have similar average heterozygosities between different population groups, but SNPs with strong contrasts in allele frequency distributions between population groups were discarded despite often having high average heterozygosity levels. For example, Fig. 1 shows the allele distributions in rs1695865, which has an average heterozygosity of 0.5124, but is a SNP better suited to ancestry inference; with strong contrasts in allele frequency between Europeans vs Africans and East Asians.

Loci were selected from positions on each chromosome that occupied a 1–5 megabase (Mb) segment and were a minimum 1 centimorgan (cM) map distance to the next SNP site, running from the 5′-most position to the following position in the 3′ direction. All cM map distances were estimated using the HapMap recombination map, applying the system outlined by Phillips et al. [22]. Therefore, in regions with low recombination rates such as around the centromere, physical distance between selected tri-allelic SNPs was often much longer than average, although the cM value between sites did not fall below one (with only 7 exceptions of less than 1 cM spacing for a set of SNPs on the X chromosome).

2.3. Sequence quality checks of MPS Panel candidate SNPs

Brief preliminary sequence quality screening was applied to candidate tri-allelic SNPs for the MPS Panel, which consisted of: (i) a brief visual inspection of the sequence with the Ensembl and Santa Cruz genome browsers [23,24] to discard candidates showing sequence characteristics that could hinder alignment, including long poly-tracts, repetitive regions, Indels and structural variants close to the target SNP site; and (ii) sequences aligning to multiple genomic positions in nucleotide BLAST [25] to avoid duplicated regions. Although a small proportion of SNPs were removed and replaced with slightly less informative but closely sited alternatives having better flanking sequence, the bulk of the candidate SNPs were submitted for MPS assay design by QIAGEN, who applied their QIAseq primer design pipelines to assemble the final large-scale PCR multiplex.

2.4. Kinship testing simulations

Simulations were run to assess the statistical power of tri-allelic SNP sets for analyzing complex kinship testing scenarios. Pairwise relationship tests were made, which compared two individuals in the absence of a pedigree (i.e. with no other family members for reference) and included full siblings; half siblings; first cousins; and second cousins. The simulation model implemented in FamLink [26] was applied to obtain distributions of likelihood ratios (LRs) for each kinship hypothesis (H1) vs the values for the unrelated hypothesis (H2), per panel. Fixed allele frequencies were used to generate artificial SNP sets and European allele frequency estimates from 1000 Genomes were used to simulate genotype data for the set of 1,377 autosomal SNPs selected to go into the final MPS identification panel. These allele frequencies dictated the simulated genotypes in 5,000 cases to generate an LR each time, calculated as: $LR = Pr(DNA|H1)/Pr(DNA|H2)$, where Pr is the probability of the genotype combinations given H1 (for each of the four relationships explored), compared to the probability given H2. For the analysis of the kinship informativeness of the MPS Panel, the cM estimates, as described in Section 2.2. were applied using the ILIR linkage adjustments in FamLink, as previously outlined [15,26]. LR distributions were estimated using the density plot function in the ggplot2 library [27] in R (www.r-project.org). Exceedance probabilities (i.e. the probability to obtain an LR larger than a given threshold) were

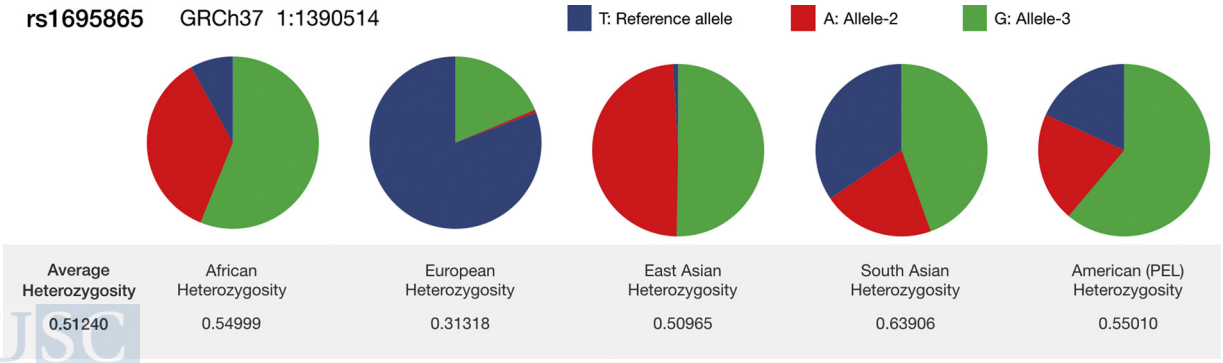


Fig. 1. The tri-allelic SNP rs1695865 has an overall heterozygosity of 0.5124, which is higher than the binary SNP maximum value of 0.5. Although listed as a potential candidate for the MPS Panel, it has a highly imbalanced allele frequency distribution and a much lower than average European heterozygosity, so was not considered further. However, the rs1695865 allele frequencies in five 1000 Genomes population groups reveal it would be an informative forensic ancestry marker.

estimated using the method described by Kruijver [28].

2.4.1. Simulations with artificial SNP sets to assess the impact of different allele frequency distributions

As a large proportion of the tri-allelic SNPs we compiled had a low frequency allele-3, we decided to explore the informativeness of a set of such SNPs compared to "perfect" tri-allelic SNPs that consist of three equally frequent alleles. While a balanced allele distribution appears to be the most informative type of polymorphism for kinship analyses, when a very large panel can be used, potentially as much power could be gained from having a small number of SNPs disproportionately contributing to the relationship likelihood when the rarest allele is the one shared between putative related individuals. To test this hypothesis, we constructed a 300-SNP artificial dataset with perfect distributions of 0.33-0.33-0.34 and another with rare allele-3 distributions of 0.5-0.45-0.05 and ran simplified simulations for half sibling and parent-child pair tests.

2.4.2. Simulations with allele frequency data of tri-allelic SNPs selected for the MPS Panel

For the kinship testing simulations of the MPS Panel tri-allelic SNPs, European allele frequency estimates for the original candidate pool of autosomal loci were used, i.e. the 1,377 'cM parsed' autosomal SNPs listed in Table 2, column 9; not the 1,241 autosomal loci that gave problem-free genotypes during MPS Panel development, and excluding the 34 X SNPs selected as candidates. Evaluations of the power of such a panel to enhance kinship tests were made by a direct comparison with an artificial dataset consisting of the same number of perfect binary SNPs (0.5:0.5 allele frequencies).

2.5. Evaluation of the ancestry inference capabilities of MPS Panel SNPs

Although the tri-allelic SNPs incorporated into the MPS Panel were selected to have low levels of allelic divergence between the 1000 genomes population groups, larger than average-scale panels of SNPs potentially offer population differentiation power equivalent to much smaller panels of carefully chosen ancestry informative SNPs (AIMs). We assessed the ability of the autosomal SNPs of the MPS Panel to assign ancestry to six continentally-defined population groups using 1000 Genomes allele frequency data for: AFR; EUR; SAS; EAS; plus a combination of Native American samples from SGDP and twenty 1000 Genomes Peruvians (PEL Peruvians from Lima) that lacked non-American co-ancestry (see Section 2.1); plus a total of 21 Oceanian samples from SGDP.

Population analyses to identify genetic clusters used STRUCTURE v.2.3.4 [29] consisting of three iterations of 100,000 burnin steps and 100,000 MCMC steps, correlated allele frequencies under the Admixture model at K = 6. Cluster membership proportion plots were constructed with CLUMPAK v.1.1 [30]. MDS analysis and construction of Neighbor-joining trees was implemented using R software v.3.5.0 [31] over an allele-distance matrix computed using the R package *pegas* [32].

3. Results and discussion

3.1. Genome-wide compilation of tri-allelic SNPs from 1000 Genomes data

In the Phase III variant data of 1000 Genomes, 271,934 SNPs have three alleles at varied frequencies (down to the lowest value of 0.0002 for single observations). Table 2 outlines the numbers found on each

Table 2
Left hand columns: total multiple-allele, tri-allelic and tetra-allelic SNPs collected per chromosome from 1000 Genomes Phase III variant data; numbers of tri-allelic SNPs with overall (i.e. 1000 Genomes-wide) Heterozygosity above the binary maximum value of 0.5, between 0.5–0.6, and greater than 0.6. Right hand columns: large scale MPS multiplex candidate SNPs with average Heterozygosity (Av. Het. combined average African-European-East Asian values) above 0.5 taken from most informative SNPs (col. 5); reduced by parsing to ensure evenly distributed cM spacing; 141 SNPs in columns 10–13 discarded from the multiplex due to low coverage, high levels of non-specific reads, discordant genotype calls or other MPS issues such as alignment problems; to produce the final optimum, balanced MPS multiplex of 1,270 tri-allelic SNPs. Chr. Chromosome; Het. Heterozygosity; cM centimorgan; QC quality control.

1000 Genomes Phase III variant compilations							Large-scale MPS multiplex construction						
Chr.	Total multiple-allele SNPs	Total tri-allelic SNPs	Total tetra-allelic SNPs	Tri-allelic SNPs with overall Het. > 0.5	Tri-allelic SNPs with overall Het. 0.5–0.6	Tri-allelic SNPs with overall Het. > 0.6	Candidate pool with 3-group Av. Het. > 0.5	Parsed for balanced cM spacing	Very low sequence coverage	Three alleles per individual	Discordant control DNA genotypes	Other MPS QC issues	Final MPS Panel
1	18,883	18,756	127	631	502	129	263	113	3	3	2	3	100
2	21,412	21,261	151	651	537	114	210	104	2	1	1	3	97
3	17,849	17,758	91	588	477	111	233	91	0	1	2	0	88
4	18,273	18,180	93	610	513	97	239	96	1	3	0	6	86
5	16,756	16,686	70	560	463	97	209	88	2	1	3	0	82
6	15,942	15,847	95	598	484	114	328	78	1	3	1	1	72
7	14,719	14,635	84	481	372	109	217	79	1	3	1	3	71
8	16,369	16,284	85	558	435	123	218	81	2	2	2	4	71
9	11,893	11,821	72	404	324	80	150	67	1	4	3	2	57
10	12,770	12,705	65	462	373	89	188	80	0	2	2	2	74
11	13,397	13,319	78	406	330	76	163	58	0	0	3	1	54
12	11,541	11,469	72	353	292	61	122	61	0	2	1	5	53
13	8,665	8,627	38	280	233	47	105	44	2	1	0	1	40
14	8,479	8,421	58	237	195	42	77	40	0	1	0	1	38
15	7,723	7,673	50	248	194	54	114	56	1	1	2	8	44
16	10,552	10,482	70	324	271	53	113	48	1	2	3	1	41
17	7,197	7,145	52	227	181	46	73	35	4	1	0	0	30
18	7,000	6,944	56	240	188	52	98	43	1	2	0	4	36
19	6,258	6,216	42	214	169	45	94	41	1	0	1	1	38
20	5,585	5,545	40	159	131	28	47	29	0	1	0	1	27
21	3,892	3,869	23	137	109	28	67	21	0	1	0	1	19
22	4,062	4,024	38	132	111	21	47	24	0	2	0	1	21
X	14,349	14,267	82	205	184	21	43	34	0	1	2	2	29
Sum:	**273,566**	**271,934**	**1,632**	**8,705**	**7,068**	**1,637**	**3,415**	**1,411**	**23**	**38**	**29**	**51**	**1,270**

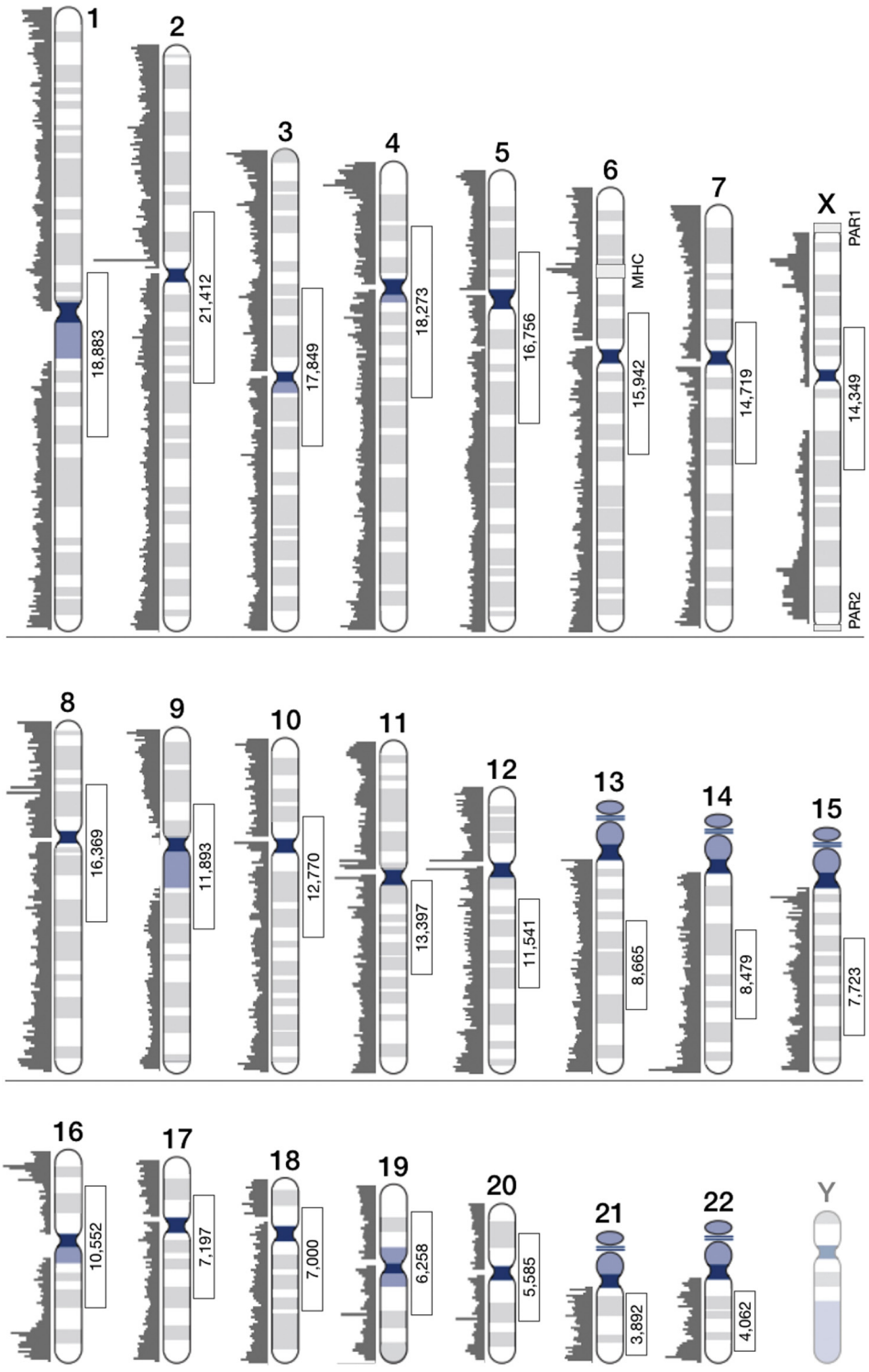


Fig. 2. Density plots of multiple-allele SNPs identified in the human genome. Each dark gray bar on the left represents total SNP counts per megabase of chromosome sequence. Y SNP data was not collected. Dark blue regions are facultative heterochromatin; light blue constitutive heterochromatin. Regions where SNPs were compiled but not selected for the MPS panel are PAR1/2: the pseudo-autosomal regions on X, and MHC: the major histocompatibility complex on C6. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

chromosome, except the Y chromosome. Due to the scale of the variant data compiled, we have processed the full VCF file into 23 detailed Excel tables listing all multiple-allele SNPs discovered, i.e. including the previously compiled 1,632 tetra-allelic SNPs [11]; provided as individual files per chromosome. Individual Excel files per chromosome are available at ***Mendeley Data*** (Mendeley Data V1, "A compilation of human tri-allelic SNPs from 1000 Genomes", DOI: 10.17632/46srpvw9xb.1). Each Excel file provides the SNP's GRCh37 genome build location; rs-number; the 2,504 genotypes of 1000 genomes samples; and overall heterozygosity values estimated from all genotypes. Each Excel file can be processed easily to identify tri-allelic SNPs that would be suitable for either forensic identification or forensic ancestry inference purposes. It should be noted that tri-allelic SNPs are not necessarily more informative for ancestry than the best binary SNP AIMs

or microhaplotype loci [33]; although dedicated searches for tri-allelic SNPs applicable to forensic ancestry analysis were not part of this study.

Fig. 2 displays the distribution of the compiled SNPs on each chromosome in the form of density bar charts representing total SNP counts per Mb. Some distributions match known patterns of human SNP variability; e.g., the broad peak of hypervariability observed for all SNPs around the major histocompatibility complex (MHC) on C6. Other high-density SNP peaks seen for the tri-allelic SNPs patterns; e.g. at the q-arm telomere of C14, or p-arm and q-arm telomeres of C16, may indicate regions with higher than average levels of segmental duplication. All other tri-allelic SNP densities are quite evenly distributed, and SNP deserts generally align with the position of constitutive and/or facultative heterochromatin on each chromosome, with the exception of the X centromere and some displacement on C6 and C18, which might be accounted for by inaccurate positioning of the centromeres on some ideograms used in the figure.

Remarkably, all the SNPs identified to be tri-allelic and assembled in the 1000 Genomes Phase III variant data release, have assigned rs-numbers. The only exceptions were 358 tri-allelic SNPs on C12 with interim "ss" numbers (e.g. the 5′-most tri-allelic SNP without an rs-number has been assigned ss1388022860; ss1388022861 temporary descriptors), and 612 X chromosome loci currently identified by co-ordinates only (e.g. X:144916948). One highly polymorphic SNP identified by such numbers: ss1388042575; ss1388042576 at 12:7645262 (GRCh37), was selected for inclusion in the MPS Panel.

Proportions of tri-allelic SNPs with a nucleotide substitution allele seen in just one genotype (more accurately described as a mutation rather than an allele at the variant site) were observed at a consistent, but very low level in all the autosomes. Values ranged from 1.02 % singleton alleles observed amongst all tri-allelic SNPs detected on C7, up to 1.47 % on C16. The X chromosome showed a completely contrasting pattern, with single allele-3 genotypes seen in 4,631 of 14,267 tri-allelic SNPs, and single allele-2 and allele-3 genotypes seen in a further 1,594: totalling ~43.6 % of tri-allelic SNPs on the X. This could be partly due to the lower SNP density in general on the X compared to the autosomes or may relate to the way SNPs were characterized in hemizygous males of 1000 Genomes and compared to female patterns. It may also be relevant that the X shows lower levels of SNP polymorphism across the chromosome as a whole - we note the proportion of SNPs with heterozygosity levels > 0.6 on C7 is 5-fold higher than in a similar number of tri-allelic SNPs on the X: 109 of 14,635 (0.74 %) vs 21 of 14,267 (0.15 %).

From the 271,934 tri-allelic SNPs identified, more than 3 % (4,234) had overall heterozygosity values higher than the binary SNP maximum of 0.5, while a total of 1,637 SNPs had the highest variation levels of 0.6–0.666 heterozygosity (i.e. column 7, Table 2). This data creates a rich source of highly polymorphic SNPs for forensic applications, which can provide short amplicon PCR, and in many loci, additional scope for population differentiation from doubling the number of potential genotypes, which may more often vary in frequency between populations through random genetic drift. We concentrated on the use of the most polymorphic tri-allelic SNPs found in 1000 Genomes for forensic identification purposes and recognize that these loci also offer a more secure system for detecting mixed DNA, from the presence of three alleles in a proportion of the tri-allelic SNPs in a large forensic multiplex.

3.2. Relationship test simulations

3.2.1. Kinship informativeness of 1,377 autosomal tri-allelic SNPs

Note the results of the kinship simulations described below were based on autosomal loci only (i.e. 1,411 – 34 X SNPs = 1,377 SNPs).

Likelihood distributions from relationship test simulations to evaluate the informativeness of the 1,377 SNP set parsed for cM spacing are shown for full siblings, half siblings, first cousins and second cousins in Fig. 3. For full siblings, half siblings and first cousins the likelihood distributions clearly separate related and unrelated individuals, demonstrating that such relationship scenarios would be readily distinguished using such a panel. For the second cousin test, the medians of the H1:H2 LRs were different for related and unrelated individuals, but distributions overlapped to some extent - indicating that a proportion of second cousin tests would be inconclusive: unable to distinguish related and unrelated hypotheses. Using an arbitrary threshold of 10,000 for the LR, the exceedance probabilities were estimated to be 1.000, 1.000, 0.9885, and 0.1056 for full sibling, half sibling, first cousin and second cousin tests, respectively.

We briefly explored the effect of taking linkage into account when calculating kinship LRs from 1,377 tri-allelic SNPs, by using the ILIR framework of FamLink developed by Tillmar and Phillips [15]. Linkage adjustments were applied for the two scenarios of full siblings vs unrelated pairs and first cousins vs unrelated pairs, and the LRs produced were compared to identical calculations ignoring linkage. The resulting pairwise LR comparison plots are shown in Supplementary Fig. S1 (plots A and C: $\log_{10}$ LR accounting for linkage plotted against $\log_{10}$ LR ignoring linkage; plots B and D: $\log_{10}$ LR accounting for linkage plotted against $\log_{10}$ [LR ignoring linkage/LR accounting for linkage]). The plots in Supplementary Fig. S1 indicate that accounting for linkage has a strong effect on LR calculations: lowering the LR values in unrelated pairs (i.e. bringing them closer to zero), while reducing their spread of values. In contrast, simulated related pairs have increased LR values and in the case of first cousins, when accounting for linkage there is no overlap with unrelated pairs, but when ignoring linkage significant levels of LR overlap occur. Since the cM-parsed SNP set had an average spacing of 2.5 cM for loci on the same chromosome, it is not unexpected that adjustment for linkage has a major effect on the distribution of LR values shown in Supplementary Fig. S1, and we assert the need to apply these linkage adjustments in relationship testing using such a dense SNP set as the MPS Panel.

3.2.2. Binary SNPs vs tri-allelic SNPs

The likelihood distributions from relationship test simulations exploring the power of 1,377 tri-allelic SNPs compared to the same number of perfect binary SNPs are shown in Fig. 4. The 1,377 autosomal SNPs used for the simulations gave much higher likelihoods in analyses of full sibling, half sibling, and first cousin relationship scenarios, but were only slightly higher for the second cousin tests. Despite the lack of power to adequately test the most challenging second cousin relationships, SNPs with more than two alleles per locus are evidently more efficient in separating IBD-alleles (identity by descent) from IBS-alleles (identity by state). For this reason, SNPs developed for forensic identification can give similar discrimination power to STRs when sufficient numbers are analysed (approximately 50 SNPs can match the power of 16 STRs); but for relationship testing where allele sharing, not genotype sharing, is the basis for the statistical tests used, many more SNPs are needed to reach the power of an STR multiplex in such tests. In the identification of missing persons, relationship testing is critical in establishing links between the missing and their surviving relatives, yet longer STR amplicons will commonly fail in degraded DNA. Therefore, the assembly of a large number of tri-allelic SNPs in an MPS-based assay consisting of short amplified fragments, represents a major step forward in this field.

3.2.3. Balanced tri-allelic SNP allele frequencies vs loci with one low frequency allele

The impact of tri-allelic SNP allele frequency distributions on relationship testing likelihoods are shown in Fig. 5. In general, the likelihoods indicate it is more efficient to use a uniform allele frequency distribution compared loci with one allele at a relatively low frequency. However, these differences are only slight, and the data suggests that having a proportion of SNPs with low allele-3 frequencies has a minimal effect on the relationship testing power of the MPS Panel we designed for the purpose. No noticeable differences could be observed

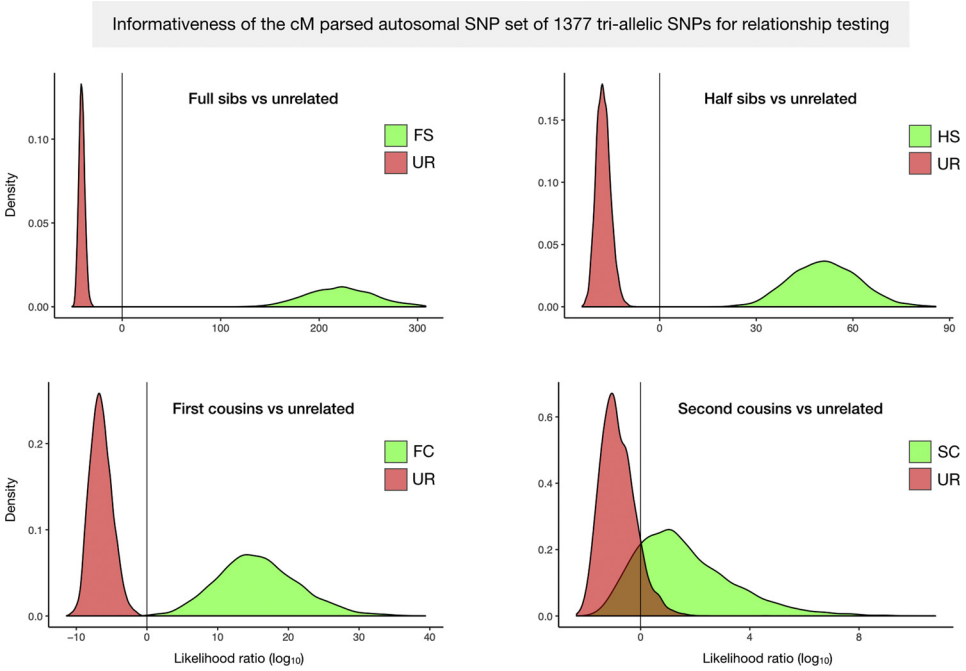


Fig. 3. Log LR distribution plots obtained from 5,000 simulations of relationship tests using European allele frequencies of 1,411 autosomal tri-allelic SNPs compiled for the MPS panel (i.e. after cM parsing but including SNPs discarded for sequence quality reasons). Relationship tests comprised full siblings vs unrelated (upper left), half sibs vs unrelated (upper right), first cousins vs unrelated (lower left) and second cousins vs unrelated (lower right). All distributions were separated from the zero line of balanced odds of H1/H2, apart from second cousins with ~10–15 % of LR values giving erroneous assignments.

for the sizes of the SNP sets or for the different case scenarios. Although we avoided SNPs with this pattern in one or more 1000 Genomes population groups, the loss of power is minimal from their use in large-scale panels.

3.3. Construction of the MPS Panel: a large-scale massively parallel sequencing assay of the most polymorphic tri-allelic SNPs

To compile a candidate set of the most polymorphic tri-allelic SNPs, all loci with an average heterozygosity greater than 0.5, calculated from 1000 Genomes African-European-East Asian population data, were identified. The South Asian heterozygosities tended to be higher than these average values, but lower in Americans, leading to many potentially useful SNPs with quite skewed allele frequencies. Numbers of compiled candidate SNPs per chromosome are shown on the right-hand pane of Table 2, which totalled 3,372 autosomal SNPs plus 43 X SNPs. The full list of selected tri-allelic SNPs is given in Supplementary Table S1A.

As a significant number of SNP clusters of closely sited loci were detected amongst the candidates, the minimum-cM-separation parsing step reduced the candidates to 1,411 SNPs. These SNPs were submitted to Qiagen for custom primer design for the QIAseq MPS system, with all SNPs successfully targeted with the QIAseq primer designs. The genomic positions of the 1,411 SNPs selected for the MPS Panel are shown in Fig. 6. The reduced list of candidate SNPs sent for QIAseq primer design after cM parsing is given in Supplementary Table S1B.

Despite a complete set of QIAseq primer designs, almost 10 % of the SNPs gave one of a range of sequencing or genotype issues. These were classified as: very low sequence coverage; detection of all three nucleotide substitution alleles in single individuals; discordant genotypes with control DNAs; and other MPS quality issues, including alignment problems due to polymeric tracts close to the target site and general

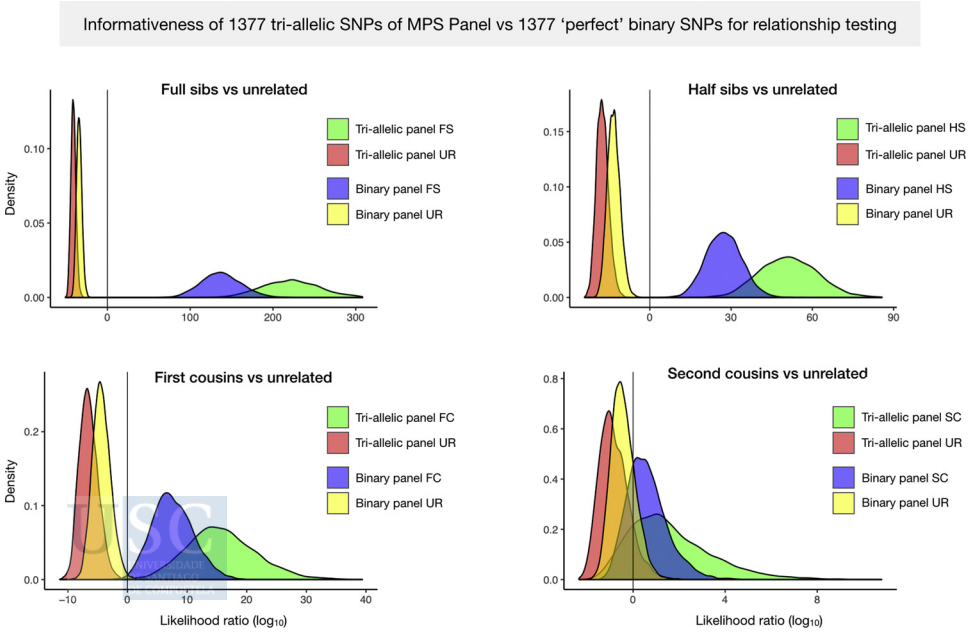


Fig. 4. The same Log LR plots as Fig. 3 shown with distributions from an equivalent number of perfect binary (0.5:0.5 bi-allelic) SNPs for each relationship testing scenario. While the binary SNP distributions are less separated, only second cousins have significant proportions of LRs that cross the zero line with both SNP sets, but there are approximately twice as many erroneous assignments from binary loci.

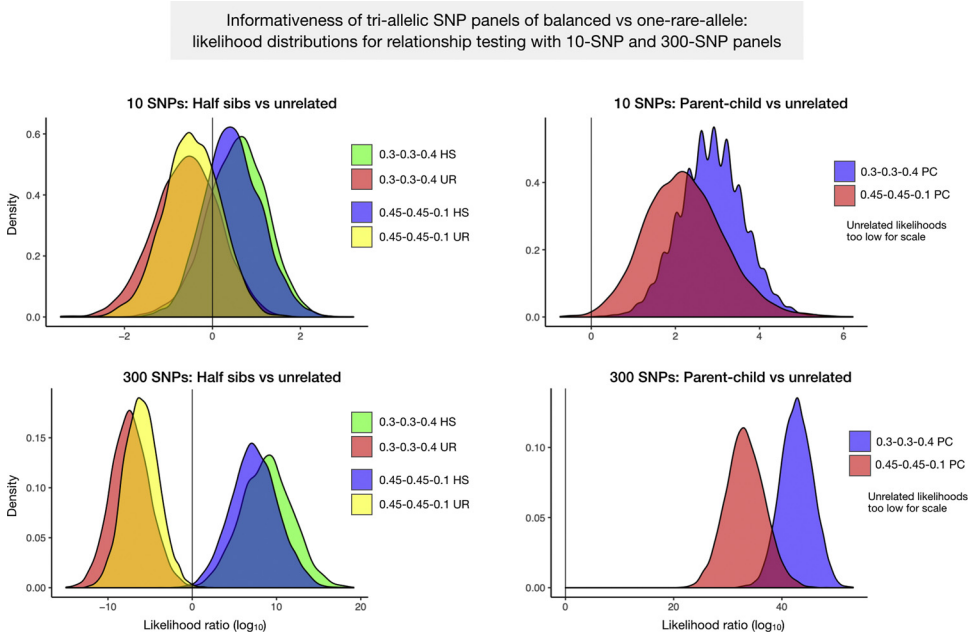


Fig. 5. Log LR plots from relationship test simulations exploring two contrasting allele frequency distributions in tri-allelic SNPs. A 10-SNP panel (upper) and 300-SNP panel (lower) were analyzed each with balanced allelic distributions of 0.34, 0.33, 0.33 frequencies, and one rarer allele of 0.5, 0.45, 0.05 frequencies. Generally, the differences in LRs were much smaller than might be expected, given the reduced polymorphism levels of SNPs with a rare allele-3. Note the two Parent-child test simulations on the right gave unrelated Log LR values too small to show, so the differences between balanced and one-rare-allele SNP LRs are slightly exaggerated.

flanking sequence quality problems. Total numbers in each class of discarded SNP are given in Table 2, with the 23 SNPs showing low sequence coverage used to calculate the QIAseq MPS assay conversion rate (ACR) of 98.4 %. An additional 51 SNPs with genotypes discounted due to flanking sequence issues are detailed in Supplementary File S1C, and because they were not strictly defined as loci that failed in MPS library preparation, were treated as 'converted to assay' in the above ACR calculation. With so many original candidate SNPs removed from the cM parsing step, it would have been relatively straightforward to replace the failing SNP with another from the same linkage cluster, but as the multiplex scale was still very high, we did not pursue this process further. Details of the individual issues in each SNP are given in Supplementary Table S1D; comprising 23 SNPs with low coverage (although this was observed in additional SNPs that had other sequence quality issues); 38 with three alleles per individual DNA genotyped; 29 with discordant genotypes in control DNAs; and 51 with other sequencing issues.

Such sequence quality issues are commonplace in any MPS test and can be expected - particularly in such a large multiplex. However, the observation of three alleles in some sequenced individuals in 38 SNPs and discordant genotypes in a further 29 raises more important concerns about the specificity of the genomic segments carrying some of the tri-allelic SNP sites we identified; i.e. whether they represent single positions in the genome or are the same SNP site on segmental duplications/multiple copy replications. Divergence of one of the replicated SNP alleles from allele-2 to a new allele-3, could mimic a single polymorphism with three alternative alleles, although it would be expected to be detected by the detailed scrutiny of 1000 Genomes data, albeit at lower coverage levels than those typical of MPS. In all 38 three-allele SNPs the detected nucleotide substitutions at the target site matched those expected from the 1000 Genomes genotype data. Despite BLAST analysis of candidate SNPs, the observation of three-allele patterns or genotype discordancy in over 5 % of sequenced loci indicates this proportion of tri-allelic SNPs are likely to represent non-specific genomic sites, or something unusual about the SNP variation at the targeted site, and many of these may escape detection with traditional sequence specificity checks. Fortunately, the 1,270 remaining SNPs were specific with no discernable MPS sequence quality issues, and assuming random failure regarding levels of polymorphism per SNP, the final multiplex design retained an estimated 89 % of the differentiation power of the original 1,411 SNP set, 1,377 of which were assessed in the kinship simulations. The 141 unreliably genotyped SNPs had primers kept in the PCR multiplex but their target site data was not retained and therefore these loci do not strictly form part of the final panel.

The distribution of average heterozygosity values amongst the 1,270 tri-allelic SNPs retained in the final MPS Panel is shown in Fig. 7. A total of 145 (11.4 %) had the highest values exceeding 0.6 (mid blue bars); 1,105 (87 %) had values above the binary SNP maximum of 0.5 (light blue bars); and only 20 (1.6 %) had average values slightly below 0.5 (pink bars). For comparison, the 46 microhaplotypes that were also incorporated into the MPS Panel are shown on the right side of Fig. 7. Microhaplotypes are generally more informative per locus than tri-allelic SNPs as most, although not all, of them have more than three common haplotypes. The equivalent average heterozygosity ranges for these loci were 21 (45.6 %); 8, (17.4 %) and 2 (4.4 %). An additional 15 (32.6 %) had average heterozygosity values exceeding the tri-allelic maximum value of 0.666, indicating the best microhaplotypes are worth incorporating into such a panel, albeit in much smaller numbers than can be compiled for tri-allelic SNPs.

3.4. Ancestry informativeness of the MPS panel

The detailed tri-allelic SNP genotypes from the combination of 1000 Genomes and SGDP population data for the 1,270 markers (1,241 autosomal) of the final MPS Panel are given in Supplementary Table S2A. Allele frequency estimates are provided for the six main population groups studied from SGDP, to which estimates from a limited number of Middle East population samples were added. The STRUCTURE input data is given in Supplementary Table S2B, and comprised the genotypes of Supplementary Table S2A, but removing Middle East, Central South Asian and North East Asian (Siberia) genotypes for clarity.

Assessments of the ancestry informativeness of the MPS Panel are summarized in Fig. 8A–C, the optimum cluster number from CLUMPAK analysis of STRUCTURE data was K:6 (Fig. 8A). The South Asian populations and SGDP sample set of mainly Pakistani populations are labelled, as this was the most varied set of genetic cluster patterns. Individual sample cluster membership proportions are listed in Supplementary Table S2C, with atypical values highlighted in red.

We adopted the simple approach of treating the 1000 Genomes data from the four population groups and SGDP data from American and Oceanian populations as representative of their origins, while including

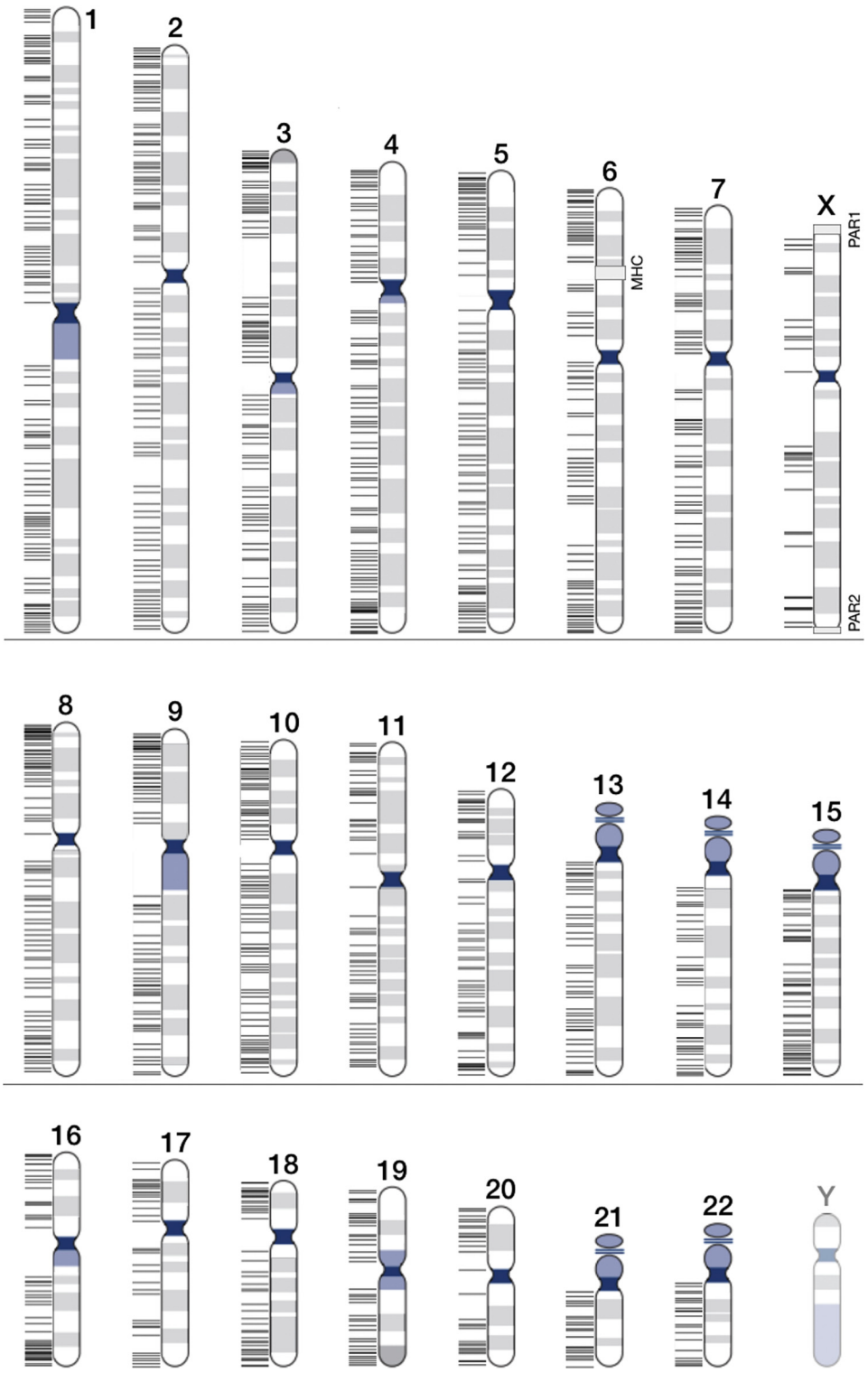


Fig. 6. Genomic positions of 1,411 tri-allelic SNPs selected for the MPS panel, scaled to the same chromosome ideograms as Fig. 2.

a widely dispersed set of other SGDP population samples from AFR, EUR, EAS and SAS regions. The bulk of samples gave very high STRUCTURE cluster membership proportions > 98 % to a single cluster and matching patterns in STRUCTURE, MDS analysis and Neighbor-joining tree positions. In the 1000 Genomes and SGDP AME − OCE samples the average cluster membership proportions were: AFR = 0.989 (min. 0.911, max. 0.999); EUR = 0.981 (0.82–0.998); EAS = 0.989 (0.877–0.999); SAS = 0.934 (0.562–0.934); AME = 0.963 (0.73-0.999); OCE = 0.958 (0.657–0.958).

Most SGDP SAS samples had lower cluster membership proportions than their 1000 Genomes counterparts, as these represented mainly HGDP-CEPH Pakistani population samples which are divergent to varying degrees from other Indian sub-continental populations. The three sets of samples representing population outliers are indicated by numbers 1–3 in the Neighbor-joining tree plot of Fig. 8C, as these are easier to point out than atypical MDS positions or STRUCTURE cluster proportions (individual columns are thin). Group 1 of SAS points in EUR comprises SGDP Balochi (0.5–0.6 SAS cluster proportions); Brahmi

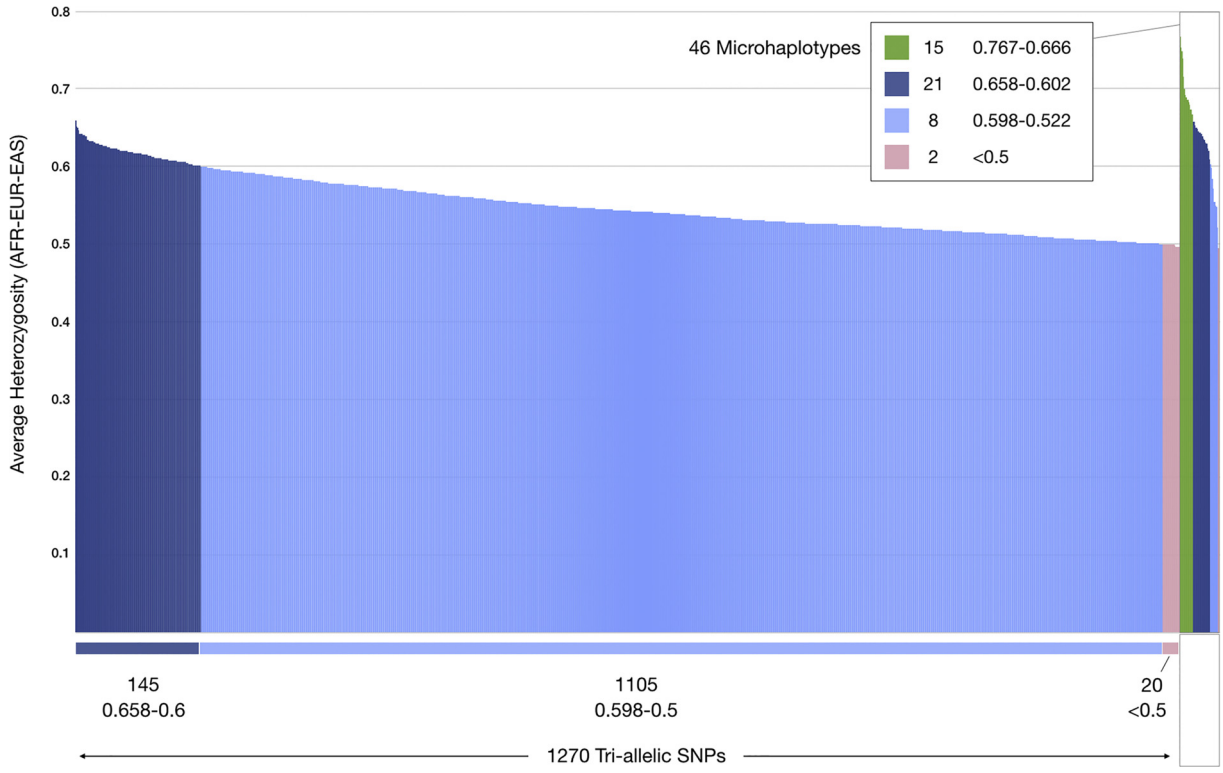


Fig. 7. Distribution of average heterozygosity values amongst 1,270 tri-allelic SNPs (bars left to center) and 46 microhaplotypes (bars on the far right), selected for the final MPS Panel. Of the tri-allelic SNPs, 145 (11.4 %) had the highest values exceeding 0.6 (mid blue bars); 1,105 (87 %) had values above the binary SNP maximum of 0.5 (light blue); and only 20 (1.6 %) were slightly below 0.5 average heterozygosity (pink). The microhaplotypes had 21, 8 and 2 (45.6 %, 17.4 %, 4.4 %) of loci in the equivalent heterozygosity ranges plus an additional 15 (32.6 %) with average heterozygosity exceeding the tri-allelic maximum value of 0.666 (green bars). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

(0.5); Burusho (0.6–0.7); Kalash (0.5); Makrani (0.4–0.6) and Sindhi (0.5–0.7) all from Pakistan, (note several points overlay each other). Group 2 of SAS points in OCE comprises SGDP Hazara (0.2–0.37 SAS cluster) and Khonda Kusunda (0.4–0.5) from Pakistan and Nepal respectively. Group 3 of OCE points in EAS are SGDP Igorot from Philippines (there are two pink columns in the right side of the OCE clusters). A further two SGDP Iglirit population samples were wrongly labelled as AME (as they share the same name as a Native North American population from Canada), when they are actually from Central Asia, but are not evident on the Neighbor-joining tree, although with zero AME membership proportions.

Although just representing a snapshot of the ancestry inference capacity of 1,241 autosomal tri-allelic SNPs in the final MPS Panel, the patterns of population divergence the above analyses detected, notably in SGDP samples, follow the known demographics of many of these samples and indicates that a large-scale panel, despite being a SNP set selected to have low population differentiation, actually provides useful levels of ancestry information when analyzed as one substantial dataset of variation.

4. Concluding remarks

The original searches for tri-allelic SNPs that this study made of the 1000 Genomes Phase I variant database were informed by the SNPs we had already identified, adopted for forensic use and comprehensively characterized by capillary electrophoresis (CE) - which confirmed their tri-allelic status [5,8]. In addition, a dedicated forensic CE test to genotype established tri-allelic SNPs has since been developed [34]. However, these established tri-allelic SNPs represent just a handful of loci, and the need to create a better framework for mixed DNA analysis in forensic SNP genotyping tests, and to gain more information per marker, provided the motivation to screen 1000 Genomes Phase I data in detail. It was soon apparent that this stage of the project was reporting SNPs known to have tri-allelic variation as bi-allelic loci and we put further searches aside. It is likely that so much variant data was generated in the Phase I whole-genome sequencing (sample numbers were 1,092 individuals, yielding 36.7 M SNP sites [3]) that it was not viable at that stage of the 1000 Genomes project to distinguish true tri-allelic SNPs from sequencing or alignment artefacts. For example, the juxtaposition of a binary SNP substitution site with a single nucleotide deletion at the adjacent 5′ or 3′ position is a common cause of incorrect identification of three alleles at the SNP site. The absence of tri-allelic SNP data from other large-scale variant discovery programs (e.g. EGDP: Estonian Biocentre Genome Diversity Panel analyses [35]) is exacerbated by the widespread use of PLINK [36], a toolbox for processing sequence and variant data. PLINK removes tri-allelic SNP genotype calls with the inference that the majority of such variant sites are likely to be misalignment, base misincorporation or dual SNP-Indel artefacts. EGDP has applied PLINK and reports no tri-allelic sites in their sequence analysis of 402 individual genomes in 126 populations [35].

The search for tri-allelic SNPs potentially useful for forensic analysis was resumed when Complete Genomics published whole-genome sequences for 427 individuals [37] and reported three alleles in the forensic SNPs known to be tri-allelic (Table 1). As Complete Genomics had used the same population samples sequenced in 1000 Genomes Phase I, it was possible for both projects to cross-check their variant calls across the whole genome of each individual. Therefore, when the 1000 Genomes Phase III variant data was released, it was evident that many of the previously discounted tri-allelic SNPs were now described, including those we had validated with CE genotyping. Phase III analyses had the additional advantage that more populations were added (notably five from South Asia), allowing the expanded identification of tri-allelic SNPs with population specific, but often low frequency, allele-3 variants.

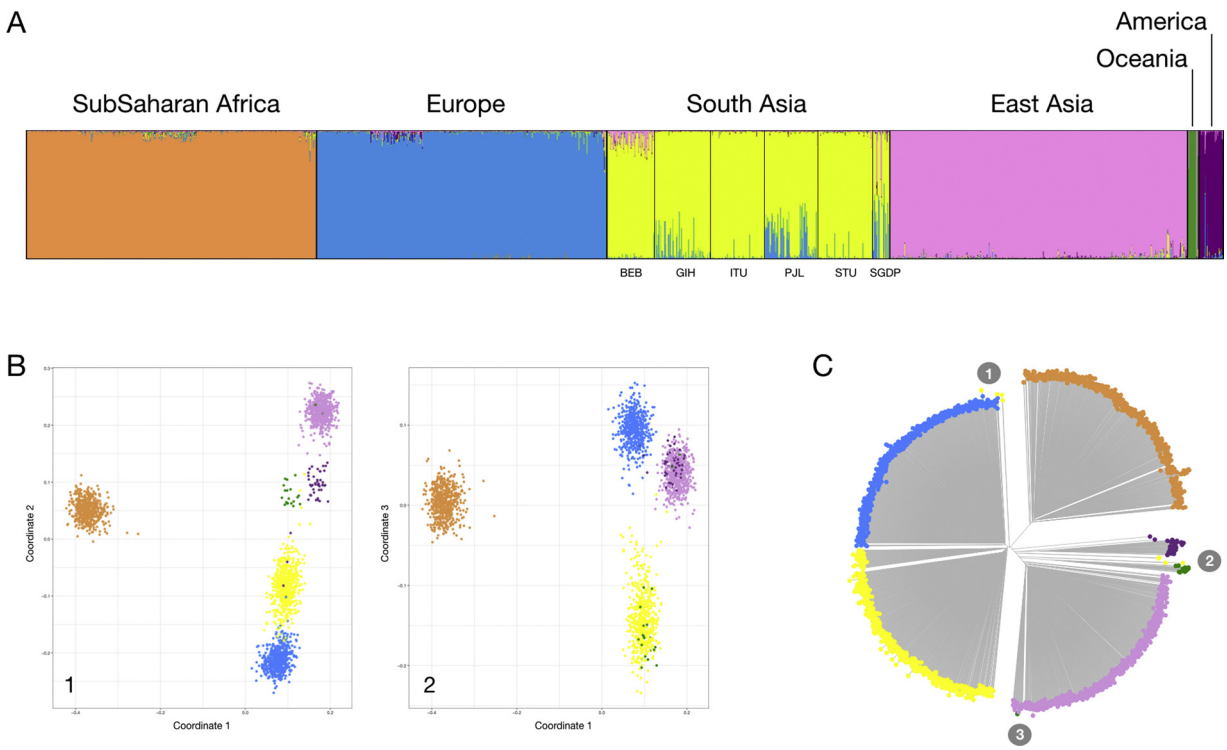


Fig. 8. An ancestry informativeness 'snapshot' of the 1,241 autosomal tri-allelic SNPs of the MPS Panel analyzed with STRUCTURE, MDS and Neighbor-joining tree tests. Samples from the six main population groups were taken from 1000 Genomes and Simons Foundation genome diversity project (SGDP). South Asian populations showed the highest levels of intra-group divergence so individual populations are indicated: BEB: Bengali from Bangladesh; GIH: Gujarati Indian from Houston, Texas; ITU: Indian Telugu from the UK; PJL: Punjabi from Lahore, Pakistan; STU: Sri Lankan Tamil from the UK. (A) STRUCTURE cluster plot for K:6 form CLUMPAK analysis. (B) Representation of the 3 dimensions of MDS analysis: coordinates 1 vs 2 (left) and 1 vs 3 (right). (C) Neighbor-joining tree. Three sample sets labelled are: set 1, SAS points in EUR comprising SGDP Balochi; Brahmin; Burusho; Kalash; Makrani and Sindhi (all from Pakistan); set 2, SAS points in OCE comprising SGDP Hazara and Khonda Kusunda from Pakistan and Nepal respectively; set 3 of OCE points in EAS of SGDP Igorot from Philippines. Note each set has points overlaying each other.

The full compilation of human tri-allelic SNPs presented here complements the study made of human tetra-allelic SNPs [11], which identified 961 loci passing sequence QC thresholds in 1000 Genomes Phase III data. Although only a handful of SNPs from this study had forensically useful levels of polymorphism, the fact that more than 280-fold higher numbers of tri-allelic SNPs have now been collected from 1000 Genomes means a much larger pool of tri-allelic SNPs can be assembled. These are then easily ranked by heterozygosity to identify the most informative markers for forensic identification tests. We were not expecting so many SNPs to be incorporated into the prototype MPS Panel built around QIAseq chemistry. Given the size of this multiplex, the > 98 % ACR to assemble more than 1,300 loci (including micro-haplotype markers) with sufficient sequence coverage was a major achievement from the Qiagen primer design team. The need to discard some 5 % of the selected SNPs because of indications of unreliable genotyping performance, or their non-specific sequences from what are likely to be multiple positions in the genome, should not be surprising. About 20 % of the tetra-allelic SNPs we have analyzed further in MPS tests since their discovery give indications of non-specific positions (personal communication, Peter de Knijff, Leiden University, Netherlands). These findings highlight the need for caution when scrutinizing the sequence characteristics and mapping details of tri-allelic variants, as such patterns can be caused by common structural variation in the genome.

Nevertheless, assembling 1,270 of the most polymorphic tri-allelic SNPs into a single MPS multiplex represents a significant development in the application of this sensitive and data-rich sequencing technology to forensic identification.

Acknowledgements

This study is 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. 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).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2020.102232.

References

[1] 1000 Genomes Project Consortium, G.R. Abecasis, D.L. Altshuler, A. Auton, L.D. Brooks, R.M. Durbin, R.A. Gibbs, M.E. Hurles, G.A. McVean, A map of human genome variation from population-scale sequencing, Nature 467 (2010) 1061–1073.
[2] 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, et al., A global reference for human genetic variation, Nature 526 (2015) 68–74.
[3] 1000 Genomes Project Consortium, G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, et al., An integrated map of genetic variation from 1,092 human genomes, Nature 491 (2012) 56–65.
[4] J. Amigo, A. Salas, C. Phillips, ENGINES: exploring single nucleotide variation in entire human genomes, BMC Bioinf. 12 (2011) 105.
[5] C. Phillips, M.V. Lareu, A. Salas, Á. Carracedo, Nonbinary single-nucleotide polymorphism markers, Int. Congress Ser. 1261 (2004) 27–29.
[6] M. Cao, J. Shi, Ji. Wang, J. Hong, B. Cui, G. Ning, Analysis of human triallelic SNPs by Next-Generation Sequencing, Ann. Hum. Genet. 79 (2015) 275–281.
[7] C. Phillips, A. Salas, J.J. Sanchez, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios,

M. Calaza, M. Casares de Cal, D. Ballard, M.V. Lareu, Á. Carracedo, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci. Int. Genet. 1 (2007) 273–280.
[8] A.A. Westen, A.S. Matai, J.F.J. Laros, H.C. Meiland, M. Jasper, W.J.F. de Leeuw, P. de Knijff, T. Sijen, Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples, Forensic Sci. Int. Genet. 3 (2009) 233–241.
[9] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Børsting, P. Johansen, M. Fondevila, et al., Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set, Forensic Sci. Int. Genet. 11 (2014) 13–25.
[10] Z. Gao, X. Chen, Y. Zhao, X. Zhao, S. Zhang, Y. Yang, Y. Wang, J. Zhang, Forensic genetic informativeness of an SNP panel consisting of 19 multi-allelic SNPs, Forensic Sci. Int. Genet. 34 (2018) 49–56.
[11] C. Phillips, J. Amigo, A. Carracedo, M.V. Lareu, Tetra-allelic SNPs: informative forensic markers compiled from public whole-genome sequence data, Forensic Sci. Int. Genet. 19 (2015) 100–106.
[12] S. Cornelius, Y. Gansemans, A.S. Vander Plaetsen, J. Weymaere, S. Willems, D. Deforce, F. Van Nieuwerburgh, Foerensic tri-allelic SNP genotyping using Nanopore sequencing, Forensic Sci. Int. Genet. 38 (2019) 204–210.
[13] A.J. Hodgkinson, A. Eyre-Walker, Human tri-allelic sites: evidence for a new mutational mechanism? Genetics 184 (2010) 233–241.
[14] P.A. Jenkins, J.W. Mueller, Y.S. Song, General triallelic frequency spectrum under demographic models with variable population size, Genetics 196 (2014) 295–311.
[15] A.O. Tillmar, C. Phillips, Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets, Forensic Sci. Int. Genet. 26 (2017) 58–65.
[16] J.J. Sanchez, C. Phillips, C. Borsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P.M. Schneider, A. Carracedo, N. Morling, A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713–1724.
[17] 1000 Genomes FTP site [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/] accessed 2014-2019.
[18] L. Clarke, S. Fairley, X. Zheng-Bradley, I. Streeter, E. Perry, E. Lowy, A.M. Tassé, P. Flicek, The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data, Nucleic Acids Res. 45 (2017) D854–D859.
[19] M. Lek, K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A.H. O'Donnell-Luria, J.S. Ware, A.J. Hill, B.B. Cummings, et al., Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (2016) 285–291.
[20] J.A. Brody, A.C. Morrison, J.C. Bis, J.R. O'Connell, M.R. Brown, J.E. Huffman, D.C. Ames, A. Carroll, M.P. Conomos, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, et al., Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology, Nat. Genet. 49 (2017) 1560–1563.
[21] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, et al., The Simons Genome Diversity Project: 300 genomes from 142 diverse populations, Nature 538 (2016) 201–206.
[22] C. Phillips, D. Ballard, P. Gill, D. Syndercombe Court, Á. Carracedo, M.V. Lareu, The recombination landscape around forensic STRs: accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data, Forensic Sci. Int. Genet. 6 (2012) 354–365.
[23] D.R. Zerbino, P. Achuthan, W. Akanni, M.R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C.G. Giron, et al., Ensembl 2018, Nucleic Acids Res. 46 (2018) D754–761.
[24] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The Human Genome Browser at UCSC, Genome Res. 12 (2002) 996–1006.
[25] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[26] D. Kling, T. Egeland, A.O. Tillmar, FamLink - a user friendly software for linkage calculations in family genetics, Forensic Sci. Int. Genet. 6 (2012) 616–620.
[27] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016.
[28] M. Kruijver, Efficient computations with the likelihood ratio distribution, Forensic Sci. Int. Genet. 14 (2015) 116–124.
[29] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.
[30] N.M. Kopelman, J. Mayzel, M. Jakobsson, N.A. Rosenberg, I. Mayrose, Clumpak: a program for identifying clustering modes and packaging population structure inferences across K, Mol. Ecol. Resour. 15 (2015) 1179–1191.
[31] R: A language and environment for statistical computing. http://www.r-project.org/.
[32] E. Paradis, pegas: an R package for population genetics with an integrated-modular approach, Bioinformatics 26 (2010) 419–420.
[33] E.Y.Y. Cheung, C. Phillips, M. Eduardoff, M.V. Lareu, D. McNevin, Performance of ancestry-informative SNP and microhaplotype markers, Forensic Sci. Int. Genet. 43 (2019) 102141.
[34] C. Phillips, L. Manzo, M. de la Puente, M. Fondevila, M.V. Lareu, The MASTiFF panel - a versatile multiple-allele SNP test for forensics, Int. J. Legal Med. (2019), https://doi.org/10.1007/s00414-019-02233-8.
[35] L. Pagani, D.J. Lawson, E. Jagoda, A. Mörseburg, A. Eriksson, M. Mitt, F. Clemente, G. Hudjashov, M. DeGiorgio, L. Saag, et al., Genomic analyses inform on migration events during the peopling of Eurasia, Nature 538 (2016) 238–242.
[36] PLINK guidelines: [https://www.cog-genomics.org/plink2] accessed November 2019.
[37] N. Rieber, M. Zapatka, B. Lasitschka, D. Jones, P. Northcott, B. Hutter, N. Jäger, M. Kool, et al., Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies, PLoS One 11 (2013) e66621.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Article 6. Inter-platform evaluation of the MPSplex large-scale tri-allelic SNP panel for forensic identification

Article R1 (revised and re-submitted)

Reference: NA

ISSN: NA

DOI: NA

Web link: NA

Authors: Ruiz-Ramírez J[1], Bittner F[2], Parsons TJ[2], Tillmar A[3,4], Vangeel L[2], Grandell I[3], Eduardoff M[2], Peck M.A.[2], Ambroa-Conde A[1], Mosquera-Miguel A[1], Freire-Aradas A[1], Lareu MV[1], Phillips C[1,5], de la Puente M[1]

Author information:
1. Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain.
2. Fundación Pública Galega de Medicina Xenómica SERGAS, Grupo de Medicina Xenómica USC, IDIS, Santiago de Compostela, Spain.
3. Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden.
4. Department of Clinical and Experimental Medicine, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden.
5. International Commission on Missing Persons, Koninginnegracht 12, The Hague, Netherlands.

Quality metrics: NA

Contribution: methodology, MPS sequencing, collection of population data, data curation, investigation, formal analysis, investigation, resources, software, visualization, manuscript writing

Authorization: the article was submitted to Forensic Science International. Genetics, where authors retain the right to include the article within a thesis or dissertation, provided it is not published commercially. Permission is not required as long as the journal is referenced as the original source.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Inter-platform evaluation of the MPSplex large-scale tri-allelic SNP panel for forensic identification

Ruiz-Ramírez J[1*], Bittner F[2], Parsons TJ[2], Tillmar A[3,4], Vangeel L[2], Grandell I[3], Eduardoff M[2], Peck M.A.[2], Ambroa-Conde A[1], Mosquera-Miguel A[1], Freire-Aradas A[1], Lareu MV[1], Phillips C[1,5], de la Puente M[1*]

1 Forensic Genetics Unit, Institute of Forensic Sciences, Universidade de Santiago de Compostela, Santiago de Compostela, Spain
2 International Commission on Missing Persons, Koninginnegracht 12, The Hague, Netherlands
3 Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
4 Department of Biomedical and Clinical Sciences, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden
5 King's Forensics, Faculty of Life Sciences and Medicine, King's College, London, UK
*Corresponding authors.
E-mail addresses: jruiz.ramirez@usc.es (J. Ruiz-Ramírez), mdelcarmendela.puente@usc.es (M. de la Puente)

Keywords: Forensic Genetics; Tri-allelic SNPs; Unique Molecular Indices (UMIs); Missing Persons Identification; Massively Parallel Sequencing

Highlights

- The MPSplex SNP panel evaluated using MiSeq, Ion S5 and GeneReader MPS platforms
- Several underperforming markers were identified
- A casework kinship case is used to exhibit the power of the multiplex for relationship testing
- Results indicate robust and reliable performance of MPSplex as a medium-scale forensic SNP genotyping assay

Abstract

MPSplex is a large-scale forensic massively parallel sequencing (MPS) panel with 1,270 tri-allelic SNPs, 44 microhaplotypes (MH) and 55 ancestry-informative bi-allelic SNPs (aiSNPs) designed for missing persons identification. We have evaluated MPSplex with the most widely used MPS platforms in the forensic field: the Illumina MiSeq, the Thermo Fisher Scientific Ion S5 and the Qiagen GeneReader. The tri-allelic SNPs of MPSplex were previously identified from the most polymorphic loci with three common alleles in 1000 Genomes Phase III data and combined with the 44 MH and 55 aiSNPs, then implemented into a QIAseq Targeted DNA Custom Panel (Qiagen), a marker panel which uses Unique Molecular Indices or UMIs. The UMI random-sequence DNA molecules are incorporated onto DNA fragments before the Target Enrichment PCR, allowing the identification of reads that originated from the same template and consequently they can be used to correct the errors that may arise within the PCR or the sequencing process. In this study, we present the results of an inter-platform evaluation of the MPSplex panel, characterizing its performance in different forensic scenarios, which assessed aspects that include sensitivity, genotyping accuracy and mixture analysis. MPSplex aims to provide a tool designed for kinship analysis that can be applied beyond the resolution of first- or second-degree relationships, avoiding the need for much bigger forensic panels designed for genealogy purposes, which usually require significantly more sequencing resources. This study provides evaluation of MPSplex using the MPS systems in routine use for forensic genotyping of large-scale panels of SNPs.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1. Introduction

The implementation of massively parallel sequencing (MPS) technologies in the forensic field has led to the creation of many different large-scale panels, including expanded STR panels (e.g., ForenSeq MainstAY from Verogen, PowerSeq® 46GY System from Promega or Precision ID GlobalFiler™ NGS STR Panel v2 from Thermo Fisher Scientific (TFS)), panels of single nucleotide polymorphisms (SNPs), and panels comprising microhaplotypes (MHs) [1-3], a type of marker which cannot be analysed through traditional capillary electrophoresis (CE) procedures. Many such panels are solely composed of SNPs, which can be used not only for identification purposes, but also to predict bio-geographic ancestry (BGA) and externally visible characteristics (EVCs) [4-7].

While SNPs have lower polymorphism levels than STRs so are less informative per locus, the capability of MPS to simultaneously analyse many more markers per run than conventional multiplexed assays based on capillary electrophoresis, and the abundance of SNPs in the human genome, has led to the development of panels targeting up to a thousand or more SNPs. Panels of SNPs at this scale have the most utility in kinship studies where the putative relationship analysed is distant and STRs cannot provide enough power to resolve the degree of relatedness - commonly beyond second-degree relatives. Currently, a variety of MPS SNP panels developed by the community or commercially available, range in scale from ~100 SNPs [8, 9] to the bigger multiplexes specifically designed for genealogical purposes: Verogen's ForenSeq Kintelligence Kit comprising 10,230 SNPs; and the FORCE panel, containing 3,929 kinship SNPs [10].

We previously developed and optimised the MPSplex panel, designed to enhance the identification of missing persons when STRs could not resolve the putative relationship of human remains to reference samples of relatives [11]. MPSplex consists of 1,270 carefully selected tri-allelic SNPs compiled from 1000 Genomes Phase III data supplemented with 44 MHs and 55 established ancestry informative SNPs (aiSNPs) [12], implemented in a QIAseq custom panel. The main characteristic of the QIAseq MPS panel design is the use of Unique Molecular Indices (UMIs), random nucleotide sequences that, when incorporated into DNA fragments prior to PCR amplification, can be used to identify all sequence reads that originated from the same template molecule. During posterior bioinformatic analysis, all these sequence reads are unified into a single read (the UMI read), allowing bioinformatic removal of errors introduced during PCR and the sequencing process. In addition, the panel utilizes a single region-specific primer for each locus, while the other primer is complementary to the adapter ligated (Supplementary Fig. S1). Although this technology is mainly used outside of forensic applications, such as somatic variant detection or transcript counting in RNA sequence analysis, UMIs have begun to be successfully implemented in several forensic panels [13-16].

In the present study, we evaluated the forensic performance of the MPSplex panel sequenced in the three MPS platforms which have been applied to forensic DNA analysis: Thermo Fisher Scientific's Ion S5, Verogen's MiSeq FGx and Qiagen's GeneReader.

2. Materials and methods

2.1. The MPSplex panel

The version of MPSplex used for this evaluation was developed as a QIAseq Targeted DNA Custom Panel (Catalogue number: CDHS-15861Z-2897) targeting a total of 1445 SNPs. In addition to the 1,270 tri-allelic SNPs originally selected in [11] (1,241 autosomal plus 29 X-SNPs) and 120 key targeted SNPs within 44 microhaplotypes (MHs) selected from [1] and [17], it includes 55 aiSNPs for ancestry prediction from [12]. Chromosomal locations and SNP

frequencies for the tri-allelics were previously compiled in Supplementary Table S1C from [11] (note that SNP allele frequencies for EUR and AFR populations were inverted, so incorrectly described in this data) and information on the 55 aiSNPs was compiled in the original publication [12]. Several MHs were shortened to reduce the targeted sequence amplicon size while designing their primers, using an approach similar that followed in [18]. A list of microhaplotype loci with their respective composite SNPs is in Supplementary Table S1.

For all data analysis in this study, the MH composite SNPs were treated as individual SNPs rather than as components of discrete sets forming the targeted microhaplotypes. Additionally, it was noted that aiSNPs rs870347, rs192655 and rs917115 had incorrect coordinates (e.g., rs870347 was incorrectly positioned on chromosome 6, not 5). Thus, primers for these three SNPs did not target the correct location and were excluded from further analysis.

2.2. Inter-laboratory validation

Evaluation of the performance of MPSplex across the three platforms involved MPS systems sited in three different locations. These were: the Ion S5 System (TFS) at Universidade de Santiago de Compostela, the MiSeq FGx (Verogen) at the National Board of Forensic Medicine from Sweden and the GeneReader (Qiagen) at the International Commission on Missing Persons. A set of shared DNA control samples were selected to be run independently with each platform.

2.2.1. Reference DNA sample set

The reference DNA set comprised: (i) the 2800M Control DNA (Promega); (ii) the NIST 2391d Standard Reference Material (SRM), components A and C; and (iii) 7 Coriell cell-line control DNAs of NA12878 and two trios comprising NA24143 (mother) - NA24149 (father) - NA24385 (son) (a CEU sample trio) and, NA06994 (father) - NA07000 (mother) - NA07029 (son) (an Ashkenazim sample trio). These samples were utilized for the assessment of sequencing quality metrics including sequence coverage (both total coverage and number of UMI reads), base misincorporation, strand bias and allele read frequency balance. The non-template negative control was used to measure the sequencing baseline.

2.2.2. Concordance analysis

The Coriell samples were used for both confirmation of mendelian inheritance across all markers and for concordance studies. The reference genotypes for these samples were retrieved from the 1000 Genomes project's high coverage sequencing data [19] for NA06994, NA07000 and NA12878; and the Genome-in-a-Bottle (GIAB) benchmark variant calls v4.2.1 [20] for the Ashkenazim Trio.

2.2.3. Sensitivity

A dilution series of 2800M was used to evaluate the sensitivity of MPSplex to low-level DNA, with DNA input amounts in the range: 1000 pg; 500 pg; 250 pg; 125 pg; 62 pg; 31 pg; and 7 pg. The consensus profile of the 1000 pg replicate from each platform was used to harmonize the inter-platform analysis of the percentage of concordant loci – i.e., for each SNP, a genotype was assigned when coinciding for at least two of the MPS platforms.

2.2.4. Mixture detection

To evaluate the ability of MPSplex to detect and analyse mixed DNA, SRM 2391d component D (NIST), a precisely constructed 3:1 mixture of components A and C, was analysed together with components A and C from the SRM 2391d reference set.

2.3. Proof-of-concept kinship analysis

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

A set of 3 samples from an indirect kinship case (details in section 3.5) showing low LR values when analysed with two CE-STRs kits of GlobalFiler™ (TFS) and Investigator HDplex (Qiagen) were run on the Ion S5 instrument to test the applicability of MPSplex for typical relationship testing casework. Samples had been previously extracted from buccal swabs using the AutoMate *Express*™ Forensic DNA Extraction System and the PrepFiler *Express*™ Forensic DNA Extraction kit (Applied Biosystems) and quantified with Quantifiler™ Trio DNA Quantification Kit (Applied Biosystems).

2.4. Library preparation and sequencing

Libraries were prepared using the QIAseq Targeted DNA Panel and QIAseq Indexes (Qiagen) for the Ion S5 and MiSeq, following the manufacturer's protocol for standard DNA conditions for each specific platform. The GeneRead QIAact Library Preparation Target Enrichment Reagents and GeneRead QIAact Lung DNA UMI Panel and Adaptor Kit (Qiagen) kits were used for GeneReader library preparation, using a 1.4x QIAseq beads-DNA ratio for adapter-ligated DNA cleanup and 1.2x for target enrichment and Universal PCR cleanups instead of the 1x ratio recommended in the protocol.

Although the QIAseq assay specifications stated that a minimum of 10 ng input DNA is required, given the forensic applications that the panel was designed for (i.e., samples with low concentration and/or highly degraded), and that the identification of low-frequency variants was not intended, most samples were amplified with lower amounts of input DNA. The specific input DNAs for each sample and MPS platform are outlined in Supplementary Table S2.

2.5. Sequencing

2.5.1. Ion S5

Ion S5 libraries were quantified with the Ion Library TaqMan™ Quantitation Kit on a 7500 Real-Time PCR System (TFS) following the manufacturer's protocol and subsequently pooled at equimolar concentrations at a final library pool of 50 pM. Templating was carried out in an Ion Chef Instrument (TFS) with the Ion 510 & Ion 520 & Ion 530 Kit-Chef using a 400 bp templating protocol. Sequencing was performed on an Ion 530™ Chip (with a capacity of 15-20 million reads per chip) using the Ion S5™ Sequencing Reagents (TFS) on the Ion S5™ Instrument (TFS).

2.5.2. MiSeq

MiSeq libraries were quantified using the Qubit 2.0 fluorometer (TFS) and DNA integrity was evaluated using the High Sensitivity DNA kit on the 2100 Bioanalyzer (Agilent Technologies). Samples were further diluted to 4 nM based on the quantification and fragment size distribution. Libraries were then pooled, denatured, and further diluted to 5 pM. The sequencing process was carried out on a MiSeq FGx instrument with the MiSeq FGx Reagent Kit, providing a flow cell capacity of 12.5 million reads (Verogen) using the paired-end 2 × 151 bp sequencing protocol. In addition, a QIAseq A Read 1 Custom Primer I was loaded according to the manufacturer's protocol.

2.5.3. GeneReader

Libraries for the GeneReader were quantified with the Qubit 3.0 fluorometer (TFS) and DNA integrity was verified with TapeStation 4200 and D5000 ScreenTape and D5000 reagents (Agilent Technologies). These were further normalised to 100 pg/µL, pooled in equimolar concentrations and diluted to 1 pg/µL. Clonal amplification was performed on the GeneRead QIAcube and sequencing was carried out on the GeneReader platform with the GeneRead UMI Advanced Sequencing Q kit on a GeneRead UMI Advanced Sequencing Q Flow Cell (Qiagen).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2.5. Sequence analysis and genotype calling

2.5.1. UMI analysis

CLC Genomics Workbench v.20.0.1 bioinformatics suite (herein CLC) was used for analysis of the sequencing data. Certain settings were adapted for each platform – e.g., use of paired-end reads for the MiSeq. FASTQ files were imported to CLC and processed with *Remove and Annotate with Unique Molecular Index* and *Universal adapter trimming* tools. Next, reads were aligned to the reference genome (hg19) using the *Map Reads to Reference*; and subsequently the *Calculate Unique Molecular Index Groups* and *Create UMI Reads from Grouped Reads* tools were used to generate UMI reads. Alignments were processed with the *Remove Ligation Artifacts* tool and then *Indel and Structural Variants* and *Prepare Guidance Variant Track* tools were used to identify structural variants. The *Local Realignment* tool was used to improve alignment quality and then primer sequences were removed from reads with *Trim Primers of Mapped Reads.* Finally, the *Identify Variants on Known Positions* utility was used to extract the data for each locus.

Genotype calling was performed on R software 4.3.1 (https://www.r-project.org/, [21]) using a 10 UMI threshold and, for reference samples, a 90% minimum allele frequency was applied for homozygous calls and a 20% minimum allele frequency for heterozygous calls. If a marker showed three alleles with allele frequencies above 20%, the genotype was not called. For mixture analysis, allele calling was made using a 10 UMI threshold and a minimum allele frequency of 10% for allele calling. This threshold was established considering prior knowledge of the mixture ratio, that represents a minimum allele frequency for the minor component of 12.5% for the 3:1 ratio (e.g., the genotype from the major contributor is AA and the minor contributor AG).

2.5.2. Independent sequence analyses without using UMI data

Raw FASTQ files were processed independently for the three kinship samples, using a pipeline based on open-source software and disregarding the UMI information. First, UMI tags, along with the common sequence, were trimmed from reads. Then, FASTQ files were aligned to hg19 reference genome using the bwa mem algorithm and later analysed with bam-readcount [22], applying a minimum base quality of 20 and mapping quality of 30. The output was subsequently imported into R and used for genotype calling. Parameters for genotype calling comprised an analytical threshold of 10 reads and an allele frequency of 90% and 20% for calling homozygous and heterozygous genotypes, respectively.

2.6. Statistical analyses

R scripts from ILIR and FamLink (v. 1.1) [23, 24] were used to generate simulations to obtain kinship probability distributions using the 1000 Genomes European haplotype frequency estimates and map distance estimates from the HapMap project provided in [11]. A posterior version of FamLink (v. 2.3) was used for LR calculation. Familias R package [25] was used to generate the pedigree plots.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3. Results

3.1. Inter-platform comparisons and overall performance assessments of MPSplex.

3.1.1. Total sequence coverage, UMI coverage and sequence read/UMI ratio

Fig. 1A represents the distribution of UMI coverage per marker for the reference samples that were amplified with 1-2 ng DNA input – 9, 3 and 10 samples for the Ion S5, MiSeq and GeneReader, respectively. These samples had an average of 159, 137 and 123 UMI reads per locus for the Ion S5, MiSeq and GeneReader, respectively. The proportions of the sequenced target loci of 99.8%, 99.8% and 99.0%, respectively, had a UMI coverage value above the calling threshold. Fig. 1B shows the coverage per marker for the full reference DNA sample set ordered by mean UMI coverage (excluding the high input samples). The Fig. 1B plot shows the differential amplification among markers typically seen in targeted sequencing MPS panels (e.g. [26, 27]), in which some markers display a coverage above the average while a few other markers fail to reach the 10 UMI threshold. Supplementary Fig. S2 contains individual boxplots for total coverage, UMI coverage and reads per UMI distribution for each of the reference samples plus the no template control.

It is noteworthy that most reference samples typed on the MiSeq and one on the Ion S5 (NA12878) were amplified using ≥10 ng input DNA (Supplementary Table S2). With the greater initial number of DNA copies as input, the UMI coverage for these samples was significantly higher despite all samples having similar depth of coverage, given all samples were added at equimolar concentrations, as indicated in Fig. 1B and Supplementary Fig. S2 for the above samples. Consequently, most samples with ≥10 ng input had an average of 1-2 reads per UMI, whereas for most of the samples with lower DNA input each UMI family contained >3 reads. These results outline the necessity of allocating different sequencing resources according to initial DNA input, and samples with higher DNA input should be sequenced with a higher depth of coverage to maintain the reads per UMI proportion, i.e., several reads for each UMI must be used for error correction.

3.1.2. Base misincorporation

The base misincorporation rate was estimated from the percentage of reads of non-called alleles divided by total coverage for the marker. Mean misincorporation values were relatively low, with mean values below 1%; i.e., levels similar to previous studies evaluating MPS panels for forensic use [26-28]. We obtained 0.564 ± 1.350, 0.253 ± 0.619 and 0.318 ± 0.920 percentage values for the Ion S5, MiSeq and the GeneReader, respectively; while a median value of zero was recorded across the whole panel due to the large proportion of markers showing no base misincorporation.

Overall, the Ion S5 exhibited the highest number of markers with high misincorporation values, with 73 loci presenting a mean misincorporation rate above 2%, 17 of which were above 5%. In contrast, the MiSeq had 15 loci above 2% and 1 above 5%, and the GeneReader had 18 above 2% and 2 above 5%. Three markers, rs4991456, rs60216576 and rs6764984, were observed to have high misincorporation rates for the three platforms. Another five showed high rates in two platforms: rs10795925, rs58318997 (MiSeq and GeneReader); rs13172903, rs370892 (S5 and MiSeq); and rs1762533 (S5 and GeneReader). Supplementary Fig. S3 contains misincorporation boxplots for each marker and Supplementary Table S3 lists all the SNPs with mean values above 2% and their respective mean, standard deviation and median values.

Although no SNP surpassed a 20% misincorporation rate (the genotype calling threshold used for heterozygotes), rs4991456 in the MiSeq and GeneReader and rs9933604, rs2184289 and rs2843150 in the Ion S5 gave noticeably high base misincorporation rates of ~18-19% for particular samples, so these five SNPs have a substantially higher potential to produce erroneous genotype calls, particularly in mixed DNA samples. The three markers with a high

misincorporation rate on the Ion S5 are all located close to homopolymeric tracts (poly-T tract near rs2184289, poly-G tracts near rs9933604 and rs2843150), so manual curation of these genotypes may be required. Inspection of rs4991456 did not reveal any alignment issues.

3.1.3. Strand bias

Supplementary Fig. S4 outlines the percentage of sequence reads in the forward strand vs reverse strand. Overall, markers showed a good balance between forward and reverse strand coverage. The Ion S5 showed the most balanced strand ratios, with 55.7% of markers having median values in the 45-55% range. Similarly, the MiSeq had 53.6% of markers in this range, while the GeneReader had a noticeably less balanced output with 35.4% of loci in the 45-55% strand ratio range, likely to be caused by lower UMI coverage.

It should be noted that SNPs that tend to be read only in one strand on the Ion S5 and the GeneReader, show a tendency to be read on the other strand on the MiSeq – i.e., markers that are mostly read in the forward strand on MiSeq are read on the reverse strand on the other platforms and vice versa. Although the reason for this phenomenon remains unknown, we suggest this could be caused by the way paired-end reads in the MiSeq are interpreted by the software, rather than related to the chemistry.

3.1.4. Allele read frequency balance

Major allele read frequency (MAF) data, defined in this study as UMI reads for the allele with the highest sequence coverage divided by total UMI coverage, is represented in Supplementary Fig. S5. Ideally, homozygous genotype calls should be close to 100% and heterozygous calls around 50% per allele, meaning that the sequencing assay is perfectly balanced for each allele. The MiSeq had the most balanced output out of the three platforms, with 96.0% of SNP calls falling in the homozygote coverage >90% and heterozygote coverage of 50-60%. The Ion S5 had 91.6% of loci in these ranges, and the GeneReader 88.6%. The proportion of genotype calls with more acute imbalance, that is, with major allele frequencies between 50-70%, was 1.4%, 2.8% and 3.2% for the MiSeq, Ion S5 and GeneReader, respectively. Despite most of the MiSeq samples having a higher DNA input, which may have led to better performance for this platform, the comparison of samples with ~1ng input versus samples with higher input did not reveal any differences (data not shown).

Although observing a small proportion of markers with imbalanced number of reads is expected, especially in a multiplex of this size, several markers consistently gave imbalanced genotypes across the different platforms (i.e., within the 70-90% MAF range). SNPs rs66525445, rs7623855, rs9617584, rs1097454, rs58318997, rs370892, rs10151146, rs4906213, rs8029886, rs10795925 and rs11103107 produced the highest number of imbalanced genotypes. SNPs rs370892 and rs10795925 were observed to also have high levels of base misincorporation, whilst rs58318997 had additional issues (detailed later). Although no other issues were observed in the eight remaining SNPs, their genotypes should also be interpreted with caution, as intermittent allele dropouts might be caused by undocumented neighbouring sequence features such as segmental duplications or copy number variation.

3.1.5. Baseline noise

Estimates of baseline sequencing noise from UMI coverage rates in the non-template negative control are outlined in Supplementary Fig. S6. Mean UMI coverage for the negative control was lower than 1 UMI per marker in the three platforms, with 29, 147 and 4 loci showing reads (representing ~2, 10 and 0.3% of the markers in the multiplex) for the Ion S5, the MiSeq and the GeneReader, respectively. The Ion S5 and GeneReader had a maximum number of 7 and 1 UMI reads, respectively; both below the 10 UMI calling threshold. In contrast, the MiSeq negative control had several markers above this calling threshold (rs7657799, rs784685, rs11531789 and rs10147017), detecting up to 23 UMI reads for rs10147017. Although these measures are based

on a single non-template control and an internal validation is required when implementing this tool in the analysis of casework DNA, these SNPs should have manual inspection.

3.2. Genotype concordance analysis

Genotype concordance assessed using reference samples with online genotype data from the independent sequencing analyses of the 1000 Genomes Project and GIAB Consortium, produced 7435 comparisons for each platform (the reference data did not contain full profiles in all cases). A concordance rate of 99.5% was obtained for the Ion S5, 99.7% for the MiSeq and 99.6% for the GeneReader, discounting 63, 31 and 129 no-calls; respectively. We found 33 discordances for the Ion S5 (consisting in 26 drop-ins, 1 drop-out and 6 incorrect genotypes), 19 for the MiSeq (11 drop-ins, 5 drop-outs and 3 incorrect genotypes) and 26 for the GeneReader (9 drop-ins, 15 drop-outs and 2 incorrect genotypes). Despite the overall high genotyping concordance rates, several markers were found to produce consistent genotype discordancies, some platform-specific, and others consistently observed across all three MPS systems. Alignments for SNPs that were suspected to present alignment artifacts were manually inspected with IGV and are outlined in Supplementary File S1.

On the Ion S5, SNPs rs319305 and rs6764984 produced erroneous calls due to the presence of homopolymeric tracts in the neighbouring region. In rs319305, a 3' 7-T tract next to the target site, combined with an InDel, coincides with an A drop-in, and in rs6764984, a 5' InDel next to the target site and the 3' GGGAAAAAAAA tract next to the target site are likely to affect reliable sequence alignment. The presence of other high-frequency variants in the vicinity of the target SNPs also hindered correct alignment and manual revision was challenging due to the complexity of the alignment (see Supplementary File S1). Another SNP, rs7623855, which was observed previously to have acute imbalance in several samples, had allelic dropouts in several samples typed on the GeneReader due to the differences in coverage between both alleles. Additionally, various allele dropouts were found for rs1809429 on the MiSeq and GeneReader, while Ion S5 genotype calls in this SNP were concordant. Reads containing the missing allele were found to align to an off-target region (coinciding with the SNP rs183110408), indicating that the drop-outs observed could be caused by the incorrect alignment of sequences due to this homology.

The results obtained in SNPs rs58318997 (previously observed to have high misincorporation rates and to be highly imbalanced) and rs80266372 gave rise to incorrect genotypes or drop-ins in several samples, independent of the platform. IGV inspection did not reveal any features hindering alignment despite some alleles with sequence reads for more than two different alleles. A BLAT query for the reads showed similarities to other regions, corresponding to SNPs rs201552694 and rs1253935158 respectively, where reads aligning in these regions were found (Supplementary File 1). These findings suggested off-target amplification causing incorrect genotyping, therefore these SNPs should be removed from any analysis. Finally, the genotypes obtained for rs7629097 and rs4461915 were discordant with the 1000 Genomes data for NA06994 and NA07000. However, the results were concordant amongst the sequencing platforms of our study, so this discordancy remains without explanation.

The two Coriell control DNA trios produced some departures from mendelian inheritance patterns. In the CEU trio, non-mendelian incompatibilities were observed in rs1809429 due to the dropout of markers (e.g., rs1809429 on the MiSeq caused by imbalanced reads for both alleles, rs6478720 on the Ion S5 due to the high misincorporation values due to the target SNP being close to a poly-C tract). Nonetheless, two persistent incompatibilities were found in the CEU trio for SNPs rs2009433 and rs399737. Both parents' genotypes were concordant with the genotypes from the 1000 Genomes Project and all examinations made of the sequencing output failed to provide an explanation (no apparent dropin/dropout, no hinderance to alignments). SNP rs399737

is reported to map to multiple locations in the 1000 Genomes Ensembl browser, so should be discounted as a component marker from all future analyses using MPSplex.

3.3. Sensitivity analysis

Fig. 2A shows the UMI coverage distribution for the 2800M dilution series and Fig. 2B plots the number of reads in each UMI family. Additional information about the total read depth observed per UMI is outlined in Supplementary Fig. S7. As would be expected, there was a decrease of UMI counts when the input DNA was reduced due to the correlation of UMIs detected with the original DNA copies. Total read depth on the other hand also shows a tendency to decrease as the DNA input is reduced, although this effect is not so perceptible due to the normalisation step, where all samples are quantified and pooled at equimolar concentrations. It is noteworthy that the 125 pg sample on the Ion S5 has a higher UMI coverage than the 250 pg, which could be explained by the higher depth of coverage from the sample that, in turn, could derive from an underestimation in the quantification.

Fig. 2C illustrates the proportion of loci that were concordant with the consensus 1000 pg profile from the three platforms. Both the Ion S5 and MiSeq presented high levels of sensitivity up to 125 pg, with concordance levels of ~95% or higher. At lower DNA inputs levels, the Ion S5 exhibited a higher sensitivity, with ~50% correctly called genotypes at 15 pg input and ~30% with 7 pg of DNA. In comparison, the MiSeq the reduction in sequence data with DNA input observed is more acute, with ~20 % correctly called genotypes at 31 pg input and less than 10% concordant calls with 15 pg. The GeneReader performed the least well with lower levels of DNA, displaying lower concordance rates than the other two platforms, with > 90% concordance at 500 pg input DNA, but less than 50% correctly called genotypes at 125 pg.

3.4. Mixture detection

Although tri-allelic SNPs are not as proficient as STRs or microhaplotypes for mixture detection, when compared to bi-allelic SNPs, they enhance the detection of mixtures, as not only are their allele read frequency ratios imbalanced, but more importantly, in the MPSplex panel which compiled tri-allelic SNPs with the maximum heterozygosity levels, a significant proportion of loci in mixed DNA will show sequence data consisting of three distinct alleles.

For the assessment of the mixture detection capabilities of MPSplex, the sequence patterns of SRM 2391d component D were analysed in detail. Fig. 3 summarises the percentage of loci where three alleles were detected, the proportion of alleles from the major and minor component detected in the mixture, and the major allele read frequency distribution of the single-source samples and the mixture for each of the MPS platforms. Clear signals of mixed DNA were observed in all platforms, with three alleles detected in ~18-21% of tri-allelic SNPs, and more than 90% of the genotypes of the minor contributor were obtained. Additionally, MAF patterns were altered compared to single-source samples, with many SNPs showing MAF ratios beyond their expected ranges observed in unadmixed DNA analyses.

In addition, even though we did not formally assess the capability of MPSplex for mixture deconvolution, it should be possible to retrieve partial profiles from the minor contributor based on skewed allele read frequency ratios from loci where three alleles are detected. However, factors such as high misincorporation for some SNPs or the possible effect of using UMI reads must be considered if this approach is applied further.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.5. Proof-of-concept kinship analysis

A kinship case in which routine STR kits (Globalfiler™ and Investigator HDplex) did not provide enough statistical power for its resolution consisted of an indirect paternity test in which the paternity was analysed using two putative half-sisters. The hypotheses tested are detailed in Fig. 4A.

A total of 1061, 1124 and 1153 genotype calls out of 1235 autosomal tri-allelic SNPs were obtained for the CH1 and suggested sibships CH2 and CH3, respectively. It should be noted that aiSNPs, tri-allelic SNPs from the X chromosome and the microhaplotype composite SNPs were not used for the analyses. Additionally, the 6 autosomal tri-allelic SNPs identified as problematic for the Ion S5 (rs2009433, rs399737, rs58318997, rs80266372, rs319305 and rs6764984, as detailed in section 3.2) were also disregarded.

Fig. 4B shows the probability distributions from the simulations performed with FamLink for both relationship hypotheses, using all the markers in the panel and with profiles containing calls for 70, 80 and 90% of the SNPs. Due to the large number of markers present in the panel and the nature of the pedigrees assessed, several simulations gave genetic incompatibilities (therefore, failing to give an LR value), when the profiles were generated assuming H2 to be true, ~5,900 for the 10,000 simulations assuming a full profile. In other words, for the profiles simulated for CH1, CH2 and CH3 assuming H2, with CH2-CH3 being full siblings and CH1 as non-related to CH2 and CH3; some markers presented genetic inconsistences that could not be explained if H1 were true (e.g., CH2 and CH3 are homozygotes for alleles 2 and 3, respectively, and CH1 is homozygous for allele 1).

The LR value obtained from FamLink was 1.339e+044, around 42 orders of magnitude greater than the results obtained for STRs (an LR of 89.17) and falling within the expected LR distribution for the suggested relationship. Nonetheless, it was also observed that the LR value was positioned in the lower end of the LR distribution when compared to the expected value from simulation data using the whole set of SNPs, as the case samples did not result in genotypes for all MPSplex SNPs, while the simulations assumed complete SNP profiles. A total of 80.16% of the single-site autosomal tri-allelic SNPs in MPSplex overlapped between CH1 and CH2. While CH1 and CH3 shared calls for 82.10% of the markers and the three (CH1-3) shared 76.84%. Therefore, simulations using partial profiles provided a better adjustment to the LR value obtained.

4. Discussion

From our results and experience of running multiple MPS chemistries and sequence detection systems we observed that the MPSplex assay developed with QIAseq technology gives robust and reliable performance as a forensic SNP genotyping assay, and at a larger genotyping scale than most such assays. Each of the three MPS systems evaluated had individual beneficial aspects – e.g., the Ion S5 displayed higher sensitivity, the MiSeq had lower base misincorporation values, and the GeneReader had fewer non-specific sequence reads. Despite these differences in sequencing performance, overall, more than 99% of the SNPs had sequence coverage levels above the calling threshold. High levels of sensitivity were observed with 250 pg of input DNA and as low as 31-62 pg on the MiSeq and Ion S5 platforms. Mixture detection, albeit evaluated with a limited control DNA experiment, was also successful in each platform, although this is, in part, due to the detection of three alleles for a sizeable proportion of the tri-allelic component SNPs of MPSplex.

Despite having high genotype concordance values (≥ 99.5% concordant results), some SNPs were identified as problematic, in many cases because primers are targeting SNPs previously not identified to be sited in duplicated regions. SNPs rs58318997 and rs80266372 were found to have three alleles in some samples; and rs2009433 and rs399737 were found to depart from mendelian inheritance patterns, so these four SNPs should be excluded from further analyses using the current MPSplex primer sets. In terms of individual marker performance, SNP rs4991456 on the MiSeq and GeneReader; and SNPs rs9933604, rs2184289 and rs2843150 on the Ion S5 were

found to have high levels of base misincorporation, which would require manual scrutiny of sequence output using IGV, and thus involving a labour-intensive process. SNPs rs319305 and rs6764984 also produced erroneous calls on the Ion S5 due to the presence of allele drop-ins, while rs7623855 on the MiSeq and rs1809429 on the MiSeq and GeneReader had miscalls due to allele drop-out. Supplementary Table S4 summarises all the underperforming SNPs identified in our study along with details of the issues that were identified. In addition, IGV screenshots for the markers for which alignment issues were observed are included in Supplementary File S1.

The kinship study we performed served as a simple proof-of-concept of the potential of MPSplex, and demonstrated its applicability to relationship tests where routine STR and mtDNA analyses cannot provide sufficient power of discrimination. Although in the current field, the Kintelligence and FORCE forensic panels have now been developed, they are focussed on genealogical applications and the information they provide may exceed what would be required in many relationship testing scenarios. These panels also require the allocation of extra capacity in the sequencing run, increasing the cost per sample. Therefore, MPSplex represents a simpler, but highly informative “middle ground” solution for cases in which the application of MPS technologies is required but the degree of relatedness between missing persons and relatives is less distant.

In comparison to other smaller scale MPS forensic panels, MPSplex displayed comparable levels of performance. The error rate was low when considering the whole multiplex, shown by the median misincorporation value of 0 obtained for the three platforms and the high levels of genotype concordance. Some SNPs nonetheless had high levels of base misincorporation, especially for the Ion S5 platform. Although the number of markers showing high levels of misincorporation was higher than other MPS panels, this factor should be considered in the light of the bigger size of the MPSplex multiplex. Previous studies also report a similar behaviour using this platform, on many occasions due to the presence of homopolymeric tracts target SNP site that hinder correct alignment [2, 26]. In terms of sensitivity, we observed similar or less sensitive outputs when compared to lower-scale SNP multiplexes on the Ion S5 and MiSeq [9, 13, 26, 27]. Bigger multiplexes on the other hand showed reduced sensitivity with less complete results in terms of percentage of markers genotypes, although the total number of markers genotyped was higher in some instances due to the larger number of SNPs present in those panels [14, 29].

Although we did not formally assess the improvement of sequencing results obtained when the UMI information is considered, the studies of Woerner et al., 2023, Crysup et al., 2023, and Staadig et al., 2021 [14-16] indicate that this technology increases the accuracy of the results. Nevertheless, several factors should be taken into consideration when working with UMI-based sequencing technology in a forensic context. First, UMI technology was previously developed for non-forensic applications, where neither DNA concentration nor DNA integrity are a problem. Therefore, the UMI workflow contains several steps that may pose a problem when dealing with forensic samples. A total of four magnetic bead clean-ups are required for library preparation, which is likely to increase the risk of target DNA loss. We note that the recently introduced QIAseq PRO kit could potentially alleviate this issue, as it replaces the bead clean-ups after adapter ligation and target enrichment with enzymatic clean-ups. In addition, a fragmentation step is performed at the beginning of the protocol, and this procedure could present issues for the analysis of degraded or low-level DNA. Although the protocol contains an adaptation for FFPE DNA with reduced fragmentation time and Staadig’s study obtained promising results with a UMI panel used to analyse bone samples, the applicability of UMI-based MPS for this type of sample needs further evaluation of the effects of variation in fragmentation time by sequencing a sufficiently broad-based collection of highly degraded samples.

Another issue that is likely to arise with UMI technology in forensic applications is the increased complexity of analysis demanded for the interpretation of sequencing data, as not only coverage needs to be considered, but also the UMI count for each locus and the number of reads for each UMI family. Additionally, prior information on the likely quantity and quality of the DNA is required

to correctly pool each library. Otherwise, as evident in the samples amplified with ≥ 10 ng of DNA, each UMI family could be under-represented and consequently that will affect the quality of the sequencing output.

Despite the above reservations, our findings indicate that the use of MPSplex for missing persons analysis based on UMI technology will fulfil the need for a medium-scale marker panel that can provide informative likelihoods in challenging kinship tests where relationships are either too distant for conventional STR data to resolve, or the pedigree of surviving relatives is too incomplete and requires a much higher number of markers. At the time of writing, Qiagen are withdrawing their GeneReader technology, but our study shows the workflows of MiSeq and Ion S5 MPS platforms can be readily adapted to perform UMI-based sequencing with the expectation that a large proportion of complex kinship tests can be resolved when use of STRs alone fails to provide high enough likelihoods.

References

[1] N. Hiroaki, F. Koji, K. Tetsushi, S. Kazumasa, N. Hiroaki, S. Kazuyuki, Approaches for identifying multiple-SNP haplotype blocks for use in human identification, Legal Medicine 17(5) (2015) 415-420.

[2] M. de la Puente, C. Phillips, C. Xavier, J. Amigo, A. Carracedo, W. Parson, M.V. Lareu, 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 (2020) 102213.

[3] F. Oldoni, D. Bader, C. Fantinato, S.C. Wootton, R. Lagace, K.K. Kidd, D. Podini, A sequence-based 74plex microhaplotype assay for analysis of forensic DNA mixtures, Forensic Sci Int Genet 49 (2020) 102367.

[4] K. Breslin, B. Wills, A. Ralf, M. Ventayol Garcia, M. Kukla-Bartoszek, E. Pospiech, A. Freire-Aradas, C. Xavier, S. Ingold, M. de La Puente, K.J. van der Gaag, N. Herrick, C. Haas, W. Parson, C. Phillips, T. Sijen, W. Branicki, S. Walsh, M. Kayser, HIrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms, Forensic Sci Int Genet 43 (2019) 102152.

[5] M. Diepenbroek, B. Bayer, K. Schwender, R. Schiller, J. Lim, R. Lagace, K. Anslinger, Evaluation of the Ion AmpliSeq PhenoTrivium Panel: MPS-Based Assay for Ancestry and Phenotype Predictions Challenged by Casework Samples, Genes (Basel) 11(12) (2020).

[6] J. Ruiz-Ramirez, M. de la Puente, C. Xavier, A. Ambroa-Conde, J. Alvarez-Dios, A. Freire-Aradas, A. Mosquera-Miguel, A. Ralf, C. Amory, M.A. Katsara, T. Khellaf, M. Nothnagel, E.Y.Y. Cheung, T.E. Gross, P.M. Schneider, J. Uacyisrael, S. Oliveira, M.D.N. Klautau-Guimaraes, C. Carvalho-Gontijo, E. Pospiech, W. Branicki, W. Parson, M. Kayser, A. Carracedo, M.V. Lareu, C. Phillips, V. Consortium, Development and evaluations of the ancestry informative markers of the VISAGE Enhanced Tool for Appearance and Ancestry, Forensic Sci Int Genet 64 (2023) 102853.

[7] M. Kayser, W. Branicki, W. Parson, C. Phillips, Recent advances in Forensic DNA Phenotyping of appearance, ancestry and age, Forensic Sci Int Genet 65 (2023) 102870.

[8] I. Grandell, R. Samara, A.O. Tillmar, A SNP panel for identity and kinship testing using massive parallel sequencing, Int J Legal Med 130(4) (2016) 905-914.

[9] M. de la Puente, C. Phillips, C. Santos, M. Fondevila, A. Carracedo, M.V. Lareu, Evaluation of the Qiagen 140-SNP forensic identification multiplex for massively parallel sequencing, Forensic Sci Int Genet 28 (2017) 35-43.

[10] A. Tillmar, K. Sturk-Andreaggi, J. Daniels-Higginbotham, J.T. Thomas, C. Marshall, The FORCE Panel: An All-in-One SNP Marker Set for Confirming Investigative Genetic Genealogy Leads and for General Forensic Applications, Genes (Basel) 12(12) (2021).

[11] C. Phillips, J. Amigo, A.O. Tillmar, M.A. Peck, M. de la Puente, J. Ruiz-Ramirez, F. Bittner, S. Idrizbegovic, Y. Wang, T.J. Parsons, M.V. Lareu, A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel, Forensic Sci Int Genet 46 (2020) 102232.
[12] K.K. Kidd, W.C. Speed, A.J. Pakstis, M.R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F.R. Friedlaender, J.R. Kidd, Progress toward an efficient panel of SNPs for ancestry inference, Forensic Sci Int Genet 10 (2014) 23-32.
[13] D. Truelsen, A. Freire-Aradas, M. Nazari, A. Aliferi, D. Ballard, C. Phillips, N. Morling, V. Pereira, C. Borsting, Evaluation of a custom QIAseq targeted DNA panel with 164 ancestry informative markers sequenced with the Illumina MiSeq, Sci Rep 11(1) (2021) 21040.
[14] A. Staadig, J. Hedman, A. Tillmar, Applying Unique Molecular Indices with an Extensive All-in-One Forensic SNP Panel for Improved Genotype Accuracy and Sensitivity, Genes (Basel) 14(4) (2023).
[15] B. Crysup, S. Mandape, J.L. King, M. Muenzler, K.B. Kapema, A.E. Woerner, Using unique molecular identifiers to improve allele calling in low-template mixtures, Forensic Sci Int Genet 63 (2023) 102807.
[16] A.E. Woerner, S. Mandape, J.L. King, M. Muenzler, B. Crysup, B. Budowle, Reducing noise and stutter in short tandem repeat loci with unique molecular identifiers, Forensic Sci Int Genet 51 (2021) 102459.
[17] K.K. Kidd, W.C. Speed, A.J. Pakstis, D.S. Podini, R. Lagace, J. Chang, S. Wootton, E. Haigh, U. Soundararajan, Evaluating 130 microhaplotypes across a global set of 83 populations, Forensic Sci Int Genet 29 (2017) 29-37.
[18] C. Phillips, D. McNevin, K.K. Kidd, R. Lagacé, S. Wootton, M. de la Puente, A. Freire-Aradas, A. Mosquera-Miguel, M. Eduardoff, T. Gross, L. Dagostino, D. Power, S. Olson, M. Hashiyada, C. Oz, W. Parson, P.M. Schneider, M.V. Lareu, R. Daniel, MAPlex - A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations, Forensic Science International: Genetics 42 (2019) 213-226.
[19] M. Byrska-Bishop, U.S. Evani, X. Zhao, A.O. Basile, H.J. Abel, A.A. Regier, A. Corvelo, W.E. Clarke, R. Musunuri, K. Nagulapalli, S. Fairley, A. Runnels, L. Winterkorn, E. Lowy, C. Human Genome Structural Variation, F. Paul, S. Germer, H. Brand, I.M. Hall, M.E. Talkowski, G. Narzisi, M.C. Zody, High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios, Cell 185(18) (2022) 3426-3440 e19.
[20] J. Wagner, N.D. Olson, L. Harris, Z. Khan, J. Farek, M. Mahmoud, A. Stankovic, V. Kovacevic, B. Yoo, N. Miller, J.A. Rosenfeld, B. Ni, S. Zarate, M. Kirsche, S. Aganezov, M.C. Schatz, G. Narzisi, M. Byrska-Bishop, W. Clarke, U.S. Evani, C. Markello, K. Shafin, X. Zhou, A. Sidow, V. Bansal, P. Ebert, T. Marschall, P. Lansdorp, V. Hanlon, C.A. Mattsson, A.M. Barrio, I.T. Fiddes, C. Xiao, A. Fungtammasan, C.S. Chin, A.M. Wenger, W.J. Rowell, F.J. Sedlazeck, A. Carroll, M. Salit, J.M. Zook, Benchmarking challenging small variants with linked and long reads, Cell Genom 2(5) (2022).
[21] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2023.
[22] Bam-readcount software at: https://github.com/genome/bam-readcount.).
[23] D. Kling, T. Egeland, A.O. Tillmar, FamLink--a user friendly software for linkage calculations in family genetics, Forensic Sci Int Genet 6(5) (2012) 616-20.
[24] A.O. Tillmar, C. Phillips, Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets, Forensic Science International: Genetics 26 (2017) 58-65.
[25] T. Egeland, P.F. Mostad, B. Olaisen, A computerised method for calculating the probability of pedigrees from genetic data, Sci Justice 37(4) (1997) 269-74.
[26] C. Xavier, M. de la Puente, A. Mosquera-Miguel, A. Freire-Aradas, V. Kalamara, A. Ralf, A. Revoir, T.E. Gross, P.M. Schneider, C. Ames, C. Hohoff, C. Phillips, M. Kayser, W. Parson, V.

Consortium, Development and inter-laboratory evaluation of the VISAGE Enhanced Tool for Appearance and Ancestry inference from DNA, Forensic Sci Int Genet 61 (2022) 102779.
[27] C. Xavier, M. de la Puente, M. Sidstedt, K. Junker, A. Minawi, M. Unterlander, Y. Chantrel, F.X. Laurent, A. Delest, C. Hohoff, I. Bastisch, J. Hedman, K.J. van der Gaag, T. Sijen, W. Parson, Evaluation of the VISAGE basic tool for appearance and ancestry inference using ForenSeq(R) chemistry on the MiSeq FGx(R) system, Forensic Sci Int Genet 58 (2022) 102675.
[28] M. Eduardoff, T.E. Gross, C. Santos, M. de la Puente, D. Ballard, C. Strobl, C. Borsting, N. Morling, L. Fusco, C. Hussing, B. Egyed, L. Souto, J. Uacyisrael, D. Syndercombe Court, A. Carracedo, M.V. Lareu, P.M. Schneider, W. Parson, C. Phillips, E.U.-N. Consortium, W. Parson, C. Phillips, Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM, Forensic Sci Int Genet 23 (2016) 178-189.
[29] J. Antunes, P. Walichiewicz, E. Forouzmand, R. Barta, M. Didier, Y. Han, J.C. Perez, J. Snedecor, C. Zlatkov, G. Padmabandu, L. Devesse, S. Radecke, C.L. Holt, S.A. Kumar, B. Budowle, K.M. Stephens, Developmental validation of the ForenSeq(R) Kintelligence kit, MiSeq FGx(R) sequencing system and ForenSeq Universal Analysis Software, Forensic Sci Int Genet 71 (2024) 103055.

Figures

Fig. 1A. Violin plot of the distribution of UMI counts for reference samples amplified using 1-2 ng DNA input. Maroon dashed line represents the 10 UMI calling threshold. N = 9, 3 and 10 for the Ion S5, MiSeq and GeneReader, respectively. **Fig 1B.** Raster plot with UMI coverage for each marker for the reference sample set. Markers in grey represent loci with a UMI count lower than the calling threshold of 10 UMIs.

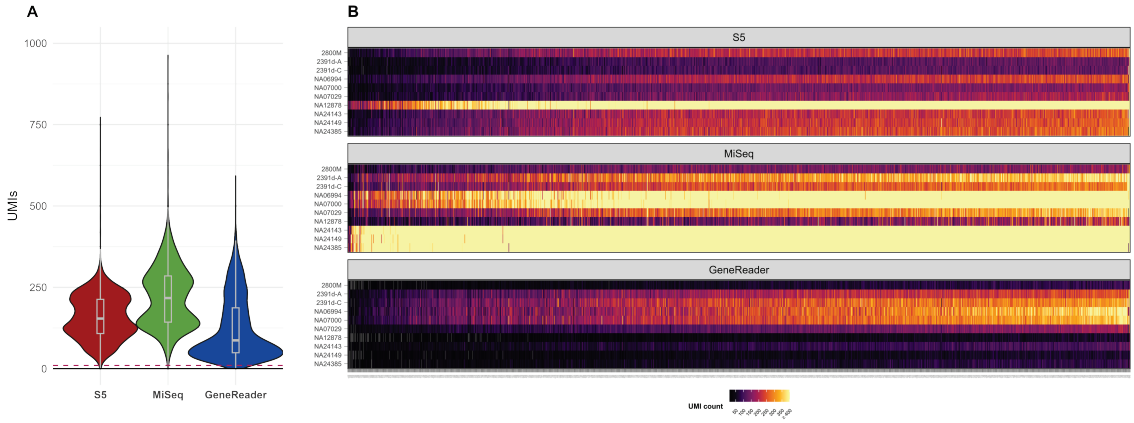
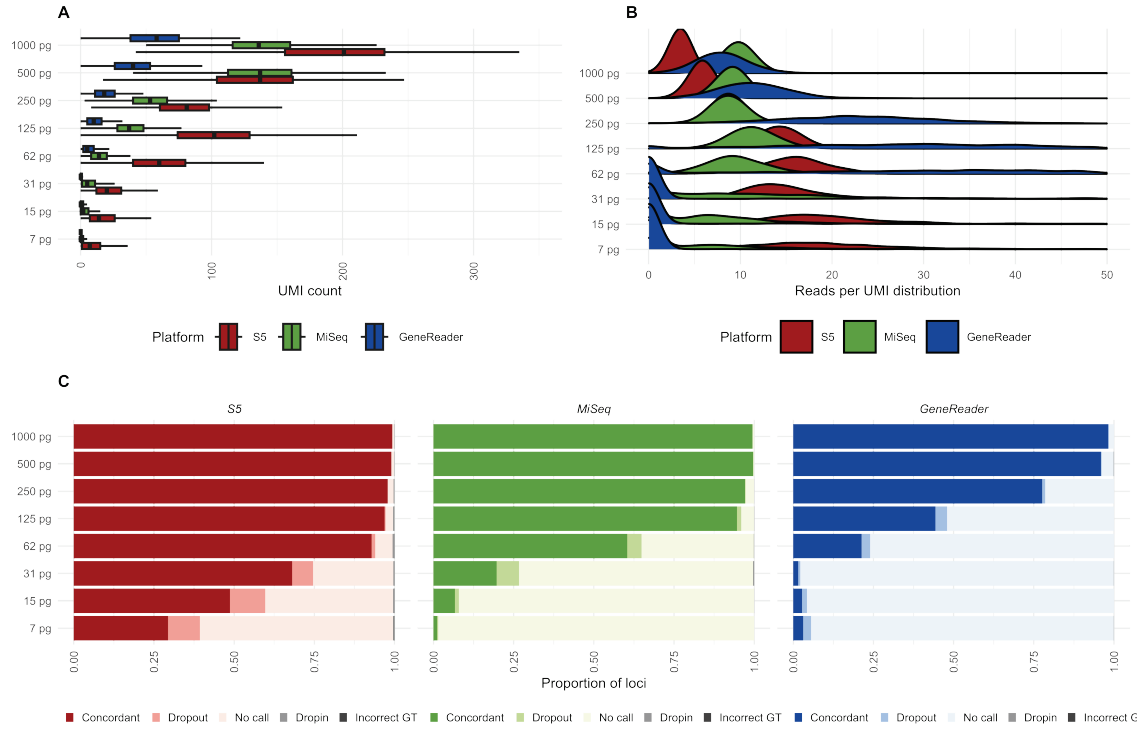


Fig. 2 Overall performance of the 2800M dilution series in each of the three platforms**. Fig. 2A**. Boxplots of UMI coverage distribution per locus. **Fig. 2B** Ridgeline plot of number of reads per UMI read distribution. **Fig. 2C** Proportion of concordant genotypes obtained for each sample compared to the consensus 1000 pg reference profile. Incorrect GT: incorrect genotype.



USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Fig. 3. Mixture analysis with MPSplex single-site autosomal tri-allelic SNPs on each sequencing platform. Upper left bar charts show percentage of heterozygous (HT), homozygous (HM), three-allele and no genotype calls for the two single-source samples and the mixture from the SRM 2391d control DNA set. Upper right circular plots show proportion of alleles called in the mixture for the major (2391d-A) and the minor (2391d-C) components. Bottom scatter plots show the major allele frequency for the SNPs analysed in three sample panes. Red dashed lines represent the optimum 50-60% and >90% intervals for heterozygous and homozygous genotype calls, respectively.

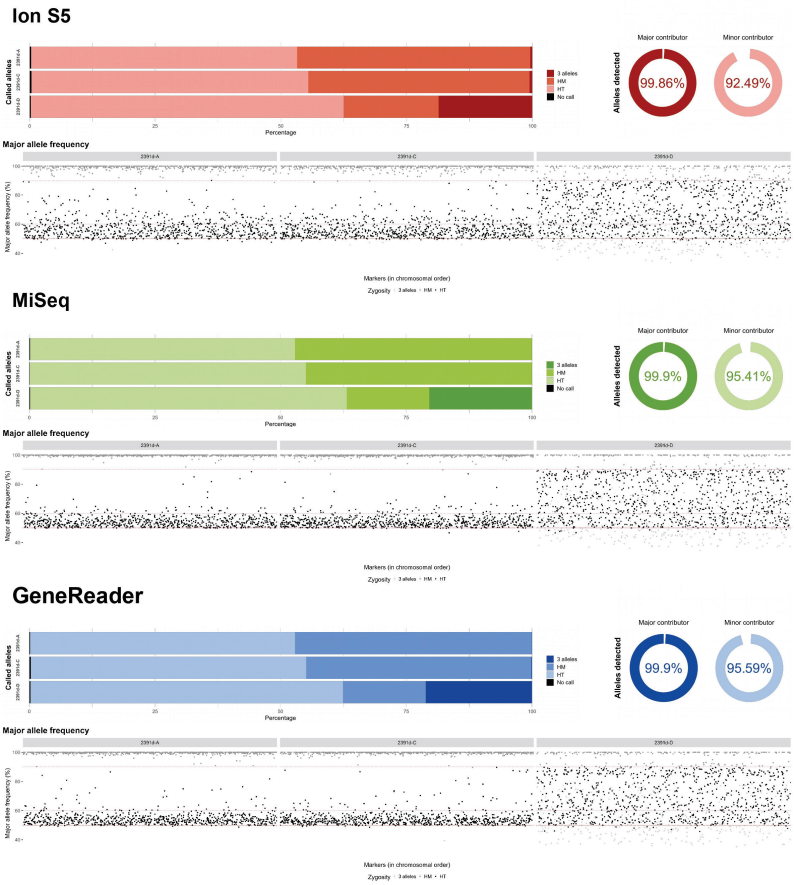
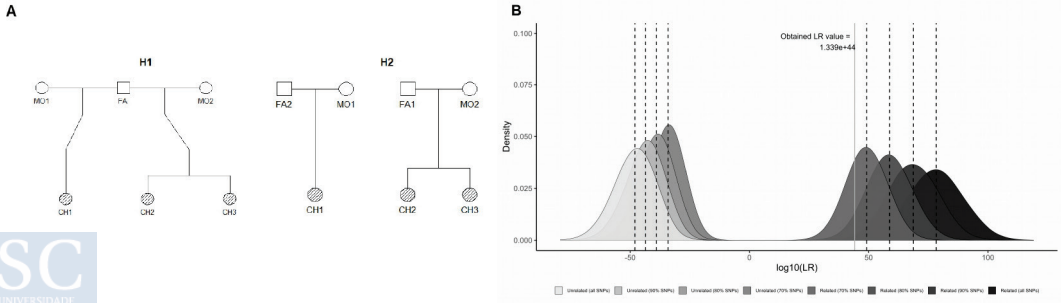


Fig. 4A. Pedigrees of the two hypotheses assessed in the kinship study, comprising: CH1 is related to CH2 and CH3 (H1) vs CH1 unrelated (H2), with shaded symbols representing individuals that were genotyped. **Fig. 4B**. Log LR distributions for the simulations performed with FamLink. Log LR distributions have graded greyscale shading to represent different degrees of SNP genotype data completeness of 70%, 80%, 90% and 100% used in the simulations. Black dashed lines represent mean LR values for the simulations and the grey line represents the LR value obtained in support of H1 with MPSplex.



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Acknowledgements

J.R. was supported by the “Programa de axudas á etapa predoutoral” funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481A-2020/039). MVL is supported by grant PID2019–107876RB-I00 funded by the Ministerio de Educación, Cultura y Ciencia, Spain (MCIN/AEI/10.13039/501100011033) and grant PID2022-141224OB-I00 funded by MCIN/AEI/10.13039/501100011033 and “ERDF A way of making Europe”. MdlP is supported by grant IJC2020–042638-I funded by the Gobierno de España MCIN/AEI/10.13039/501100011033 and the European Union “NextGenerationEU/PRTR”. Data analysis was partially undertaken in the FinisTerrae III supercomputer from the Centro de Supercomputación de Galicia (CESGA). We would like to thank Keith Elliot and Holger Karas from Qiagen for their support and technical assistance. A special thank you goes to the ISFG board for supporting J.R. with the Short Time Fellowships program allowing this study to be accomplished at ICMP, The Hague, Netherlands.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5. DISCUSSION

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5. DISCUSSION

In the last decade, the interest around MPS and its forensic applications has grown exponentially, in part owing to the large multiplexing capacity that these instruments offer, allowing the simultaneous analysis of thousands of markers within a single reaction; and also due to their versatility, allowing the accommodation of different types of markers that can even be analysed together, such as STRs, SNPs and InDels – a possibility which was not feasible with CE methodologies. This, in turn, has permitted development of a wide variety of tools for applications such as human identification or the inference of externally visible characteristics (EVCs), age or ancestry (BGA) and even multiplexes combining markers for several of these purposes. The results of this Doctoral Thesis compile six different studies where five different MPS panels for forensic applications are studied: the VISAGE Basic Tool for Appearance and Ancestry prediction (herein VISAGE BT or BT), the VISAGE Enhanced Tool for Appearance and Ancestry inference (herein VISAGE ET or ET), a 113 MH marker panel, the *Eurasiaplex*-2 panel, and the MPSplex panel.

Overall, the results can be grouped into two different blocs divided by the main application that was explored: the panels in which mainly their application for BGA analyses was studied and the development and validation of the studied panel for kinship testing.

5.1 Assessing panels for biogeographic ancestry inference

Although STRs have proved to be invaluable in a very wide range of forensic identification analyses due to, among other characteristics, their high discrimination power. However, the information that they can provide in scenarios where there is no other profile available to compare can be quite limited. In terms of ancestry inference, the application of routine STR kits was explored in the studies of Lowe et al., Phillips et al., and Pereira et al. This potential use was appealing as extra information could be extracted from routine analyses without having to use more DNA sample, no additional techniques would be required to set up in the laboratory and could be useful even in instances of mixed DNA profiles. All the above studies achieved, to a certain extent, the differentiation between some of the main population groups, although overall they all lacked enough power to differentiate all populations successfully and produced a high misclassification rate (Lowe et al., 2001; Pereira et al., 2011; Phillips et al., 2011). The work by Rosenberg et al. accomplished the differentiation of the main worldwide genetic clusters by using 377 highly informative tetra-nucleotide, tri-nucleotide and di-nucleotide STRs (Rosenberg et al., 2002). Although dimeric STRs can discriminate more effectively among population groups than tetrameric STRs, these type of STRs are not as suitable for forensic

USC UNIVERSIDADE DE SANTIAGO

analyses. Further studies by Londin et al., Phillips et al. explored the creation of panels that incorporated STRs specifically selected for this purpose (Londin et al., 2010; Phillips et al., 2013a) but in order to achieve satisfactory results either relied on dinucleotide STRs or supplemented the STR information with additional markers, such as SNPs.

Dedicated SNP panels on the other hand can provide accurate estimations even when small sets of these markers are used, which can be typed through SBE and CE, and successful classification of individuals has been accomplished up to a continental level – sub-Saharan Africans, Europeans, native Americans, Oceanians and East Asians (de la Puente et al., 2016; Phillips et al., 2007). However, the differentiation of other populations which are more closely related, in which there are no big geographic barriers such as oceans or deserts that hinder gene flow between populations, has proved to be more challenging and requires the study of a wider set of markers. Thus, the large number of markers that MPS panels can comprise makes them a suitable tool for the development of more accurate and functional tools to provide a higher degree of detail.

The availability of extensive and comprehensive population data has also, up until recently, posed a challenge in the selection of ancestry informative markers for several populations, including Oceanians, unadmixed native Americans and South Asians/Middle East individuals (Phillips, 2015). But within the last few years, and thanks to the rise of MPS sequencing, forensic DNA analysis has at its disposal a much wider collection of populational samples from these regions, as well as more curated data from already-existing whole-genome sequencing projects (Almarri et al., 2021; Bergström et al., 2020; Byrska-Bishop et al., 2022) which can be used to select more powerful makers.

5.1.1 Development of new panels for ancestry inference

The VISAGE BT and ET marker sets are two panels that were designed within the VISAGE framework pursuing the same goal: providing an all-in-one tool that integrates markers for inferring physical traits and ancestry from an unknown individual, but each of the tools with a different degree of detail. Both panels underwent forensic validation studies (Palencia-Madrid et al., 2020; Xavier et al., 2022; Xavier et al., 2020), in which the performance of the panel was assessed. Instead, the studies described in this Thesis focused on the selection of the markers and their applicability in the performance of BGA prediction analyses.

The selection of markers for the Basic Tool was based on previously established SNP sets, encompassing the 41 HIrisPlex-S markers for the inference of eye, hair and skin colour prediction (Breslin et al., 2019); and 115 SNPs for ancestry inference selected to target 6 population groups (AFR, AME, SAS, EAS, EUR and OCE). These aiSNPs included 15 tri-allelic SNPs collected from Phillips et al. (2020) together with 100 bi-allelic SNPs selected from more than 700 autosomal AIMs from other well-established SNP panels: (i) the Global

USC DE SANTIAGO DE COMPOSTELA

AIMs panel (Eduardoff et al., 2016), (ii) the LACE panel (Galanter et al., 2012), (iii) the Kiddlab ancestry panel (Kidd et al., 2014), (iv) the Thermo Fisher's Precision ID Identity panel, (iv) the EUROFORGEN NAME panel (Pereira et al., 2019), (vi) the US admixture mapping panel from Shriver's group (Mao et al., 2007) and (vii) the *Eurasiaplex* panel (Phillips et al., 2013b).

The VISAGE ET panel, in turn, combined an extended number of markers for both applications, targeting the differentiation of an extended range of traits – adding eyebrow colour, freckles, hair shape and male baldness patterns – while the ancestry markers aimed for the differentiation of Middle Eastern and North African populations. Instead of relying on the previous panel design and increasing the number for markers for the new target populations, the panel was created anew, with an extensively revised selection of autosomal aiSNPs; while also widening the types of markers, including combining X-SNPs and Y-SNPs, as well as ancestry informative MHs. Candidate loci included once again markers from forensic panels developed for ancestry prediction – NAME (Pereira et al., 2019) *Eurasiaplex* (Phillips et al., 2013b) and *Pacifiplex* panels (Santos et al., 2016) and PIMA (Carvalho Gontijo et al., 2020) panels – plus an increased selection of the tri-allelic SNPs; as well as novel markers selected from the 1000 Genomes Phase III data. Despite the overall increase in the total number of ancestry markers when compared to the BT panel, the bi-allelic autosomal SNPs were reduced in favour of the additional types of markers, not only maintaining the informativeness in BT, but increasing the number of population groups that could be differentiated. This was achieved through a careful adjustment of the number of markers targeting each population, on the one hand, reducing the number of markers for AFR and OCE, as many SNPs selected for other populations were already highly informative for these two non-targeted populations; and on the other hand, selecting more powerful markers and expanding the markers targeting the differentiation of more closely related populations (like SAS and EUR). In both BT and ET panel designs, the selection of markers sought to balance the cumulative Population Specific Divergence so as to avoid strong differentiation bias towards specific populations (see Figure 1 from Article 1 as example).

The exploration of the classification for the autosomal marker sets in both panels achieved the separation of each of the targeted populations. With the BT data, several classification methods for the classification of the 6 target population groups were explored, including STRUCTURE, MDS, neighbour-joining trees (NJT); while also performing a one-out cross-validation using *Snipper*. Conversely, the analysis with the VISAGE ET focused mainly on the STRUCTURE analyses, as this had proved to provide extra detail on co-ancestry (plus, this method accepts the combination of different types of markers). Unlike BT, a two-stage STRUCTURE analysis was performed, comprising a continental analysis tier with the 5 main continental groups (excluding North African, SAS and ME populations in the reference data), and a secondary run performed dependent on the results of the first run, for Eurasian population groups separation (including North African, SAS and ME data). This method allowed for the distinction of Middle Eastern populations, which also produced a consistent separation of a

USC DE SANTIAGO DE COMPOSTELA

cluster for North African populations. However, the lack of a larger collection of reference data for North Africa, which is currently composed solely of CEPH Algerians from a single location, hinders the accurate assessment of whether the separation observed could be attributed to variation exclusive to this population or could be generalised to the whole region – although the variation cline observed in the study Moroccans may indicate the first alternative.

The ancestry predictive capacity of the panels was further underlined by the analysis of admixed samples, comprising individuals with co-ancestral backgrounds. Co-ancestry proportions obtained from STRUCTURE were benchmarked using the estimates obtained with much larger SNP sets (i.e., the Human Origins array for BT and the 1000 Genomes genome-wide SNP data estimates for ET). In general, these studies corroborated the good performance of the panels – overall high correlation values of co-ancestry proportion estimates were obtained, which proved to be more precise when the major component was predominant (> 90%), but populations with individuals presenting three-way admixture patterns proved to be more challenging for these SNPs sets, as correlations were less precise in such instances. Conversely, GDA analyses in BT failed to provide good correlations in admixed samples, showing a tendency to underestimate EUR co-ancestry. Additionally, some of the study populations indicated the importance of careful interpretation of admixed individuals and how the reference set must be adapted in accordance with the results in each, such as the non-detection of the SAS component from Indo-Fijian in the first STRUCTURE run due to the absence of SAS reference populations. It should be noted that additional testing with the VISAGE ET data also sought to improve the quality of the results with the inclusion of the EVC markers, but the addition of extra markers did not yield more informative analyses, but it rather decreased the quality of the results. Regarding the MH markers, the reduced number of these markers in the multiples could not provide the same level of resolution as the autosomal markers. Yet, initial explorations of ancestry inference in mixtures provided good results of the applicability of the panel using a simple 3-population classification of AFR, EUR and EAS in all instances, despite access to partial profiles resulting from mixture deconvolution.

The inclusion of non-autosomal markers was one of the main differences of the Enhanced Tool from the Basic Tool, which only contained autosomal aiSNPs. The use of this type of marker for ancestry inference has more restricted use when no co-ancestry patterns are detected, but they can provide more detailed insight into the individual's ancestry when there are indications of admixture from the autosomal AIMs. Y-SNP analysis can sometimes indicate a more restricted geographic distribution of the variation in a smaller region of the broader extent of continental population groups (e.g. Eastern Europe often has distinct Y-SNP distributions compared to the north and west of Europe). The Y-SNPs included in the multiplex, included for patrilinear ancestry inference, were a selection of one of the most comprehensive compilations of these markers for forensic purposes (Ralf et al., 2019), balancing the number of markers incorporated into the ET multiplex while trying to provide the maximum informativeness. The X-SNPs on the other hand represented the counterpart of the Y-SNPs to evaluate the maternal lineage. The exploration of maternal ancestry with X chromosome markers is mainly

USC
DE SANTIAGO
DE COMPOSTELA

disregarded as it has a complex pattern of heredity, while mtDNA analysis, though more complex to sequence, is well established and can provide much easier-to-interpret data. The X-AIMs nonetheless constitute an alternative to mtDNA as they can be easily co-amplified with autosomal markers, whereas the high number of copies of the mitochondrial genome prevents its analysis in a single-library multiplex. Two methods of analysis were explored for the 16 X-SNPs selected for the ET assay: using the whole set of SNPs taking into account the linkage between them, and the use of 5 SNPs in the centromere region treated as a haplotype block that is generally well-preserved form the root co-ancestral populations in admixed individuals, due to almost no recombination activity around the X centromere.

The use of these markers was explored in admixed populations, including the African Americans from SW USA and African Caribbeans from Barbados, from 1000 Genomes, plus three in-house study populations – Brazilians from Kalunga, Brazilians from Brasilia and Fijians. The Brazilian population is a highly diverse population resulting from a complex process of admixture, involving American, European and African components from the indigenous aboriginal population, the European colonizers and the trade of African slaves respectively. The Kalunga are an example of Quilombos, isolated communities of African slaves that were made to act as refuges to escape from enslavement, and to this day, still holding a strong African component (Gontijo et al., 2018), while the Brasilia urban individuals represent a more diverse population in which there is a much higher influence of the European immigrants. The Fijians also constitute a complex case of an admixed population with Melanesian and Polynesian individuals settling the island group during the expansion of the latter into Remote Oceania ~3500 years ago (Wollstein et al., 2010), with later settlement by South Asians in some parts of Fiji.

In the same fashion as the MH markers, the X-SNPs were not as proficient as the autosomal AIMs to distinguish 7-8 population groups, but the analysis of 5 populations yielded successful results using both the whole set of SNPs through PCA analysis, and the study of the 5-SNP centromere haplotypes. Although the centromeric SNPs showed strong linkage among the markers, the physical distance between them prevents the possibility to obtain the phase of the SNPs with the current sequencing technologies, so the configuration of the haplotype block in females needs to be inferred. Nonetheless, in the examination of these markers in male individuals from the 1000 Genomes and HGDP-CEPH data population-specific haplotypes were found, which accounted for a large proportion of the haplotypes present in each of them. The haplotypes found in the admixed populations further confirmed the presence of these haplotypes from the co-ancestral root populations and suggested that, despite the physical distance and time since the admixture event of these populations, the haplotypes had not been disrupted by recombination. Thus, these findings suggested that the analysis of the haplotype block could be a suitable tool to analyse the admixture on the X chromosome even when there is no indication of when the admixture event may have occurred. Finally, thanks to the study of the ancestry from the X chromosome in conjunction with the Y-SNP data, it proved possible to recreate the admixture patterns of the African American and admixed American populations

USC DE SANTIAGO DE COMPOSTELA

from 1000 Genomes and the Brazilian Kalungas, as the uneven ancestry distribution of variation in the X-SNPs and Y-SNPs in males gave indications of sex bias during admixture events, findings which had been previously reported for these populations (Gontijo et al., 2018; Lind et al., 2007).

In the current field, there are several other available tools which have been developed for the inference of bio-geographic ancestry. Both VISAGE tools discussed above constitute some of the most advanced tools in the forensic community, with ET representing the most complete forensic panel to date and the first of this kind to allow the simultaneous analysis of X-AIMs. The ForenSeq DNA Signature Prep Kit from Verogen (most specifically, primer mix B) or the PhenoTrivium panel (Diepenbroek et al., 2020) are two examples of "all-in-one" panels targeting different applications, including ancestry and ECVs prediction. The Signature Prep Kit has the advantage of combining several types of routine markers (autosomal, X- and Y-STRs) with identity, BGA and phenotypic (EVC) SNPs. The selection of AIMs nonetheless is restricted and encompasses the 55 aiSNPs described in Kidd et al. (2014), allowing just a 5-group differentiation. The PhenoTrivum panel on the other hand constitutes one of the most useful multiplexes with the ET. It includes the HIrisPlex-S markers, 120 Y-SNPs and a total of 163 autosomal AIMs, a 50% increase from the ET AIM SNP pool (most of those in PhenoTrivium shared with the Precision ID Ancestry Panel from TFS). These SNPs demonstrated a similar capacity to ET to distinguish up to seven population groups, including the five major continental groups plus South Asia and Southwest Asia/Middle East. The results of the analysis of admixed individuals however were not completely successful, in line with previous studies for the same set of markers (Al-Asfi et al., 2018).

Despite the promising results obtained in the analyses, and even if these tools may suffice in most instances, demonstrating outstanding achievements in the field of admixture analysis, they still rely on the use of a small scale set of SNPs, far from the large capacity offered by MPS sequencers. The recent publication of panels suitable for forensic samples comprising thousands of SNPs (e.g., the MPSplex or FORCE panels, further discussed in section 5.2.2) serve as precedent of the possibility to develop ancestry panels of much larger sets of markers, to fully utilise the capabilities of modern-day sequencers and to target the distinction of a wider range of populations and to enhance the analysis of admixed individuals (Kayser et al., 2023).

5.1.2 Exploring the applicability of an already existing panel for BGA prediction

The construction of panels of markers for identification purposes, although constructed to be as informative as possible regardless of the population, are seen to provide, to a certain extent, a degree of population stratification. The 113 MH panel developed and validated by de la Puente and collaborators (de la Puente et al., 2020), which was validated on two different MPS platforms (Ion S5 and MiSeq), was developed as an identification panel with small amplicon sizes containing highly polymorphic markers. As discussed previously, even routine

USC DE COMPOSTELA

STR panels have some applicability in ancestry prediction due to the differences in the allelic frequencies among the main population groups. This constituted the main principle behind this project: exploring the applicability for BGA analysis of the MH panel to further extend its forensic applications.

To gauge the informativeness of the panel and provide haplotype frequency estimations, population data from the 1000 Genomes Phase III (The 1000 Genomes Project Consortium, 2015) and the whole genome sequencing data of 929 HGDP-CEPH samples (Bergström et al., 2020) were collected. Similarly to results observed in the BT study, when it was initiated, the whole-genome sequencing data of the HGDP-CEPH panel had not been released. So, in order to expand the population data of some geographical regions, 246 OCE, AME and ME samples from the HGDP-CEPH Human Genome Diversity Panel (Cann et al., 2002) were sequenced. Although this data was not used as initially intended, we took advantage of the already generated data to re-assess the performance and the concordance between the results obtained from our genotyping based on MPS analysis and the newly generated CEPH sequence-based data.

The overall performance observed for the CEPH sample MPS sequence analysis was very similar to that observed during the validation of the panel. Four MHs that had been previously flagged as underperforming due to low coverage levels, 6pB, 17qC, XpB, and 16pB, presented the same tendency in this study – several samples did not reach the threshold for genotype calling for these markers and imbalanced allele read frequencies were observed for these same systems. Some variation was detected for certain markers that presented strand bias and allele read frequency imbalance that had not been previously observed, whereas several other markers, contrary to the previous study, showed improved allele balance. Nonetheless, the differences obtained in this evaluation, could be due to a higher number of samples from the previous validation.

The concordance rate obtained was higher than 99% when the genotypes from the tested samples were compared to the whole-genome sequencing data from the SGDP and HGDP sample sets. Consistent genotype discordancies with the SGDP data were observed for some loci, which were suspected to be caused by the use of probabilistic phasing software, either because of the software not admitting multi-allelic SNPs or due to incorrect phasing. HGDP data also presented discordancies which could also be accounted for by the phasing algorithm, although the number was significantly lower. Underperforming markers, that presented low coverage or imbalanced reads, were also observed to be the cause of another portion of the discrepancies. These results illustrate the higher accuracy of using MPS for phasing as opposed to using probabilistic software to reconstruct the haplotypes from individual genotyping, while also highlighting the importance of manual curation of underperforming markers to ensure correct genotyping.

USC UNIVERSITAT DE VALÈNCIA

The assessment of the ancestry informativeness of the panel was carried out adopting a nested approach similar to the strategy followed with the VISAGE ET panel: a first stage

comprising the 5 major continental populations (AFR EUR, EAS, OCE and AME), followed by an analysis with populations from Eurasia and North Africa (NAF, ME, EUR, SAS and EAS populations). Despite the MHs chosen showing comparable levels of polymorphisms for all the populations (denoting the applicability of the panel regardless of the population), we achieved the separation of the five main population groups and, to some extent, the differentiation in Eurasian populations. Nonetheless, contrary to what was observed with ET, the separation between NAF and ME populations was less successful given the geographic proximity of these populations. Additionally, we also explored the application of the panel for ancestry inference in mixtures, in a similar approach to that performed for the VISAGE ET panel, but in this case employing the whole MH panel. The three simulated mixed profiles, at 1:3, 1:5 and 1:7 ratios, encompassed individuals from the same population or with different geographical backgrounds. Despite most of the deconvoluted profiles being partial, due to some shared haplotypes amongst contributing individuals, we achieved the correct ancestry classification in all instances. It must also be noted that deconvolution of mixtures with individuals with shared ancestry had an equal degree of success, due to the fact that the panel was constructed as an identification panel selecting highly polymorphic loci – applicable to all populations.

The results obtained in this study have helped contrast the hypothesis of how even large identification panels can be used to perform ancestry analyses, obtaining similar or enhanced predictions compared to other BGA-oriented small-scale panels. Despite that, and although MH panels are a field in forensic genetics that is explored actively and multiple panels for MPS typing have been developed in the recent years (a small compilation is included in section 1.3.1.4 of panels either based on already-discovered markers or comprising novel MHs), the majority of these studies focused on their use for identification and mixture analyses. Although these are the applications for which MHs are repeatedly selected, some studies have also explored their application for BGA prediction as well. The 16 loci compiled by van der Gaag and colleagues showed some degree of ancestry differentiation, separating African, European and East Asian populations. The 21 MHs from the Zou et al. compiled specifically for ancestry inference in Chinese subpopulations; the panel of 13 MHs by Zhu et al.; and the 20 MHs compiled by Kureshi et al. based on a selection of tri-allelic SNPs also accomplished, to some extent, the differentiation of 5-6 continental populations, although in general this panel struggles to separate closely related populations. The panel of 90 highly polymorphic MHs described by Gandotra and collaborators and the complementary study from Pakstis et al. (Pakstis et al., 2021) was the only panel in which the differentiation of 6 major populations was achieved.

Several other selections of MHs, e.g., the 130-MH panel by Kidd et al. (Kidd et al., 2017), showed promising results when the ancestry informativeness was evaluated unfortunately were either not implemented into MPS multiplexes or only a fraction of the markers were included in the final design of the multiplex. Other assays, such as the MAPlex panel (Phillips et al., 2019) or the previously discussed VISAGE ET panel, incorporate MHs that are used for ancestry inference, but in those instances either the power of the MHs is combined with those

USC
DE SANTIAGO
DE COMPOSTELA

of the aiSNPs and not analysed on their own or reserved for other applications given that, as previously observed for the BT panel, including extra markers on top of the carefully selected ancestry SNPs tends to increase the noise in the population analyses.

5.1.3 Towards the creation of a population-specific maker panel

The most common strategy when selecting AIMs is using metrics to measure the population specific divergence, and finding those that have strong contrasts between the frequencies in targeted population and the frequency of the same allele outside of that population. The two VISAGE panels and the original *Eurasiaplex* panel are a clear example of this: the markers selected present alleles that tend not to be exclusive, but we observe significant differences of their frequencies for the targeted population. With a new version of this panel: *Eurasiaplex-2*, we sought to construct a panel for the differentiation of South Asian populations based not on AIMs with high contrasts, but pursuing the adoption of markers showing alleles uniquely present in the target population, i.e., fully specific to that population only. In Figure D1 there are examples of the frequency distributions for these two types of markers: rs2026999 and rs3827760, in the VISAGE BT and ET panels, and rs191008849 and rs75721934, markers which present alleles unique for the SAS and EAS populations, respectively, in the 1000 Genomes Phase III data.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



Figure D1. Example of four SNPs suitable for BGA inference. The top two markers show high allele frequency contrasts between populations, but lack specific alleles; whereas the two markers on the bottom show specific alleles for a specific population.

However, the adoption of allele-specific SNPs required the construction of the panel *ab initio* and the compilation of a brand-new set markers. To search for candidate loci, the 1000 genomes Phase III SNP dataset was accessed, restricting the search to bi-allelic markers that had a considerable alternative allele frequency in the target population (> 0.1 frequency), here, South Asians, and virtually no presence in African, European, East Asian and American admixed populations (< 0.01 frequency). The results of the query were subsequently cross-checked with the genotype calls from the NYGC high coverage sequence-based SNP data to ensure the accuracy of the original 1000 Genomes genotyping, as the increased coverage can potentially identify errors in the Phase III data. Some markers which presented several genotype discordances or that lost their specificity to the SAS population were identified and removed from the candidate loci – rs3857620 is an example of such corrected genotype data, as a marker in the NYGC data that lost entirely the SAS specificity still shown in the SNP report page[4]. After genome-wide compilation, markers were chosen to be physically separated, taking into account possible linkage disequilibrium between them (except for six very informative SNPs on chromosome 16).

A total of 36 highly specific SNPs formed the final selection, with alternative allele frequencies reaching ~0.4 for the 1000 Genomes South Asian populations, and an average

[4] http://oct2024.archive.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=6:60527829-60528829;v=rs3857620;vdb=variation;vf=406391034

number of ~14 of South Asian-specific alleles per random South Asian individual in the 1000 Genomes sample set (excluding the Punjabi in Lahore). Population samples in other datasets, such as HGDP-CEPH or SGDP situated in regions close to, or originating near, the Indian subcontinent, like Uyghur or Roma, presented some South Asia-specific alleles, but the total count was significantly lower given the very low frequencies outside India or Bangladesh for these SNPs' specific (alternative) alleles. The Pakistani populations represented an intermediate pattern: the Punjabi in Lahore from 1000 Genomes in general displayed lower frequency estimates than the rest of the SAS populations (and, consequently, lower numbers of average specific alleles), and that lower overall SNP panel specificity was particularly marked in the 8 Pakistani populations from the HGDP-CEPH panel (Balochi, Brahui, Burusho, Hazara, Kalash, Makrani, Pathan and Sindhi). Across the rest of the world, the specific allele count was found to be zero or negligible in almost all cases.

The analysis of the 6 SNPs in chromosome 16 also represented a challenge, as they clustered around the centromere– an area in which recombination occurs at a lower rate, as observed in Phillips et al. (2012a). After calculating the recombination fraction from Kosambi, three SNPs, rs368479296-rs376893831-rs370130302, were found to have very a low recombination frequency, potentially indicating that those markers were in strict linkage disequilibrium. However, the haplotype TTG, which was the SAS specific haplotype, was observed to be disrupted in a high proportion of the SAS haplotypes, indicating that the association between these markers estimated from the HapMap data is lower.

The assessment of these markers using Bayes analysis, STRUCTURE and PCA successfully separated the SAS individuals from the other population groups of 1000 genomes – although, unlike the original Eurasiaplex panel, no other population could be differentiated apart from SAS, as the frequencies for the rest of population groups was virtually zero. Nonetheless, another objective that was pursued when trying to configure a panel based on population-specific markers was trying to simplify the analyses that are carried out to infer the ancestry, as these analyses require the use of specific software and expertise in the interpretation of the results. Instead, a much simpler approach, based on the absolute count of specific alleles was explored when using this type of marker. The distribution of the number of specific alleles from the SAS and non-SAS populations of 1000 genomes do not overlap (see Fig. 7 from article 4) and only Pakistani populations, with intermediate allele counts, present the problem of low specific allele counts in the CEPH samples analysed. In this instance, nevertheless, further testing will be required when more of these markers are compiled, and ultimately in case of uncertainty these samples could be left unassigned or perform subsequent analysis using more sophisticated software.

In brief, although the *Eurasiaplex-2* panel only constitutes an in-silico selection of markers which have still not been implemented into an MPS assay, this work constitutes a proof-of-concept for the adoption of this type of 'signature' AIM and simplified analyses based on counting methods; while also forming the first step towards the construction of a much larger global-scale panel for ancestry inference, based on the sole use of this new type of marker.

USC UNIVERSITAT DE SANTIAGO DE COMPOSTELA

5.2 A PANEL FOR KINSHIP ANALYSIS: MPSPLEX

STRs represent the marker of choice when performing identification cases, but the power of today's CE STR kits, which typically comprise around 20-24 STR loci, is not enough to resolve distant family relationships, usually beyond 2nd or 3rd degree relatives (Nothnagel et al., 2010). In these instances, opting for large-scale SNP panels can be the key to resolving kinship cases where inconclusive results are obtained after the analysis of regular STR kits. In terms of individualization, these markers have *per se* a lower power of discrimination, as they have a more limited number of possible alleles. To reach the same level of information, over 40 well-balanced bi-allelic SNPs are estimated to be equivalent to 15 STRs (Amorim & Pereira, 2005). However, the cumulative information of hundreds of SNPs, which can be easily targeted with the large sequencing capabilities of modern-day MPS sequencing, can surpass STR levels of informativeness and be used in complex kinship studies.

In the literature, there are several MPS multiplexes that have been developed for identification purposes, as well as commercially available panels that pursue the same goal (de la Puente et al., 2017; Grandell et al., 2016; Meiklejohn & Robertson, 2017). However, these panels have limited capabilities due to the low number of SNPs combined in the assay, usually targeting ~100 SNPs. The MPSplex panel was conceived as a large-scale identification panel that could be applied in these scenarios, while also aiming to increase the power of the panel by using a unique approach: targeting highly informative tri-allelic SNPs, that potentially have six genotypes rather than the three seen in bi-allelic SNPs. Within this section, the results of two different MPSplex studies are presented, comprising the development of the panel and its validation for MPS sequence analysis.

5.2.1 Marker selection and panel development

The first stage of the project focused on the search, evaluation and selection of candidate loci and the posterior implementation of the curated marker selection into an MPS assay. An early approach to this project was attempted using the Phase I results from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2012), but given the absence of genotype data that included a third allele for SNPs previously shown to be tri-allelic markers within the results of Phase I, the project was postponed until the release of the Phase III results (The 1000 Genomes Project Consortium, 2015), where a total of 271,934 tri-allelic SNPs were identified (including both autosomal and X-chromosome SNPs).

Well-balanced SNPs are usually sought when selecting markers for identification, as the probability of two random individuals sharing the same high frequency allele as a homozygous genotype is much higher if the alternative allele has a low frequency. The overall heterozygosity was used as a parameter to select the candidate SNPs, as this metric reaches its maximum value

when all the alleles present the same frequencies. This maximum value, in turn, depends on the number of possible alleles: bi-allelic SNPs present a maximum heterozygosity value of 0.50, followed by tri-allelic SNPs with 0.66, and 0.75 for the tetra-allelic SNPs. Consequently, for the selection of markers, SNPs were filtered by an expected heterozygosity equal or greater than 0.5, that is, higher than perfectly balanced bi-allelic SNPs. This allele frequency balance, in contrast to the development of the ancestry panels previously described, was also sought to be maintained throughout the main population groups, as that makes a panel suitable for any study regardless of the population of origin. The 3,700 remaining candidate markers after applying the heterozygosity filter underwent a subsequent selection process, in which they were selected in windows of at least 1-5 Mb and, generally, with at least 1 cM separation to reduce the linkage among them. Other factors were also taken into consideration to discard markers that could present issues when implemented in a multiplex, including the scrutiny of the context sequence to detect nearby elements that could hinder correct sequence alignment, or primer binding, and sequences that aligned to multiple locations in the genome. The resulting compilation of markers were implemented onto a prototype multiplex for a preliminary assessment, allowing the removal of markers with either low sequence coverage values, presenting three alleles, having discordancies with reference data, or presenting other quality issues (e.g., homopolymeric tracts in adjacent regions). The final panel included a total 1,270 tri-allelic SNPs, supplemented with 45 microhaplotypes from the 130 MH selection from Kidd Lab (Kidd et al., 2017) and one MH from Hiroaki et al. (2015).

In addition to the marker selection, several analyses were performed to assess the informativeness of the markers selected. Relationship testing simulations performed in Famlink allowed not only to evaluate the power of discrimination offered by the collection of markers selected, but also to evaluate the adequacy of two criteria used for marker selection: first, the increase in informativeness of tri-allelic SNPs, compared to the same number of perfectly balanced bi-allelic SNPs (i.e., showing the highest heterozygosity possible); and, second, examining the differences between choosing tri-allelic SNPs with evenly distributed allele frequencies (0.33:0.33:0.33) versus tri-allelic SNPs with a low-frequency third allele (0.45:0.45:0.1). Regarding the overall informativeness of the panel, the LR distributions of the simulations performed with the full set of 1377 (that is, before removing the underperforming markers in the pilot assay) indicated good separation for unrelated versus related up to 1st cousins (3rd degree), and 2nd cousins (5th degree). With 2nd cousin simulations, MPSplex did resolve a portion of cases, some overlap existed between the related and unrelated likelihood ratio distributions. Moreover, the two additional tests performed showed that tri-allelic SNPs presented overall higher LR values, as expected given their higher degree of polymorphism. However, the results of the simulations also indicated that, despite being less informative, the impact of having a third low-frequency allele was not as high, especially when accounting for the size of the MPSplex multiplex. This is explained by the fact that the odds of sharing a rare allele are increased in larger-scale multiplexes, which greatly increases the LR value obtained and consequently compensating for the lower informativeness per marker.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Finally, the ancestry informativeness of the 1,270 tri-allelic SNPs was also assessed with STRUCTURE, MDS and Neighbour-joining tree approaches. Despite the MPSplex marker selection process choosing SNPs with high heterozygosity across all populations, and intentionally avoiding SNPs with strong contrasts, the cumulative information of small differences in allele frequencies gave rise to a panel which was also informative for ancestry, even surpassing the power of small-scale panels of AIMs – similarly to what was observed for the MH panel. A good separation of the five continental populations was observed (AFR, EUR, EAS, OCE and AME) with the three approaches used, obtaining homogeneous clusters in the STRUCTURE, MDS and NJT analyses. Successful separation of SAS populations was also obtained to a lesser extent, mainly observed for the CEPH Pakistani samples – similarly to what was observed in previous analyses given their higher divergence from other Indian populations.

5.2.2 Assessment and validation of the MPSplex panel

In the second study of the MPSplex panel, we sought to evaluate the performance of the multiplex by conducting several studies, including sensitivity, ability to analyse mixtures and genotyping concordance analyses, which were performed on three of the mainstream MPS platforms in forensic genetics – the Ion S5, the MiSeq and GeneReader. This permitted not only the forensic validation of the MPSplex panel, but also a cross-comparison of the performance of each of the three MPS sequencing platforms. The composition of the panel used in the evaluation nonetheless was slightly modified from the original design from the first study, as 55 aiSNPs had consequently been incorporated into an expanded panel.

The results of the validation showed that most markers presented, in terms of universal molecular indices (UMI) reads, sequence coverage levels above the calling threshold for all the samples included as reference. However, a more in-depth analysis highlighted some issues for a portion of the samples and indicated inadequate pooling, as samples had been added in equimolar concentrations and not accordingly to their input DNA. The nature of UMI reads, which represent the original DNA copies, means that the number of different UMIs for each locus will be proportional to the DNA input used for library preparation; so, the more DNA that is used, the more different UMIs will be obtained for each locus. Each UMI requires to be read several times so that, when merging the reads bearing the same UMI tag, they can be used to correct PCR and sequencing errors. So, contrary to what is normally done with conventional MPS multiplexes, panels incorporating UMIs require that samples that are amplified with a greater DNA input (i.e., having more different UMIs for each locus) have higher sequencing depths. Otherwise, each UMI will be under-represented, and will cause the loss of the error-correction effect. This phenomenon was observed for samples with ≥10 ng DNA input: the reads / UMI ratio (i.e., the number of reads that contained the same UMI sequence and merged into a single UMI read) was found to be inadequate, presenting an average of 1-2 reads / UMI.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Although some differences were observed between platforms, the panel had an overall robust performance in all three systems; demonstrating a remarkable level of sensitivity, with near-complete profiles even with as little as 31 pg of input DNA in the Ion S5, and proficiency in the detection of mixtures, as several markers presented three different allele calls due to the composition of the panel. The concordance study also permitted the identification of several underperforming and problematic SNPs which may potentially cause issues in downstream analyses when using MPSplex in its current composition. Some of these markers presented alignment issues, while others were suspected to be located in duplicated regions from the observation of more than three different alleles simultaneously detected per locus. Despite that, high genotype concordance values were obtained on the whole when comparing the results obtained with MPS against the genotypes retrieved from whole-genome sequencing projects, and the loss of a small proportion of loci should have a minimum impact in the identification power of the multiplex given the scale of the panel as a whole – and the reduction in power of discrimination should be mitigated, if not increased, by the MH loci when they are incorporated into the analysis. The proof-of-concept kinship study exemplifies this situation, a case in which despite having obtained partial profiles for the MPSplex runs, the LR value obtained was more than 42 orders of magnitude higher than the one obtained using two different routine CE STR kits.

Regarding the applicability of the panel for forensic samples, this type of application often involves limited amounts of DNA, and its integrity can be rather low, which would be the case for the identification of human remains in mass graves. In these instances, the application of MPSplex could prove to be very useful, not only due to the high sensitivity observed during the validation, but also because these identifications often involve the comparison of unknown DNA with samples from distant relatives as the only reference DNA available – like the identification of remains from the Spanish Civil War (Baeta et al., 2015). Nonetheless, in spite of the promising results for the MPSplex panel performance, several issues were encountered during the experiments yet to be solved before the panel can be applied to this kind of forensic analysis. The QIAseq chemistry that the panel was implemented on, was not developed for forensic use, but rather conceived for clinical applications, such as the detection of low-frequency variants. In particular, the panel requirements state that a considerably high amount of input DNA (at least 10 ng) is needed and has several critical steps in the sequence library preparation workflow where DNA can be lost, and presents a degradation step within the protocol which may compromise samples. The manufacturer's protocol states that lower amounts of DNA input can be used, but with lower UMI coverage and reduced variant detection sensitivity as a result, but our study experiments helped elucidate how even with minimal amounts of DNA very good results can be obtained. Additionally, newer versions of the QIAseq chemistry (QIAseq Pro Panels) reduce the number of bead clean-ups by replacing them with enzymatic reactions, therefore minimizing the potential DNA loss. Nevertheless, notwithstanding the high sensitivity and the reduction of cleanups for the new kits, the QIAseq chemistry still entails a fragmentation step at the beginning of the library preparation that may not be suitable for samples which are already degraded. Follow-up studies will be required in

USC DE COMPOSTELA

order to evaluate the limitations with degraded samples, or, if possible, to explore whether the adaptation of the protocol for FFPE samples or a fragmentation step adapted to the degradation state of the DNA would be beneficial for this type of sample.

At the time of writing, two large-scale SNP panels have been introduced as forensic investigative genetic genealogy (FIGG) tools, whose utility in distant kinship analyses offers more power that MPSplex, as their identification capabilities greatly exceed the one for MPSplex. The new panels comprise the FORCE panel, developed by Tillmar et al. (Tillmar et al., 2021), and the ForenSeq® Kintelligence kit, a commercial panel developed by Verogen that showed very promising results in its developmental validation (Antunes et al., 2024) These panels, which comprise a total of 3,929 and 9,867 kinship SNPs, respectively, have been reported to identify up to 5th order relatives (Tillmar et al., 2021; Watson et al., 2024). In addition, Kintelligence offers other advantages, such as performing segment analysis via GEDmatch PRO instead of just LR based methods, and imputation can be applied, to further increase the number of SNP genotypes obtained from sequence low coverage data, allowing the detection of more distant relatives (Zavala et al., 2024). In addition, WGS also constitutes an alternative to targeted MPS panels for performing FIGG analyses, and this approach has already been shown to provide good results for identification with forensic samples (Tillmar et al., 2020). However, the large capability of these panels or the requirement to perform WGS analyses have consequences for the cost of the analyses, as they require extra sequencing resources (less samples per run can be analysed) and the expense of the kits are also considerably higher. Despite adoption of FIGG methodologies for routine analysis potentially providing many benefits, aiding the resolution of many criminal cases and thus justifying its higher expense (Budowle et al., 2023), in many cases the information can be excessive. Therefore, the MPSplex panel represents a pragmatic approach for forensic laboratories as it provides a cost-effective, mid-tier solution with great power of differentiation that can be used when facing distant kinship analyses in scenarios where routine markers and small-scale panels cannot provide conclusive results, while the use of FORCE and Kintelligence exceeds the requirements of the analyses (e.g., when several closely related surviving relatives are available for comparison)

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6. CONCLUSIONS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

6. CONCLUSIONS

i. The VISAGE Basic Tool (BT) panels was designed as a tool combining markers for BGA prediction and FDP targeting 42 appearance markers and 115 ancestry informative markers, the latter selected to differentiate between the 5 main continental population groups – Africans, Europeans, East Asians, Oceanians and native Americans – plus South Asians.

ii. The aiSNPs in the VISAGE BT provided the separation of the 6 targeted population groups, while co-ancestry estimates of individuals with admixed backgrounds correlate with established estimates using much larger SNP sets.

iii. The selection of markers for the VISAGE Enhanced Tool (ET) incorporated 78 bi-allelic and 26 tri-allelic autosomal SNPs, 21 MHs, 86 Y-SNPs and 16 ancestry-informative X-SNPs – the latter introduced into a modern forensic ancestry panel for the first time.

iv. Despite reducing the total number of bi-allelic and tri-allelic SNPs for autosomal ancestry inference, the careful selection of makers for ET, and increasing the overall number of markers while reducing the number for African and Oceanian differentiation, enabled a higher degree of population differentiation including the separation of North African and Middle East individuals.

v. The inclusion of non-autosomal markers in ET provides useful extra detail on the analysis of individuals with admixed backgrounds, while the pilot study on mixtures suggests the applicability of deconvoluted microhaplotype profiles for ancestry inference of the mixture contributors, albeit with a more limited resolution due to the reduced number of markers employed.

vi. The assignation of X-chromosome ancestry was explored via the use of 3-way PCA plots and, in parallel, the identification of population-specific haplotype blocks of the 5 centromeric X-SNPs. These methods yielded successful results in the cross-comparison with the STRUCTURE data from autosomal AIMs genotypes in admixed populations from Brazil, Fiji and African Americans, and they also indicated no disturbance of the signature haplotypes several generations after the occurrence of the admixture event.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

vii. Although the X-AIMs do not fully substitute the use of mtDNA, the possibility of co-amplification within the same multiplex and the level of co-ancestry information obtained from the X-SNPs in pilot population studies support their implementation in the assay and their use in forensic ancestry analysis.

viii. The applicability of a panel comprising 113 MH markers, originally developed for identification purposes, was tested to assess the level of ancestry information of the markers by performing a two-phase STRUCTURE analysis, which accomplished the differentiation of 5 main continental population groups (AFR, EUR, EAS, OCE, and AMR) and, with less success, the separation of 5 groups in the Eurasian region (NAF, ME, EUR, SAS, EAS) - in particular, the differentiation of NAF and ME populations was not completely achieved.

ix. The use of MHs for ancestry inference in two-person mixed DNAs was assessed using simulated mixtures with a similar strategy to the one used for the VISAGE ET MHs. The use of a whole panel of microhaplotypes instead of a small subset of the panel allowed an increase in the number of populations tested and gave successful results on deconvoluted profiles with partial profiles, achieving correct classification even with admixed persons or the two contributors that shared the same ancestry.

x. A total of 36 SNPs with near-completely specific alternative alleles for the South Asian populations were identified from the 1000 Genomes Phase 3 data and selected for the *Eurasiaplex-2* panel.

xi. Populations within the Indian subcontinent displayed an overall high population-specific allele count, and distributions that did not overlap with populations outside the South Asia region (except Pakistani populations), which could lead to a simplification of ancestry analyses allowing for the development of a simple allele counting system.

xii. The *Eurasiaplex-2* panel constitutes the first step towards the construction of a larger scale Signature AIMs panel, targeting the classification of individuals based on the use of markers with population-specific alleles.

xiii. A total of 271,934 tri-allelic SNPs were identified in the search of 1000 Genomes Phase 3 data, out of which 1,377 were selected to construct the MPSplex panel, a large-scale MPS panel for identification purposes.

xiv. The panel was implemented on a QIAseq Targeted DNA Custom Panel with UMI technology, successfully integrating 1,270 of the tri-allelic SNPs, 44 MHs and, later, an additional 55 aiSNPs.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

xv. The evaluation of the MPSplex panel across three different MPS platforms indicates it provides a robust and reliable forensic performance, demonstrating good sensitivity, mixture identification and a high genotype concordance rate.

xvi. The applicability of the panel for distant kinship analysis was tested on real casework giving high informativeness on a case where the use of CE typing technologies alone yielded inconclusive results.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

7. BIBLIOGRAPHY

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

7. BIBLIOGRAPHY

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., & McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56-65. https://doi.org/10.1038/nature11632

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

Almarri, M. A., Haber, M., Lootah, R. A., Hallast, P., Al Turki, S., Martin, H. C., Xue, Y., & Tyler-Smith, C. (2021). The genomic history of the Middle East. *Cell*, *184*(18), 4612-4625 e4614. https://doi.org/10.1016/j.cell.2021.07.013

Ambroa-Conde, A., Giron-Santamaria, L., Mosquera-Miguel, A., Phillips, C., Casares de Cal, M. A., Gomez-Tato, A., Alvarez-Dios, J., de la Puente, M., Ruiz-Ramirez, J., Lareu, M. V., & Freire-Aradas, A. (2022). Epigenetic age estimation in saliva and in buccal cells. *Forensic Sci Int Genet*, *61*, 102770. https://doi.org/10.1016/j.fsigen.2022.102770

Amorim, A., & Pereira, L. (2005). Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Sci Int*, *150*(1), 17-21. https://doi.org/10.1016/j.forsciint.2004.06.018

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., & Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, *290*(5806), 457-465. https://doi.org/10.1038/290457a0

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., & Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, *23*(2), 147. https://doi.org/10.1038/13779

Antunes, J., Walichiewicz, P., Forouzmand, E., Barta, R., Didier, M., Han, Y., Perez, J. C., Snedecor, J., Zlatkov, C., Padmabandu, G., Devesse, L., Radecke, S., Holt, C. L., Kumar, S. A., Budowle, B., & Stephens, K. M. (2024). Developmental validation of the ForenSeq(R) Kintelligence kit, MiSeq FGx(R) sequencing system and ForenSeq Universal Analysis Software. *Forensic Sci Int Genet*, *71*, 103055. https://doi.org/10.1016/j.fsigen.2024.103055

Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet*, *14*(2), 113-124. https://doi.org/10.1038/nrg3366

Baeta, M., Nunez, C., Cardoso, S., Palencia-Madrid, L., Herrasti, L., Etxeberria, F., & de Pancorbo, M. M. (2015). Digging up the recent Spanish memory: genetic identification of human remains from mass graves of the Spanish Civil War and posterior dictatorship. *Forensic Sci Int Genet*, *19*, 272-279. https://doi.org/10.1016/j.fsigen.2015.09.001

Ballard, D., Winkler-Galicki, J., & Wesoly, J. (2020). Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. *Int J Legal Med*, *134*(4), 1291-1303. https://doi.org/10.1007/s00414-020-02294-0

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Bauer, M. (2007). RNA in forensic science. *Forensic Sci Int Genet*, *1*(1), 69-74. https://doi.org/10.1016/j.fsigen.2006.11.002

Beckman, J. S., & Weber, J. L. (1992). Survey of human and rat microsatellites. *Genomics*, *12*(4), 627-631. https://doi.org/10.1016/0888-7543(92)90285-z

Benecke, M. (2001). A brief history of forensic entomology. *Forensic Sci Int*, *120*(1-2), 2-14. https://doi.org/10.1016/s0379-0738(01)00409-1

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., . . . Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53-59. https://doi.org/10.1038/nature07517

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., . . . Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484), eaay5012. https://doi.org/doi:10.1126/science.aay5012

Bird, A. (2007). Perceptions of epigenetics. *Nature*, *447*(7143), 396-398. https://doi.org/10.1038/nature05913

Borsani, G., Tonlorenzi, R., Simmler, M. C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C., Willard, H. F., Avner, P., & Ballabio, A. (1991). Characterization of a murine gene expressed from the inactive X chromosome. *Nature*, *351*(6324), 325-329. https://doi.org/10.1038/351325a0

Boullon-Cassau, M., Ambroa-Conde, A., Casares de Cal, M. A., Gomez-Tato, A., Mosquera-Miguel, A., Ruiz-Ramirez, J., Cabrejas-Olalla, A., Gonzalez-Bao, J., Casanova-Adan, L., de la Puente, M., Rodriguez, A., Phillips, C., Lareu, M. V., & Freire-Aradas, A. (2025). Exploring legal age estimation using DNA methylation. *Forensic Sci Int Genet*, *74*, 103142. https://doi.org/10.1016/j.fsigen.2024.103142

Boumil, R. M., & Lee, J. T. (2001). Forty years of decoding the silence in X-chromosome inactivation. *Hum Mol Genet*, *10*(20), 2225-2232. https://doi.org/10.1093/hmg/10.20.2225

Brancato, D., Coniglio, E., Bruno, F., Agostini, V., Saccone, S., & Federico, C. (2023). Forensic DNA Phenotyping: Genes and Genetic Variants for Eye Color Prediction. *Genes (Basel)*, *14*(8). https://doi.org/10.3390/genes14081604

Breslin, K., Wills, B., Ralf, A., Ventayol Garcia, M., Kukla-Bartoszek, M., Pospiech, E., Freire-Aradas, A., Xavier, C., Ingold, S., de La Puente, M., van der Gaag, K. J., Herrick, N., Haas, C., Parson, W., Phillips, C., Sijen, T., Branicki, W., Walsh, S., & Kayser, M. (2019). HIrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms. *Forensic Sci Int Genet*, *43*, 102152. https://doi.org/10.1016/j.fsigen.2019.102152

Budowle, B., Arnette, A., & Sajantila, A. (2023). A cost-benefit analysis for use of large SNP panels and high throughput typing for forensic investigative genetic genealogy. *Int J Legal Med*, *137*(5), 1595-1614. https://doi.org/10.1007/s00414-023-03029-7

Budowle, B., & van Daal, A. (2008). Forensically relevant SNP classes. *Biotechniques*, *44*(5), 603-608, 610. https://doi.org/10.2144/000112806

Butler, J., Shen, Y., & McCord, B. (2003). The Development of Reduced Size STR Amplicons as Tools for Analysis of Degraded DNA. *Journal of Forensic Sciences*, *48*(5), 1-11. https://doi.org/10.1520/jfs2003043

Butler, J. M. (2005). *Forensic DNA typing: biology, technology, and genetics of STR markers*. Elsevier.

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Human Genome Structural Variation, C., Paul, F., Germer, S., Brand, H., Hall, I. M., . . . Zody, M. C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426-3440 e3419. https://doi.org/10.1016/j.cell.2022.08.004

Cann, H. M., de Toma, C., Cazes, L., Legrand, M. F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., . . . Cavalli-Sforza, L. L. (2002). A human genome diversity cell line panel. *Science*, *296*(5566), 261-262. https://doi.org/10.1126/science.296.5566.261b

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, *325*(6099), 31-36. https://doi.org/10.1038/325031a0

Carvalho Gontijo, C., Porras-Hurtado, L. G., Freire-Aradas, A., Fondevila, M., Santos, C., Salas, A., Henao, J., Isaza, C., Beltran, L., Nogueira Silbiger, V., Castillo, A., Ibarra, A., Moreno Chavez, F., Sochtig, J., Ruiz, Y., Barreto, G., Rondon, F., Zabala, W., Borjas, L., . . . Phillips, C. (2020). PIMA: A population informative multiplex for the Americas. *Forensic Sci Int Genet*, *44*, 102200. https://doi.org/10.1016/j.fsigen.2019.102200

Casanova-Adan, L., Mosquera-Miguel, A., Gonzalez-Bao, J., Ambroa-Conde, A., Ruiz-Ramirez, J., Cabrejas-Olalla, A., Gonzalez-Martin, E., Freire-Aradas, A., Rodriguez-Lopez, A., Phillips, C., Lareu, M. V., & de la Puente, M. (2023). Adapting an established Ampliseq microhaplotype panel to nanopore sequencing through direct PCR. *Forensic Sci Int Genet*, *67*, 102937. https://doi.org/10.1016/j.fsigen.2023.102937

Chaitanya, L., Breslin, K., Zuniga, S., Wirken, L., Pospiech, E., Kukla-Bartoszek, M., Sijen, T., Knijff, P., Liu, F., Branicki, W., Kayser, M., & Walsh, S. (2018). The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. *Forensic Sci Int Genet*, *35*, 123-135. https://doi.org/10.1016/j.fsigen.2018.04.004

Chen, P., Yin, C., Li, Z., Pu, Y., Yu, Y., Zhao, P., Chen, D., Liang, W., Zhang, L., & Chen, F. (2018). Evaluation of the Microhaplotypes panel for DNA mixture analyses. *Forensic Sci Int Genet*, *35*, 149-155. https://doi.org/10.1016/j.fsigen.2018.05.003

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alfoldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Grant, R., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., . . . Karczewski, K. J. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, *625*(7993), 92-100. https://doi.org/10.1038/s41586-023-06045-0

Cheung, E. Y. Y., Gahan, M. E., & McNevin, D. (2017). Prediction of biogeographical ancestry from genotype: a comparison of classifiers. *Int J Legal Med*, *131*(4), 901-912. https://doi.org/10.1007/s00414-016-1504-3

Chien, A., Edgar, D. B., & Trela, J. M. (1976). Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *J Bacteriol*, *127*(3), 1550-1557. https://doi.org/10.1128/jb.127.3.1550-1557.1976

Coble, M. D., Loreille, O. M., Wadhams, M. J., Edson, S. M., Maynard, K., Meyer, C. E., Niederstatter, H., Berger, C., Berger, B., Falsetti, A. B., Gill, P., Parson, W., & Finelli, L. N. (2009). Mystery solved: the identification of the two missing Romanov children

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

using DNA analysis. *PLoS One*, *4*(3), e4838. https://doi.org/10.1371/journal.pone.0004838

Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W., & Pritchard, J. K. (2009). The role of geography in human adaptation. *PLoS Genet*, *5*(6), e1000500. https://doi.org/10.1371/journal.pgen.1000500

D'Mello, S. A., Finlay, G. J., Baguley, B. C., & Askarian-Amiri, M. E. (2016). Signaling Pathways in Melanogenesis. *Int J Mol Sci*, *17*(7). https://doi.org/10.3390/ijms17071144

de la Puente, M., Phillips, C., Santos, C., Fondevila, M., Carracedo, A., & Lareu, M. V. (2017). Evaluation of the Qiagen 140-SNP forensic identification multiplex for massively parallel sequencing. *Forensic Sci Int Genet*, *28*, 35-43. https://doi.org/10.1016/j.fsigen.2017.01.012

de la Puente, M., Phillips, C., Xavier, C., Amigo, J., Carracedo, A., Parson, W., & Lareu, M. V. (2020). 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. https://doi.org/10.1016/j.fsigen.2019.102213

de la Puente, M., Santos, C., Fondevila, M., Manzo, L., The EUROFORGEN-NoE Consortium, Carracedo, A., Lareu, M. V., & Phillips, C. (2016). The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs. *Forensic Sci Int Genet*, *22*, 81-88. https://doi.org/10.1016/j.fsigen.2016.01.015

Diepenbroek, M., Bayer, B., Schwender, K., Schiller, R., Lim, J., Lagace, R., & Anslinger, K. (2020). Evaluation of the Ion AmpliSeq PhenoTrivium Panel: MPS-Based Assay for Ancestry and Phenotype Predictions Challenged by Casework Samples. *Genes (Basel)*, *11*(12). https://doi.org/10.3390/genes11121398

Du, Q., Ma, G., Lu, C., Wang, Q., Fu, L., Cong, B., & Li, S. (2023). Development and evaluation of a novel panel containing 188 microhaplotypes for 2nd-degree kinship testing in the Hebei Han population. *Forensic Sci Int Genet*, *65*, 102855. https://doi.org/10.1016/j.fsigen.2023.102855

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359-361. https://doi.org/10.1007/s12686-011-9548-7

Eduardoff, M., Gross, T. E., Santos, C., de la Puente, M., Ballard, D., Strobl, C., Borsting, C., Morling, N., Fusco, L., Hussing, C., Egyed, B., Souto, L., Uacyisrael, J., Syndercombe Court, D., Carracedo, A., Lareu, M. V., Schneider, P. M., Parson, W., Phillips, C., . . . Phillips, C. (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. https://doi.org/10.1016/j.fsigen.2016.04.008

Enoch, M. A., Shen, P. H., Xu, K., Hodgkinson, C., & Goldman, D. (2006). Using ancestry-informative markers to define populations and detect population stratification. *J Psychopharmacol*, *20*(4 Suppl), 19-26. https://doi.org/10.1177/1359786806066041

Excoffier, L., & 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*(3), 564-567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Fan, H., & Chu, J. Y. (2007). A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics*, *5*(1), 7-14. https://doi.org/10.1016/S1672-0229(07)60009-6

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Fitzgerald, R. J., Price, H. L., & Valentine, T. (2018). Eyewitness Identification: Live, Photo, and Video Lineups. *Psychol Public Policy Law*, *24*(3), 307-325. https://doi.org/10.1037/law0000164

Fleming, R. I., & Harbison, S. (2010). The development of a mRNA multiplex RT-PCR assay for the definitive identification of body fluids. *Forensic Sci Int Genet*, *4*(4), 244-256. https://doi.org/10.1016/j.fsigen.2009.10.006

Fondevila, M., Phillips, C., Naverán, N., Cerezo, M., Rodríguez, A., Calvo, R., Fernández, L. M., Carracedo, Á., & Lareu, M. V. (2008). Challenging DNA: Assessment of a range of genotyping approaches for highly degraded forensic samples. *Forensic Science International: Genetics Supplement Series*, *1*(1), 26-28. https://doi.org/10.1016/j.fsigss.2007.10.057

Fondevila, M., Phillips, C., Santos, C., Freire Aradas, A., Vallone, P. M., Butler, J. M., Lareu, M. V., & Carracedo, A. (2013). Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. *Forensic Sci Int Genet*, *7*(1), 63-74. https://doi.org/10.1016/j.fsigen.2012.06.007

Fondevila, M., Phillips, C., Santos, C., Pereira, R., Gusmao, L., Carracedo, A., Butler, J. M., Lareu, M. V., & Vallone, P. M. (2012). Forensic performance of two insertion-deletion marker assays. *Int J Legal Med*, *126*(5), 725-737. https://doi.org/10.1007/s00414-012-0721-7

Franklin, R. E., & Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, *171*(4356), 740-741. https://doi.org/10.1038/171740a0

Freire-Aradas, A., Phillips, C., Mosquera-Miguel, A., Giron-Santamaria, L., Gomez-Tato, A., Casares de Cal, M., Alvarez-Dios, J., Ansede-Bermejo, J., Torres-Espanol, M., Schneider, P. M., Pospiech, E., Branicki, W., Carracedo, A., & Lareu, M. V. (2016). Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system. *Forensic Sci Int Genet*, *24*, 65-74. https://doi.org/10.1016/j.fsigen.2016.06.005

Freire-Aradas, A., Ruiz, Y., Phillips, C., Maronas, O., Sochtig, J., Tato, A. G., Dios, J. A., de Cal, M. C., Silbiger, V. N., Luchessi, A. D., Luchessi, A. D., Chiurillo, M. A., Carracedo, A., & Lareu, M. V. (2014). Exploring iris colour prediction and ancestry inference in admixed populations of South America. *Forensic Sci Int Genet*, *13*, 3-9. https://doi.org/10.1016/j.fsigen.2014.06.007

Freire-Aradas, A., Tomsia, M., Piniewska-Rog, D., Ambroa-Conde, A., Casares de Cal, M. A., Pisarek, A., Gomez-Tato, A., Alvarez-Dios, J., Pospiech, E., Parson, W., Kayser, M., Phillips, C., & Branicki, W. (2023). Development of an epigenetic age predictor for costal cartilage with a simultaneous somatic tissue differentiation system. *Forensic Sci Int Genet*, *67*, 102936. https://doi.org/10.1016/j.fsigen.2023.102936

Friedlaender, J. S., Friedlaender, F. R., Reed, F. A., Kidd, K. K., Kidd, J. R., Chambers, G. K., Lea, R. A., Loo, J. H., Koki, G., Hodgson, J. A., Merriwether, D. A., & Weber, J. L. (2008). The genetic structure of Pacific Islanders. *PLoS Genet*, *4*(1), e19. https://doi.org/10.1371/journal.pgen.0040019

Frumkin, D., Wasserstrom, A., Davidson, A., & Grafit, A. (2010). Authentication of forensic DNA samples. *Forensic Sci Int Genet*, *4*(2), 95-103. https://doi.org/10.1016/j.fsigen.2009.06.009

Galante, N., Cotroneo, R., Furci, D., Lodetti, G., & Casali, M. B. (2023). Applications of artificial intelligence in forensic sciences: Current potential benefits, limitations and

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

perspectives. *Int J Legal Med*, *137*(2), 445-458. https://doi.org/10.1007/s00414-022-02928-5

Galanter, J. M., Fernandez-Lopez, J. C., Gignoux, C. R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A. V., Figueroa, L. U., Raska, P., Jimenez-Sanchez, G., Zolezzi, I. S., Torres, M., Ponte, C. R., Ruiz, Y., Salas, A., Nguyen, E., Eng, C., Borjas, L., . . . Consortium, L. (2012). Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet*, *8*(3), e1002554. https://doi.org/10.1371/journal.pgen.1002554

Gandotra, N., Speed, W. C., Qin, W., Tang, Y., Pakstis, A. J., Kidd, K. K., & Scharfe, C. (2020). Validation of novel forensic DNA markers using multiplex microhaplotype sequencing. *Forensic Sci Int Genet*, *47*, 102275. https://doi.org/10.1016/j.fsigen.2020.102275

Ge, J., & Budowle, B. (2020). How many familial relationship testing results could be wrong? *PLoS Genet*, *16*(8), e1008929. https://doi.org/10.1371/journal.pgen.1008929

Geserick, G., & Wirth, I. (2012). Genetic Kinship Investigation from Blood Groups to DNA Markers. *Transfus Med Hemother*, *39*(3), 163-175. https://doi.org/10.1159/000338850

Gettings, K. B., Bodner, M., Borsuk, L. A., King, J. L., Ballard, D., Parson, W., Benschop, C. C. G., Borsting, C., Budowle, B., Butler, J. M., van der Gaag, K. J., Gill, P., Gusmao, L., Hares, D. R., Hoogenboom, J., Irwin, J., Prieto, L., Schneider, P. M., Vennemann, M., & Phillips, C. (2024). Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on short tandem repeat sequence nomenclature. *Forensic Sci Int Genet*, *68*, 102946. https://doi.org/10.1016/j.fsigen.2023.102946

Gibbons, A. (2011). Anthropology. A new view of the birth of Homo sapiens. *Science*, *331*(6016), 392-394. https://doi.org/10.1126/science.331.6016.392

Gill, P., Fereday, L., Morling, N., & Schneider, P. M. (2006). The evolution of DNA databases--recommendations for new European STR loci. *Forensic Sci Int*, *156*(2-3), 242-244. https://doi.org/10.1016/j.forsciint.2005.05.036

Gill, P., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Evett, I., Hagelberg, E., & Sullivan, K. (1994). Identification of the remains of the Romanov family by DNA analysis. *Nat Genet*, *6*(2), 130-135. https://doi.org/10.1038/ng0294-130

Girón Santamaría, L. (2021). *Desarrollo de técnicas cronométricas utilizando marcadores epigenéticos para el análisis en genética forense* Universidade de Santiago de Compostela].

Gomes, I., Pinto, N., Antao-Sousa, S., Gomes, V., Gusmao, L., & Amorim, A. (2020). Twenty Years Later: A Comprehensive Review of the X Chromosome Use in Forensic Genetics. *Front Genet*, *11*, 926. https://doi.org/10.3389/fgene.2020.00926

Gontijo, C. C., Mendes, F. M., Santos, C. A., Klautau-Guimaraes, M. N., Lareu, M. V., Carracedo, A., Phillips, C., & Oliveira, S. F. (2018). Ancestry analysis in rural Brazilian populations of African descent. *Forensic Sci Int Genet*, *36*, 160-166. https://doi.org/10.1016/j.fsigen.2018.06.018

Gonzalez-Bao, J., Mosquera-Miguel, A., Casanova-Adan, L., Ambroa-Conde, A., Ruiz-Ramirez, J., Cabrejas-Olalla, A., Boullon-Cassau, M., Freire-Aradas, A., Rodriguez-Lopez, A., Roth, C., Lagace, R., Phillips, C., Lareu, M. V., & de la Puente, M. (2025). Performance comparison of a previously validated microhaplotype panel and a forensic STR panel for DNA mixture analysis. *Forensic Sci Int Genet*, *74*, 103144. https://doi.org/10.1016/j.fsigen.2024.103144

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, *17*(6), 333-351. https://doi.org/10.1038/nrg.2016.49

Goodwin, W., Linacre, A., & Hadi, S. (2010). *An introduction to forensic genetics*. John Wiley & Sons.

Gosch, A., Bhardwaj, A., & Courts, C. (2023). TrACES of time: Transcriptomic analyses for the contextualization of evidential stains - Identification of RNA markers for estimating time-of-day of bloodstain deposition. *Forensic Sci Int Genet*, *67*, 102915. https://doi.org/10.1016/j.fsigen.2023.102915

Grandell, I., Samara, R., & Tillmar, A. O. (2016). A SNP panel for identity and kinship testing using massive parallel sequencing. *Int J Legal Med*, *130*(4), 905-914. https://doi.org/10.1007/s00414-016-1341-4

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., . . . Paabo, S. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710-722. https://doi.org/10.1126/science.1188021

Greenberg, B. D., Newbold, J. E., & Sugino, A. (1983). Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. *Gene*, *21*(1-2), 33-49. https://doi.org/10.1016/0378-1119(83)90145-2

Grimes, E. A., Noake, P. J., Dixon, L., & Urquhart, A. (2001). Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. *Forensic Sci Int*, *122*(2-3), 124-129. https://doi.org/10.1016/s0379-0738(01)00480-7

Gustafsson, P. U., Lindholm, T., & Jonsson, F. U. (2022). Eyewitness accuracy and retrieval effort: Effects of time and repetition. *PLoS One*, *17*(9), e0273455. https://doi.org/10.1371/journal.pone.0273455

Haas, C., Neubauer, J., Salzmann, A. P., Hanson, E., & Ballantyne, J. (2021). Forensic transcriptome analysis using massively parallel sequencing. *Forensic Sci Int Genet*, *52*, 102486. https://doi.org/10.1016/j.fsigen.2021.102486

Hares, D. R. (2015). Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci Int Genet*, *17*, 33-34. https://doi.org/10.1016/j.fsigen.2015.03.006

Hashiguchi, K., Bohr, V. A., & de Souza-Pinto, N. C. (2004). Oxidative stress and mitochondrial DNA repair: implications for NRTIs induced DNA damage. *Mitochondrion*, *4*(2-3), 215-222. https://doi.org/10.1016/j.mito.2004.05.014

Heaton-Armstrong, A. (1995). Eye-witness testimony and judicial studies. *Med Sci Law*, *35*(2), 93-94. https://doi.org/10.1177/002580249503500201

Helmuth, R., Fildes, N., Blake, E., Luce, M. C., Chimera, J., Madej, R., Gorodezky, C., Stoneking, M., Schmill, N., Klitz, W., Higuchi, R., & Erlich, H. A. (1990). HLA-DQ alpha allele and genotype frequencies in various human populations, determined by using enzymatic amplification and oligonucleotide probes. *Am J Hum Genet*, *47*(3), 515-523.

Henn, B. M., Hon, L., Macpherson, J. M., Eriksson, N., Saxonov, S., Pe'er, I., & Mountain, J. L. (2012). Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. *PLoS One*, *7*(4), e34267. https://doi.org/10.1371/journal.pone.0034267

Hiroaki, N., Koji, F., Tetsushi, K., Kazumasa, S., Hiroaki, N., & Kazuyuki, S. (2015). Approaches for identifying multiple-SNP haplotype blocks for use in human

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

identification. *Leg Med (Tokyo)*, *17*(5), 415-420. https://doi.org/10.1016/j.legalmed.2015.06.003

Hoogenboom, J., Sijen, T., & van der Gaag, K. J. (2021). STRNaming: Generating simple, informative names for sequenced STR alleles in a standardised and automated manner. *Forensic Sci Int Genet*, *52*, 102473. https://doi.org/10.1016/j.fsigen.2021.102473

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol*, *14*(10), R115. https://doi.org/10.1186/gb-2013-14-10-r115

Huber, N., Parson, W., & Dur, A. (2018). Next generation database search algorithm for forensic mitogenome analyses. *Forensic Sci Int Genet*, *37*, 204-214. https://doi.org/10.1016/j.fsigen.2018.09.001

Ingram, C. J., Mulcare, C. A., Itan, Y., Thomas, M. G., & Swallow, D. M. (2009). Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet*, *124*(6), 579-591. https://doi.org/10.1007/s00439-008-0593-6

International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851-861. https://doi.org/10.1038/nature06258

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931-945. https://doi.org/10.1038/nature03001

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lonnerberg, P., & Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods*, *11*(2), 163-166. https://doi.org/10.1038/nmeth.2772

ISO/IEC. (2017). ISO/IEC 17025:2017 General requirements for the competence of testing and calibration laboratories. 42.

Ivanov, P. L., Wadhams, M. J., Roby, R. K., Holland, M. M., Weedn, V. W., & Parsons, T. J. (1996). Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. *Nat Genet*, *12*(4), 417-420. https://doi.org/10.1038/ng0496-417

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*(14), 1801-1806. https://doi.org/10.1093/bioinformatics/btm233

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Individual-specific 'fingerprints' of human DNA. *Nature*, *316*(6023), 76-79. https://doi.org/10.1038/316076a0

Jobling, M., Hollox, E., Kivisild, T., & Tyler-Smith, C. (2013). *Human evolutionary genetics* (2nd ed.). Garland Science.

Jobling, M., & Tyler-Smith, C. (2004). *Human Evolutionary Genetics: Origins, Peoples and Disease* (1st ed.). Garland Science. https://doi.org/https://doi.org/10.1201/9780203487211

Jobling, M. A., & Gill, P. (2004). Encoded evidence: DNA in forensic analysis. *Nat Rev Genet*, *5*(10), 739-751. https://doi.org/10.1038/nrg1455

Jobling, M. A., & Tyler-Smith, C. (2003). The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet*, *4*(8), 598-612. https://doi.org/10.1038/nrg1124

Jung, S. E., Lim, S. M., Hong, S. R., Lee, E. H., Shin, K. J., & 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. *Forensic Sci Int Genet*, *38*, 1-8. https://doi.org/10.1016/j.fsigen.2018.09.010

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Kayser, M. (2015). Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci Int Genet*, *18*, 33-48. https://doi.org/10.1016/j.fsigen.2015.02.003

Kayser, M., Branicki, W., Parson, W., & Phillips, C. (2023). Recent advances in Forensic DNA Phenotyping of appearance, ancestry and age. *Forensic Sci Int Genet*, *65*, 102870. https://doi.org/10.1016/j.fsigen.2023.102870

Kharoshah, M. A., Zaki, M. K., Galeb, S. S., Moulana, A. A., & Elsebaay, E. A. (2011). Origin and development of forensic medicine in Egypt. *J Forensic Leg Med*, *18*(1), 10-13. https://doi.org/10.1016/j.jflm.2010.11.009

Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., & Ihuegbu, N. (2013). Microhaplotype loci are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series*, *4*(1), e123-e124. https://doi.org/https://doi.org/10.1016/j.fsigss.2013.10.063

Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F. R., & Kidd, J. R. (2014). Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet*, *10*, 23-32. https://doi.org/10.1016/j.fsigen.2014.01.002

Kidd, K. K., Speed, W. C., Pakstis, A. J., Podini, D. S., Lagace, R., Chang, J., Wootton, S., Haigh, E., & Soundararajan, U. (2017). Evaluating 130 microhaplotypes across a global set of 83 populations. *Forensic Sci Int Genet*, *29*, 29-37. https://doi.org/10.1016/j.fsigen.2017.03.014

Kimpton, C., Walton, A., & Gill, P. (1992). A further tetranucleotide repeat polymorphism in the vWF gene. *Hum Mol Genet*, *1*(4), 287. https://doi.org/10.1093/hmg/1.4.287

Kimpton, C. P., Gill, P., Walton, A., Urquhart, A., Millican, E. S., & Adams, M. (1993). Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Methods Appl*, *3*(1), 13-22. https://doi.org/10.1101/gr.3.1.13

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W., & Vogelstein, B. (2011). Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A*, *108*(23), 9530-9535. https://doi.org/10.1073/pnas.1105422108

Kling, D., Egeland, T., & Tillmar, A. O. (2012). FamLink--a user friendly software for linkage calculations in family genetics. *Forensic Sci Int Genet*, *6*(5), 616-620. https://doi.org/10.1016/j.fsigen.2012.01.012

Kling, D., Phillips, C., Kennett, D., & Tillmar, A. (2021). Investigative genetic genealogy: Current methods, knowledge and practice. *Forensic Sci Int Genet*, *52*, 102474. https://doi.org/10.1016/j.fsigen.2021.102474

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, *15*(5), 1179-1191. https://doi.org/https://doi.org/10.1111/1755-0998.12387

Kureshi, A., Li, J., Wen, D., Sun, S., Yang, Z., & Zha, L. (2020). Construction and forensic application of 20 highly polymorphic microhaplotypes. *R Soc Open Sci*, *7*(5), 191937. https://doi.org/10.1098/rsos.191937

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., . . . International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860-921. https://doi.org/10.1038/35057062

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Landsteiner, K. (1900). *Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe*. https://books.google.es/books?id=67lIOAAACAAJ

Landsteiner, K. (1901). Ueber Agglutinationserscheinungen normalen menschlichen Blutes. *Wien Klin Wochenschr*, *113*(20-21), 768-769.

Lareu, M. V., Barral, S., Salas, A., Pestoni, C., & Carracedo, A. (1998). Sequence variation of a hypervariable short tandem repeat at the D1S1656 locus. *Int J Legal Med*, *111*(5), 244-247. https://doi.org/10.1007/s004140050161

Lareu, M. V., Pestoni, C., Carracedo, A., Schürenkamp, M., Rand, S., & Brinkmann, B. (1996). A highly variable STR at the D12S391 locus. *International Journal of Legal Medicine*, *109*(3), 134-138. https://doi.org/10.1007/BF01369673

Lee, H. Y., Hong, S. R., Lee, J. E., Hwang, I. K., Kim, N. Y., Lee, J. M., Fleckhaus, J., Jung, S. E., & Lee, Y. H. (2020). Epigenetic age signatures in bones. *Forensic Sci Int Genet*, *46*, 102261. https://doi.org/10.1016/j.fsigen.2020.102261

Lee, J. E., Park, S. U., So, M. H., & Lee, H. Y. (2024). Age prediction using DNA methylation of Y-chromosomal CpGs in semen samples. *Forensic Sci Int Genet*, *69*, 103007. https://doi.org/10.1016/j.fsigen.2024.103007

Lee, J. T., Davidow, L. S., & Warshawsky, D. (1999). Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet*, *21*(4), 400-404. https://doi.org/10.1038/7734

Lewontin, R. C. (1972). The Apportionment of Human Diversity. In T. Dobzhansky, M. K. Hecht, & W. C. Steere (Eds.), *Evolutionary Biology: Volume 6* (pp. 381-398). Springer US. https://doi.org/10.1007/978-1-4684-9063-3_14

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, *319*(5866), 1100-1104. https://doi.org/10.1126/science.1153717

Lind, J. M., Hutcheson-Dilks, H. B., Williams, S. M., Moore, J. H., Essex, M., Ruiz-Pesini, E., Wallace, D. C., Tishkoff, S. A., O'Brien, S. J., & Smith, M. W. (2007). Elevated male European and female African contributions to the genomes of African American individuals. *Hum Genet*, *120*(5), 713-722. https://doi.org/10.1007/s00439-006-0261-7

Lindenbergh, A., de Pagter, M., Ramdayal, G., Visser, M., Zubakov, D., Kayser, M., & Sijen, T. (2012). A multiplex (m)RNA-profiling system for the forensic identification of body fluids and contact traces. *Forensic Sci Int Genet*, *6*(5), 565-577. https://doi.org/10.1016/j.fsigen.2012.01.009

Liu, F., Hamer, M. A., Heilmann, S., Herold, C., Moebus, S., Hofman, A., Uitterlinden, A. G., Nothen, M. M., van Duijn, C. M., Nijsten, T. E., & Kayser, M. (2016). Prediction of male-pattern baldness from genotypes. *Eur J Hum Genet*, *24*(6), 895-902. https://doi.org/10.1038/ejhg.2015.220

Liu, F., Hendriks, A. E., Ralf, A., Boot, A. M., Benyi, E., Savendahl, L., Oostra, B. A., van Duijn, C., Hofman, A., Rivadeneira, F., Uitterlinden, A. G., Drop, S. L., & Kayser, M. (2014). Common DNA variants predict tall stature in Europeans. *Hum Genet*, *133*(5), 587-597. https://doi.org/10.1007/s00439-013-1394-0

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nat Rev Genet*, *21*(10), 597-614. https://doi.org/10.1038/s41576-020-0236-x

Londin, E. R., Keller, M. A., Maista, C., Smith, G., Mamounas, L. A., Zhang, R., Madore, S. J., Gwinn, K., & Corriveau, R. A. (2010). CoAIMs: a cost-effective panel of ancestry

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

informative markers for determining continental origins. *PLoS One*, *5*(10), e13443. https://doi.org/10.1371/journal.pone.0013443

Lowe, A. L., Urquhart, A., Foreman, L. A., & Evett, I. W. (2001). Inferring ethnic origin by means of an STR profile. *Forensic Sci Int*, *119*(1), 17-22. https://doi.org/10.1016/s0379-0738(00)00387-x

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., . . . Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201-206. https://doi.org/10.1038/nature18964

Mannucci, A., Sullivan, K. M., Ivanov, P. L., & Gill, P. (1994). Forensic application of a rapid and quantitative DNA sex test by amplification of the X-Y homologous gene amelogenin. *Int J Legal Med*, *106*(4), 190-193. https://doi.org/10.1007/BF01371335

Mao, X., Bigham, A. W., Mei, R., Gutierrez, G., Weiss, K. M., Brutsaert, T. D., Leon-Velarde, F., Moore, L. G., Vargas, E., McKeigue, P. M., Shriver, M. D., & Parra, E. J. (2007). A genomewide admixture mapping panel for Hispanic/Latino populations. *Am J Hum Genet*, *80*(6), 1171-1178. https://doi.org/10.1086/518564

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376-380. https://doi.org/10.1038/nature03959

Maronas, O., Phillips, C., Sochtig, J., Gomez-Tato, A., Cruz, R., Alvarez-Dios, J., de Cal, M. C., Ruiz, Y., Fondevila, M., Carracedo, A., & Lareu, M. V. (2014). Development of a forensic skin colour predictive test. *Forensic Sci Int Genet*, *13*, 34-44. https://doi.org/10.1016/j.fsigen.2014.06.017

Martin-Baniandres, P., Lan, W. H., Board, S., Romero-Ruiz, M., Garcia-Manyes, S., Qing, Y., & Bayley, H. (2023). Enzyme-less nanopore detection of post-translational modifications within long polypeptides. *Nat Nanotechnol*, *18*(11), 1335-1340. https://doi.org/10.1038/s41565-023-01462-8

Martin, P. D., Schmitter, H., & Schneider, P. M. (2001). A brief history of the formation of DNA databases in forensic science within Europe. *Forensic Sci Int*, *119*(2), 225-231. https://doi.org/10.1016/s0379-0738(00)00436-9

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, *74*(2), 560-564. https://doi.org/10.1073/pnas.74.2.560

McWilliams, T. G., & Suomalainen, A. (2019). Mitochondrial DNA can be inherited from fathers, not just mothers. *Nature*, *565*(7739), 296-297. https://doi.org/10.1038/d41586-019-00093-1

Meiklejohn, K. A., & Robertson, J. M. (2017). Evaluation of the Precision ID Identity Panel for the Ion Torrent() PGM() sequencer. *Forensic Sci Int Genet*, *31*, 48-56. https://doi.org/10.1016/j.fsigen.2017.08.009

Meissner, C., von Wurmb, N., Schimansky, B., & Oehmichen, M. (1999). Estimation of age at death based on quantitation of the 4977-bp deletion of human mitochondrial DNA in skeletal muscle. *Forensic Sci Int*, *105*(2), 115-124. https://doi.org/10.1016/s0379-0738(99)00126-7

Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segre, A. V., Djebali, S., Niarchou, A., Consortium, G. T., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G.,

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

. . . Guigo, R. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science*, *348*(6235), 660-665. https://doi.org/10.1126/science.aaa0355

Mengel-From, J., Borsting, C., Sanchez, J. J., Eiberg, H., & Morling, N. (2010). Human eye colour and HERC2, OCA2 and MATP. *Forensic Sci Int Genet*, *4*(5), 323-328. https://doi.org/10.1016/j.fsigen.2009.12.004

Mertes, F., Elsharawy, A., Sauer, S., van Helvoort, J. M., van der Zaag, P. J., Franke, A., Nilsson, M., Lehrach, H., & Brookes, A. J. (2011). Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief Funct Genomics*, *10*(6), 374-386. https://doi.org/10.1093/bfgp/elr033

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*, *16*(9), 1182-1190. https://doi.org/10.1101/gr.4565806

Mills, R. E., Pittard, W. S., Mullaney, J. M., Farooq, U., Creasy, T. H., Mahurkar, A. A., Kemeza, D. M., Strassler, D. S., Ponting, C. P., Webber, C., & Devine, S. E. (2011). Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res*, *21*(6), 830-839. https://doi.org/10.1101/gr.115907.110

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harbor symposia on quantitative biology,

Musgrave-Brown, E., Ballard, D., Balogh, K., Bender, K., Berger, B., Bogus, M., Borsting, C., Brion, M., Fondevila, M., Harrison, C., Oguzturun, C., Parson, W., Phillips, C., Proff, C., Ramos-Luis, E., Sanchez, J. J., Sanchez Diz, P., Sobrino Rey, B., Stradmann-Bellinghausen, B., . . . Syndercombe Court, D. (2007). Forensic validation of the SNPforID 52-plex assay. *Forensic Sci Int Genet*, *1*(2), 186-190. https://doi.org/10.1016/j.fsigen.2007.01.004

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297-304. https://doi.org/10.1093/genetics/156.1.297

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A*, *70*(12), 3321-3323. https://doi.org/10.1073/pnas.70.12.3321

Newman, A. M., Lovejoy, A. F., Klass, D. M., Kurtz, D. M., Chabon, J. J., Scherer, F., Stehr, H., Liu, C. L., Bratman, S. V., Say, C., Zhou, L., Carter, J. N., West, R. B., Sledge, G. W., Shrager, J. B., Loo, B. W., Jr., Neal, J. W., Wakelee, H. A., Diehn, M., & Alizadeh, A. A. (2016). Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol*, *34*(5), 547-555. https://doi.org/10.1038/nbt.3520

Nothnagel, M., Schmidtke, J., & Krawczak, M. (2010). Potentials and limits of pairwise kinship analysis using autosomal short tandem repeat loci. *Int J Legal Med*, *124*(3), 205-215. https://doi.org/10.1007/s00414-009-0413-0

Nyren, P., Pettersson, B., & Uhlen, M. (1993). Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem*, *208*(1), 171-175. https://doi.org/10.1006/abio.1993.1024

Oldoni, F., Bader, D., Fantinato, C., Wootton, S. C., Lagace, R., Kidd, K. K., & Podini, D. (2020). A sequence-based 74plex microhaplotype assay for analysis of forensic DNA mixtures. *Forensic Sci Int Genet*, *49*, 102367. https://doi.org/10.1016/j.fsigen.2020.102367

Oldroyd, N. J., Urquhart, A. J., Kimpton, C. P., Millican, E. S., Watson, S. K., Downes, T., & Gill, P. D. (1995). A highly discriminating octoplex short tandem repeat polymerase chain reaction system suitable for human individual identification. *Electrophoresis*, *16*(3), 334-337. https://doi.org/10.1002/elps.1150160155

Oxford English Dictionary. (2023). forensic, adj. & n., Etymology. In *Oxford English Dictionary*.

Pagani, L., Lawson, D. J., Jagoda, E., Morseburg, A., Eriksson, A., Mitt, M., Clemente, F., Hudjashov, G., DeGiorgio, M., Saag, L., Wall, J. D., Cardona, A., Magi, R., Wilson Sayres, M. A., Kaewert, S., Inchley, C., Scheib, C. L., Jarve, M., Karmin, M., . . . Metspalu, M. (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, *538*(7624), 238-242. https://doi.org/10.1038/nature19792

Pakstis, A. J., Gandotra, N., Speed, W. C., Murtha, M., Scharfe, C., & Kidd, K. K. (2021). The population genetics characteristics of a 90 locus panel of microhaplotypes. *Hum Genet*, *140*(12), 1753-1773. https://doi.org/10.1007/s00439-021-02382-0

Palencia-Madrid, L., Xavier, C., de la Puente, M., Hohoff, C., Phillips, C., Kayser, M., & Parson, W. (2020). Evaluation of the VISAGE Basic Tool for Appearance and Ancestry Prediction Using PowerSeq Chemistry on the MiSeq FGx System. *Genes (Basel)*, *11*(6). https://doi.org/10.3390/genes11060708

Pang, J. B., Rao, M., Chen, Q. F., Ji, A. Q., Zhang, C., Kang, K. L., Wu, H., Ye, J., Nie, S. J., & Wang, L. (2020). A 124-plex Microhaplotype Panel Based on Next-generation Sequencing Developed for Forensic Applications. *Sci Rep*, *10*(1), 1945. https://doi.org/10.1038/s41598-020-58980-x

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, *26*(3), 419-420. https://doi.org/10.1093/bioinformatics/btp696

Pearn, J. (2016). Hammurabi's Code: A primary datum in the conjoined professions of medicine and law. *Med Leg J*, *84*(3), 125-131. https://doi.org/10.1177/0025817216646038

Pereira, F., Meirinhos, J., Amorim, A., & Pereira, L. (2006). Analysis of inter-specific mitochondrial DNA diversity for accurate species identification. *International Congress Series*, *1288*, 103-105. https://doi.org/https://doi.org/10.1016/j.ics.2005.09.125

Pereira, L., Alshamali, F., Andreassen, R., Ballard, R., Chantratita, W., Cho, N. S., Coudray, C., Dugoujon, J. M., Espinoza, M., Gonzalez-Andrade, F., Hadi, S., Immel, U. D., Marian, C., Gonzalez-Martin, A., Mertens, G., Parson, W., Perone, C., Prieto, L., Takeshita, H., . . . Fonseca, N. A. (2011). PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile. *Int J Legal Med*, *125*(5), 629-636. https://doi.org/10.1007/s00414-010-0472-2

Pereira, R., Phillips, C., Pinto, N., Santos, C., dos Santos, S. E., Amorim, A., Carracedo, A., & Gusmao, L. (2012). Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS One*, *7*(1), e29684. https://doi.org/10.1371/journal.pone.0029684

Pereira, V., Freire-Aradas, A., Ballard, D., Borsting, C., Diez, V., Pruszkowska-Przybylska, P., Ribeiro, J., Achakzai, N. M., Aliferi, A., Bulbul, O., Carceles, M. D. P., Triki-Fendri, S., Rebai, A., Court, D. S., Morling, N., Lareu, M. V., Carracedo, A., Consortium, E. U.-N., & Phillips, C. (2019). Development and validation of the EUROFORGEN NAME (North African and Middle Eastern) ancestry panel. *Forensic Sci Int Genet*, *42*, 260-267. https://doi.org/10.1016/j.fsigen.2019.06.010

Phillips, C. (2015). Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci Int Genet*, *18*, 49-65. https://doi.org/10.1016/j.fsigen.2015.05.012

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

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Phillips, C., Amigo, J., McNevin, D., de la Puente, M., Cheung, E. Y. Y., & Lareu, M. V. (2021). Online population data resources for forensic SNP analysis with Massively Parallel Sequencing: An overview of online population data for forensic purposes. In E. Pilli & A. Berti (Eds.), *Forensic DNA Analysis: Technological Development and Innovative Applications*. CRC Press: Boca Raton.

Phillips, C., Amigo, J., Tillmar, A. O., Peck, M. A., de la Puente, M., Ruiz-Ramirez, J., Bittner, F., Idrizbegovic, S., Wang, Y., Parsons, T. J., & Lareu, M. V. (2020). A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel. *Forensic Sci Int Genet*, *46*, 102232. https://doi.org/10.1016/j.fsigen.2020.102232

Phillips, C., Ballard, D., Gill, P., Court, D. S., Carracedo, A., & Lareu, M. V. (2012a). The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data. *Forensic Sci Int Genet*, *6*(3), 354-365. https://doi.org/10.1016/j.fsigen.2011.07.012

Phillips, C., Fernandez-Formoso, L., Garcia-Magarinos, M., Porras, L., Tvedebrink, T., Amigo, J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Freire-Aradas, A., Gomez-Carballa, A., Mosquera-Miguel, A., Carracedo, A., & Lareu, M. V. (2011). Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. *Forensic Sci Int Genet*, *5*(3), 155-169. https://doi.org/10.1016/j.fsigen.2010.02.003

Phillips, C., Fernandez-Formoso, L., Gelabert-Besada, M., Garcia-Magarinos, M., Santos, C., Fondevila, M., Carracedo, A., & Lareu, M. V. (2013a). Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing. *Electrophoresis*, *34*(8), 1151-1162. https://doi.org/10.1002/elps.201200621

Phillips, C., Fondevila, M., & Lareau, M. V. (2012b). A 34-plex autosomal SNP single base extension assay for ancestry investigations. *Methods Mol Biol*, *830*, 109-126. https://doi.org/10.1007/978-1-61779-461-2_8

Phillips, C., Freire Aradas, A., Kriegel, A. K., Fondevila, M., Bulbul, O., Santos, C., Serrulla Rech, F., Perez Carceles, M. D., Carracedo, A., Schneider, P. M., & Lareu, M. V. (2013b). Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Sci Int Genet*, *7*(3), 359-366. https://doi.org/10.1016/j.fsigen.2013.02.010

Phillips, C., McNevin, D., Kidd, K. K., Lagace, R., Wootton, S., de la Puente, M., Freire-Aradas, A., Mosquera-Miguel, A., Eduardoff, M., Gross, T., Dagostino, L., Power, D., Olson, S., Hashiyada, M., Oz, C., Parson, W., Schneider, P. M., Lareu, M. V., & Daniel, R. (2019). MAPlex - A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. *Forensic Sci Int Genet*, *42*, 213-226. https://doi.org/10.1016/j.fsigen.2019.06.022

Phillips, C., Salas, A., Sanchez, J. J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M. C., Ballard, D., Lareu, M. V., Carracedo, A., & Consortium, S. N. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet*, *1*(3-4), 273-280. https://doi.org/10.1016/j.fsigen.2007.06.008

Pickrell, J. K., & Reich, D. (2014). Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet*, *30*(9), 377-389. https://doi.org/10.1016/j.tig.2014.07.007

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Pilin, A., Pudil, F., & Bencko, V. (2007). Changes in colour of different human tissues as a marker of age. *Int J Legal Med*, *121*(2), 158-162. https://doi.org/10.1007/s00414-006-0136-4

Plesivkova, D., Richards, R., & Harbison, S. (2019). A review of the potential of the MinION™ single-molecule sequencing system for forensic applications. *WIREs Forensic Science*, *1*(1), e1323. https://doi.org/https://doi.org/10.1002/wfs2.1323

Pospiech, E., Karlowska-Pik, J., Marcinska, M., Abidi, S., Andersen, J. D., Berge, M. V. D., Carracedo, A., Eduardoff, M., Freire-Aradas, A., Morling, N., Sijen, T., Skowron, M., Sochtig, J., Syndercombe-Court, D., Weiler, N., Schneider, P. M., Ballard, D., Borsting, C., Parson, W., . . . Consortium, E. U.-N. (2015). Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans. *Forensic Sci Int Genet*, *19*, 280-288. https://doi.org/10.1016/j.fsigen.2015.09.004

President's Council of Advisors on Science and Technology. (2016). *Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods*.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959. https://doi.org/10.1093/genetics/155.2.945

Qu, N., Lin, S., Gao, Y., Liang, H., Zhao, H., & Ou, X. (2020). A microhap panel for kinship analysis through massively parallel sequencing technology. *Electrophoresis*, *41*(3-4), 246-253. https://doi.org/10.1002/elps.201900337

Ralf, A., van Oven, M., Montiel Gonzalez, D., de Knijff, P., van der Beek, K., Wootton, S., Lagace, R., & Kayser, M. (2019). Forensic Y-SNP analysis beyond SNaPshot: High-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing. *Forensic Sci Int Genet*, *41*, 93-106. https://doi.org/10.1016/j.fsigen.2019.04.001

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol*, *19*(1), 90. https://doi.org/10.1186/s13059-018-1462-9

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., . . . Paabo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, *468*(7327), 1053-1060. https://doi.org/10.1038/nature09710

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol*, *29*(1), 24-26. https://doi.org/10.1038/nbt.1754

Rosenberg, N. A. (2004). distruct: a program for the graphical display of population structure. *Molecular Ecology Notes*, *4*(1), 137-138. https://doi.org/https://doi.org/10.1046/j.1471-8286.2003.00566.x

Rosenberg, N. A., Li, L. M., Ward, R., & Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet*, *73*(6), 1402-1422. https://doi.org/10.1086/380416

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, *298*(5602), 2381-2385. https://doi.org/10.1126/science.1078311

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G. R., Burrows, C., Bird, C. P., Frankish, A., Lovell, F. L., Howe, K. L., Ashurst,

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

J. L., Fulton, R. S., Sudbrak, R., Wen, G., Jones, M. C., Hurles, M. E., . . . Bentley, D. R. (2005). The DNA sequence of the human X chromosome. *Nature*, *434*(7031), 325-337. https://doi.org/10.1038/nature03440

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., . . . Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348-352. https://doi.org/10.1038/nature10242

Ruiz, Y., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., Casares de Cal, M., Cruz, R., Maronas, O., Sochtig, J., Fondevila, M., Rodriguez-Cid, M. J., Carracedo, A., & Lareu, M. V. (2013). Further development of forensic eye color predictive tests. *Forensic Sci Int Genet*, *7*(1), 28-40. https://doi.org/10.1016/j.fsigen.2012.05.009

Sagan, L. (1967). On the origin of mitosing cells. *J Theor Biol*, *14*(3), 255-274. https://doi.org/10.1016/0022-5193(67)90079-3

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, *4*(4), 406-425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977a). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, *265*(5596), 687-695. https://doi.org/10.1038/265687a0

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*, *94*(3), 441-448. https://doi.org/10.1016/0022-2836(75)90213-2

Sanger, F., Nicklen, S., & Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, *74*(12), 5463-5467. https://doi.org/10.1073/pnas.74.12.5463

Sankararaman, S., Patterson, N., Li, H., Paabo, S., & Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. *PLoS Genet*, *8*(10), e1002947. https://doi.org/10.1371/journal.pgen.1002947

Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R. A. H., Burchard, E. G., Schanfield, M. S., Souto, L., Uacyisrael, J., Via, M., Carracedo, A., & Lareu, M. V. (2016). Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region. *Forensic Sci Int Genet*, *20*, 71-80. https://doi.org/10.1016/j.fsigen.2015.10.003

Sapan, V., Simsek, S. Z., Filoglu, G., & Bulbul, O. (2024). Forensic DNA phenotyping using Oxford Nanopore Sequencing system. *Electrophoresis*. https://doi.org/10.1002/elps.202300252

Satoh, M., & Kuroiwa, T. (1991). Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. *Exp Cell Res*, *196*(1), 137-140. https://doi.org/10.1016/0014-4827(91)90467-9

Scientific Working Group on DNA Analysis Methods. (2016). Validation guidelines for DNA analysis methods. In: SWGDAM Washington, DC.

Segurel, L., & Bon, C. (2017). On the Evolution of Lactase Persistence in Humans. *Annu Rev Genomics Hum Genet*, *18*, 297-319. https://doi.org/10.1146/annurev-genom-091416-035340

Sharma, V., & Litt, M. (1992). Tetranucleotide repeat polymorphism at the D21S11 locus. *Hum Mol Genet*, *1*(1), 67. https://doi.org/10.1093/hmg/1.1.67-a

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., Chinwalla, A., Delehaunty, A., Delehaunty, K., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S. F., . . . Page, D. C. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, *423*(6942), 825-837. https://doi.org/10.1038/nature01722

Sochtig, J., Phillips, C., Maronas, O., Gomez-Tato, A., Cruz, R., Alvarez-Dios, J., de Cal, M. A., Ruiz, Y., Reich, K., Fondevila, M., Carracedo, A., & Lareu, M. V. (2015). Exploration of SNP variants affecting hair colour prediction in Europeans. *Int J Legal Med*, *129*(5), 963-975. https://doi.org/10.1007/s00414-015-1226-y

Sokolov, B. P. (1990). Primer extension technique for the detection of single nucleotide in genomic DNA. *Nucleic Acids Res*, *18*(12), 3671. https://doi.org/10.1093/nar/18.12.3671

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, *98*(3), 503-517. https://doi.org/https://doi.org/10.1016/S0022-2836(75)80083-0

Spinney, L. (2008). Eyewitness identification: line-ups on trial. *Nature*, *453*(7194), 442-444. https://doi.org/10.1038/453442a

Staadig, A., & Tillmar, A. (2021). Evaluation of microhaplotypes in forensic kinship analysis from a Swedish population perspective. *Int J Legal Med*, *135*(4), 1151-1160. https://doi.org/10.1007/s00414-021-02509-y

Straub, R. E., Speer, M. C., Luo, Y., Rojas, K., Overhauser, J., Ott, J., & Gilliam, T. C. (1993). A microsatellite genetic linkage map of human chromosome 18. *Genomics*, *15*(1), 48-56. https://doi.org/10.1006/geno.1993.1008

Sun, S., Liu, Y., Li, J., Yang, Z., Wen, D., Liang, W., Yan, Y., Yu, H., Cai, J., & Zha, L. (2020). Development and application of a nonbinary SNP-based microhaplotype panel for paternity testing involving close relatives. *Forensic Sci Int Genet*, *46*, 102255. https://doi.org/10.1016/j.fsigen.2020.102255

Syndercombe Court, D. (2018). Forensic genealogy: Some serious concerns. *Forensic Sci Int Genet*, *36*, 203-204. https://doi.org/10.1016/j.fsigen.2018.07.011

Termine, J. D., Belcourt, A. B., Christner, P. J., Conn, K. M., & Nylen, M. U. (1980). Properties of dissociatively extracted fetal tooth matrix proteins. I. Principal molecular species in developing bovine enamel. *J Biol Chem*, *255*(20), 9760-9768. https://www.ncbi.nlm.nih.gov/pubmed/7430099

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74. https://doi.org/10.1038/nature15393

Tillmar, A., Sjolund, P., Lundqvist, B., Klippmark, T., Algenas, C., & Green, H. (2020). Whole-genome sequencing of human remains to enable genealogy DNA database searches - A case report. *Forensic Sci Int Genet*, *46*, 102233. https://doi.org/10.1016/j.fsigen.2020.102233

Tillmar, A., Sturk-Andreaggi, K., Daniels-Higginbotham, J., Thomas, J. T., & Marshall, C. (2021). The FORCE Panel: An All-in-One SNP Marker Set for Confirming Investigative Genetic Genealogy Leads and for General Forensic Applications. *Genes (Basel)*, *12*(12). https://doi.org/10.3390/genes12121968

Tsuji, A., Ishiko, A., Takasaki, T., & Ikeda, N. (2002). Estimating age of humans based on telomere shortening. *Forensic Sci Int*, *126*(3), 197-199. https://doi.org/10.1016/s0379-0738(02)00086-5

Tucker, T., Marra, M., & Friedman, J. M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet*, *85*(2), 142-154. https://doi.org/10.1016/j.ajhg.2009.06.022

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Turchi, C., Melchionda, F., Pesaresi, M., & Tagliabracci, A. (2019). Evaluation of a microhaplotypes panel for forensic genetics using massive parallel sequencing technology. *Forensic Sci Int Genet*, *41*, 120-127. https://doi.org/10.1016/j.fsigen.2019.04.009

van der Gaag, K. J., de Leeuw, R. H., Laros, J. F. J., den Dunnen, J. T., & de Knijff, P. (2018). Short hypervariable microhaplotypes: A novel set of very short high discriminating power loci without stutter artefacts. *Forensic Science International: Genetics*, *35*, 169-175. https://doi.org/https://doi.org/10.1016/j.fsigen.2018.05.008

van Oven, M., & Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat*, *30*(2), E386-394. https://doi.org/10.1002/humu.20921

Vidaki, A., & Kayser, M. (2018). Recent progress, methods and perspectives in forensic epigenetics. *Forensic Sci Int Genet*, *37*, 180-195. https://doi.org/10.1016/j.fsigen.2018.08.008

Vigeland, M. D. (2022). QuickPed: an online tool for drawing pedigrees and analysing relatedness. *BMC Bioinformatics*, *23*(1), 220. https://doi.org/10.1186/s12859-022-04759-y

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., & Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. *Science*, *253*(5027), 1503-1507. https://doi.org/10.1126/science.1840702

Visser, M., Kayser, M., & Palstra, R. J. (2012). HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res*, *22*(3), 446-455. https://doi.org/10.1101/gr.128652.111

von Dungern, E., & Hirschfeld, L. (1911). Über Vererbung gruppenspezifischer Strukturen des Blutes. *Zeitschrift für induktive Abstammungs- und Vererbungslehre*, *5*(1), 196-197. https://doi.org/10.1007/BF01798027

Wagner, J., Olson, N. D., Harris, L., Khan, Z., Farek, J., Mahmoud, M., Stankovic, A., Kovacevic, V., Yoo, B., Miller, N., Rosenfeld, J. A., Ni, B., Zarate, S., Kirsche, M., Aganezov, S., Schatz, M. C., Narzisi, G., Byrska-Bishop, M., Clarke, W., . . . Zook, J. M. (2022). Benchmarking challenging small variants with linked and long reads. *Cell Genom*, *2*(5). https://doi.org/10.1016/j.xgen.2022.100128

Wai, T., Ao, A., Zhang, X., Cyr, D., Dufort, D., & Shoubridge, E. A. (2010). The role of mitochondrial DNA copy number in mammalian fertility. *Biol Reprod*, *83*(1), 52-62. https://doi.org/10.1095/biolreprod.109.080887

Waite, E. R., Collins, M. J., Ritz-Timme, S., Schutz, H. W., Cattaneo, C., & Borrman, H. I. (1999). A review of the methodological aspects of aspartic acid racemization analysis for use in forensic science. *Forensic Sci Int*, *103*(2), 113-124. https://doi.org/10.1016/s0379-0738(99)00081-x

Walsh, S., Liu, F., Ballantyne, K. N., van Oven, M., Lao, O., & Kayser, M. (2011). 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*(3), 170-180. https://doi.org/10.1016/j.fsigen.2010.02.004

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W., & Kayser, M. (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci Int Genet*, *7*(1), 98-115. https://doi.org/10.1016/j.fsigen.2012.07.005

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Wang, H., Kang, B., Gao, Y., Zhang, M., Jiang, J., Su, J., Zhang, B., Zhu, B., & Liao, S. (2023). Construction and evaluation of a novel set of 90 microhaplotypes for forensic applications using NGS technology. *Forensic Sci Int*, *353*, 111848. https://doi.org/10.1016/j.forsciint.2023.111848

Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M. V., Molina, J. A., Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A. M., Labuda, D., Klitz, W., Barrantes, R., Bortolini, M. C., Salzano, F. M., . . . Ruiz-Linares, A. (2007). Genetic variation and population structure in native Americans. *PLoS Genet*, *3*(11), e185. https://doi.org/10.1371/journal.pgen.0030185

Watson, J., McNevin, D., Grisedale, K., Spiden, M., Seddon, S., & Ward, J. (2024). Operationalisation of the ForenSeq(R) Kintelligence Kit for Australian unidentified and missing persons casework. *Forensic Sci Int Genet*, *68*, 102972. https://doi.org/10.1016/j.fsigen.2023.102972

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737-738. https://doi.org/10.1038/171737a0

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., & Marth, G. (2002). Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet*, *71*(4), 854-862. https://doi.org/10.1086/342727

White, J. D., Indencleef, K., Naqvi, S., Eller, R. J., Hoskens, H., Roosenboom, J., Lee, M. K., Li, J., Mohammed, J., Richmond, S., Quillen, E. E., Norton, H. L., Feingold, E., Swigut, T., Marazita, M. L., Peeters, H., Hens, G., Shaffer, J. R., Wysocka, J., . . . Claes, P. (2021). Insights into the genetic architecture of the human face. *Nat Genet*, *53*(1), 45-53. https://doi.org/10.1038/s41588-020-00741-7

Wollstein, A., Lao, O., Becker, C., Brauer, S., Trent, R. J., Nurnberg, P., Stoneking, M., & Kayser, M. (2010). Demographic history of Oceania inferred from genome-wide data. *Curr Biol*, *20*(22), 1983-1992. https://doi.org/10.1016/j.cub.2010.10.040

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, *16*(2), 97-159. https://doi.org/10.1093/genetics/16.2.97

Wright, S. (1950). Genetical structure of populations. *Nature*, *166*(4215), 247-249. https://doi.org/10.1038/166247a0

Wu, R., Li, H., Li, R., Peng, D., Wang, N., Shen, X., & Sun, H. (2021). Identification and sequencing of 59 highly polymorphic microhaplotypes for analysis of DNA mixtures. *Int J Legal Med*, *135*(4), 1137-1149. https://doi.org/10.1007/s00414-020-02483-x

Xavier, C., de la Puente, M., Mosquera-Miguel, A., Freire-Aradas, A., Kalamara, V., Ralf, A., Revoir, A., Gross, T. E., Schneider, P. M., Ames, C., Hohoff, C., Phillips, C., Kayser, M., Parson, W., & Consortium, V. (2022). Development and inter-laboratory evaluation of the VISAGE Enhanced Tool for Appearance and Ancestry inference from DNA. *Forensic Sci Int Genet*, *61*, 102779. https://doi.org/10.1016/j.fsigen.2022.102779

Xavier, C., de la Puente, M., Mosquera-Miguel, A., Freire-Aradas, A., Kalamara, V., Vidaki, A., T, E. G., Revoir, A., Pospiech, E., Kartasinska, E., Spolnicka, M., Branicki, W., C, E. A., P, M. S., Hohoff, C., Kayser, M., Phillips, C., Parson, W., & Consortium, V. (2020). Development and validation of the VISAGE AmpliSeq basic tool to predict appearance and ancestry from DNA. *Forensic Sci Int Genet*, *48*, 102336. https://doi.org/10.1016/j.fsigen.2020.102336

Y Chromosome Consortium. (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res*, *12*(2), 339-348. https://doi.org/10.1101/gr.217602

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Zardoya, R. (2020). Recent advances in understanding mitochondrial genome diversity. *F1000Res*, *9*. https://doi.org/10.12688/f1000research.21490.1

Zavala, E. I., Rohlfs, R. V., & Moorjani, P. (2024). Benchmarking for genotyping and imputation using degraded DNA for forensic applications across diverse populations. *bioRxiv*, 2024.2007.2002.601808. https://doi.org/10.1101/2024.07.02.601808

Zhang, Z., & Gerstein, M. (2003). Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res*, *31*(18), 5338-5348. https://doi.org/10.1093/nar/gkg745

Zhu, J., Lv, M., Zhou, N., Chen, D., Jiang, Y., Wang, L., He, W., Peng, D., Li, Z., Qu, S., Wang, Y., Wang, H., Luo, H., An, G., Liang, W., & Zhang, L. (2019). Genotyping polymorphic microhaplotype markers through the Illumina((R)) MiSeq platform for forensics. *Forensic Sci Int Genet*, *39*, 1-7. https://doi.org/10.1016/j.fsigen.2018.11.005

Zou, X., He, G., Liu, J., Jiang, L., Wang, M., Chen, P., Hou, Y., & Wang, Z. (2022). Screening and selection of 21 novel microhaplotype markers for ancestry inference in ten Chinese subpopulations. *Forensic Sci Int Genet*, *58*, 102687. https://doi.org/10.1016/j.fsigen.2022.102687

Zubakov, D., Liu, F., van Zelm, M. C., Vermeulen, J., Oostra, B. A., van Duijn, C. M., Driessen, G. J., van Dongen, J. J., Kayser, M., & Langerak, A. W. (2010). Estimating human age from T-cell DNA rearrangements. *Curr Biol*, *20*(22), R970-971. https://doi.org/10.1016/j.cub.2010.10.022

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ANNEXES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

ANNEXES

ANNEX I. SUPPLEMENTARY TABLES FOR ARTICLE 6

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Supplementary Table S1. Detailed information on the 44 MH loci included in the multiplex.

MH No.	Chr	5' Start SNP Pos (GRCh37/hg19)	3' End SNP Pos (GRCh37/hg19)	Size	Original code	Original publication	Reduced size?	5' original coordinates	3' original coordinates	5' Start SNP	3' End SNP	Genetic position (cM)	Composite SNPs
1	1	3743319	3743391	72	mh01KK-001 / MHA-01	Kidd et al. 2017 [17] / Phillips et al. 2019 [18]	yes - shortened on 5'	3743132	3743391	rs6663840	rs6688969	8,8	rs6663840-rs58111155-rs6688969
2	1	18722801	18722846	45	mh01KK-205	Kidd et al. 2017 [17]	yes - shortened on 5'	18722692	18722846	rs1533623	rs1533622	39,5	rs1533623-rs1533622
3	1	161955758	161955804	46	mh01KK-211	Kidd et al. 2017 [17]	yes - shortened on 5'	161955656	161955804	rs16835127	rs2341465	174,5	rs16835127-rs10918374-rs2341465
4	1	230820578	230820630	52	(no names assigned)	Hiroaki et al. 2015 [1]	no	-	-	rs2296796	rs2296798	248,3	rs2296796-rs2296797-rs2296798
5	2	161079411	161079450	39	mh02KK-134 / MHA-03	Kidd et al. 2017 [17] / Phillips et al. 2019 [18]	yes - shortened on 3'	161079411	161079514	rs12469721	rs3111398	175,5	rs12469721-rs3101043-rs3111398
6	2	228092389	228092459	70	mh02KK-136 / MHA-04	Kidd et al. 2017 [17] / Phillips et al. 2019 [18]	no	228092389	228092459	rs6714835	rs12617010	240,3	rs6714835-rs6756898-rs12617010
7	3	46352355	46352384	29	mh03KK-007	Kidd et al. 2017 [17]	no	-	-	rs4513489	rs6441961	67,6	rs4513489-rs6441961
8	3	131645972	131646001	29	mh03KK-150	Kidd et al. 2017 [17]	yes - shortened on 3'	131645972	131646156	rs1225051	rs1225050	141,9	rs1225051-rs1225050
9	4	1986938	1986972	34	mh04KK-010	Kidd et al. 2017 [17]	no	-	-	rs3135123	rs495367	2,1	rs3135123-rs495367
10	4	37858890	37858954	64	mh04KK-011	Kidd et al. 2017 [17]	no	-	-	rs6855439	rs6531591	59,5	rs6855439-rs6531591
11	4	68444180	68444257	77	mh04KK-013 / MHA-05	Kidd et al. 2017 [17] / Phillips et al. 2019 [18]	yes - shortened on both ends	68444102	68444302	rs3775866	rs3775867	84,2	rs3775866-rs11725922-rs3775867
12	5	2448146	2448160	14	mh05KK-170	Kidd et al. 2017 [17]	yes - shortened on 5'	2448024	2448160	rs370672	rs6555108	5	rs370672-rs6555108
13	5	9619905	9619936	31	mh05KK-022	Kidd et al. 2017 [17]	no	-	-	rs41461	rs41462	21,9	rs41461-rs41462
14	9	680714	680791	77	mh09KK-033	Kidd et al. 2017 [17]	no	-	-	rs10815466	rs17431629	1	rs10815466-rs9408671-rs9408672-rs17431629
15	9	85808708	85808790	82	mh09KK-152	Kidd et al. 2017 [17]	yes - shortened on 5'	85808649	85808790	rs4282648	rs7046769	85,1	rs4282648-rs10780576-rs7046769
16	9	135862563	135862632	69	mh09KK-157	Kidd et al. 2017 [17]	yes - shortened on 5'	135862479	135862632	rs56256724	rs633153	150	rs56256724-rs2073578-rs633153
17	9	137417276	137417308	32	mh09KK-035	Kidd et al. 2017 [17]	yes - shortened on 5'	137417115	137417308	rs76875728	rs10776839	155,8	rs76875728-rs7037930-rs10776839
18	10	3162410	3162486	76	mh10KK-163	Kidd et al. 2017 [17]	yes - shortened on 3'	3162410	3162669	rs3814588	rs3814589	5,7	rs3814588-rs6602026-rs3814589
19	11	1690950	1690984	34	mh11KK-180 / MHA-11	Kidd et al. 2017 [17] / Phillips et al. 2019 [18]	yes - shortened on 5'	1690791	1690984	rs4752778	rs4752777	2,2	rs4752778-rs7112918-rs4752777
20	11	5109946	5110068	122	mh11KK-036	Kidd et al. 2017 [17]	no	-	-	rs10500616	rs2499936	10,2	rs10500616-rs2499936
21	11	99880163	99880224	61	mh11KK-191 / MHA-12	Kidd et al. 2017 [17]	yes - shortened on 3'	99880163	99880352	rs12421109	rs12289401	105,2	rs12421109-rs1393073426-rs12289401
22	11	113264470	113264511	41	mh11KK-040	Kidd et al. 2017 [17]	no	-	-	rs4938013	rs11214596	117	rs4938013-rs11214596
23	12	30170306	30170382	76	mh12KK-202	Kidd et al. 2017 [17]	yes - shortened on 5'	30170229	30170382	rs4931233	rs4931234	51,4	rs4931233-rs10506053-rs4931234
24	12	118889488	118889559	71	mh12KK-046	Kidd et al. 2017 [17]	no	-	-	rs1503767	rs11068953	141,1	rs1503767-rs11068953
25	13	23765601	23765681	80	mh13KK-213	Kidd et al. 2017 [17]	yes - shortened on 5'	23765541	23765681	rs2152727	rs9510616	9,9	rs2152727-rs679482-rs9510616
26	13	46866059	46866122	63	mh13KK-217	Kidd et al. 2017 [17]	yes - shortened on 5'	46865930	46866122	rs9534373	rs2765614	48,2	rs9534373-rs9562649-rs2765614
27	13	54060827	54060892	65	mh13KK-218	Kidd et al. 2017 [17]	yes - shortened on 3'	54060827	54060972	rs1927847	rs7492234	54,6	rs1927847-rs9536429-rs7492234

28	13	66712790	66712828	38	mh13KK-225	Kidd et al. 2017 [17]	yes - shortened on 5'	66712732	66712828	rs9529023	rs7329287	61,9	rs9529023-rs4884652-rs7329287
29	13	110806699	110806759	60	mh13KK-223	Kidd et al. 2017 [17]	yes - shortened on 3'	110806699	110806852	rs1192204	rs3825483	116,4	rs1192204-rs1192205-rs3825483
30	15	46870734	46870855	121	mh15KK-067	Kidd et al. 2017 [17]	no	-	-	rs701463	rs701464	59,3	rs701463-rs7183984-rs701464
31	15	52484950	52485024	74	mh15KK-066	Kidd et al. 2017 [17]	no	-	-	rs1063902	rs4219	63,2	rs1063902-rs4219
32	16	7209208	7209267	59	mh16KK-049 / MHA-16	Kidd et al. 2017 [17] / Phillips et al. 2019 [18]	yes - shortened on 5'	7209208	7209381	rs9937467	rs17670111	16,5	rs9937467-rs17670098-rs17670111
33	16	81970407	81970495	88	mh16KK-255 / MHA-18	Kidd et al. 2017 [17] / Phillips et al. 2019 [18]	yes - shortened on 5'	81970353	81970495	rs3934956	rs4073828	108,7	rs3934956-rs62044947-rs3934957-rs4073828
34	17	47287067	47287109	42	mh17KK-054	Kidd et al. 2017 [17]	no	-	-	rs2233362	rs634370	69,4	rs2233362-rs634370
35	17	52942428	52942456	28	mh17KK-272	Kidd et al. 2017 [17]	yes - shortened on 3'	52942428	52942558	rs2934897	rs7207239	75,6	rs2934897-rs7207239
36	17	76228571	76228630	59	mh17KK-055	Kidd et al. 2017 [17]	no	-	-	rs11868709	rs9907137	114,9	rs11868709-rs58233940-rs9907137
37	18	22137396	22137454	58	mh18KK-285	Kidd et al. 2017 [17]	yes - shortened on 5'	22137319	22137454	rs17187688	rs1945150	43,3	rs17187688-rs17187695-rs1945150
38	18	76089886	76089968	82	mh18KK-293 / MHA-20	Kidd et al. 2017 [17] / Phillips et al. 2019 [18]	no	-	-	rs621320	rs621766	114,6	rs621320-rs621340-rs678179-rs621766
39	20	16513260	16513342	82	mh20KK-307	Kidd et al. 2017 [17]	yes - shortened on 3'	16513260	16513400	rs6044080	rs6044081	38,8	rs6044080-rs17674942-rs6044081
40	21	21880158	21880231	73	mh21KK-315 / MHA-21	Kidd et al. 2017 [17] / Phillips et al. 2019 [18]	yes - shortened on 5'	21880086	21880231	rs6517970	rs6517971	16,8	rs6517970-rs8131148-rs6517971
41	21	43062859	43062929	70	mh21KK-320	Kidd et al. 2017 [17]	yes - shortened on 3'	43062859	43063044	rs2838081	rs2838082	52,6	rs2838081-rs2838082
42	21	46714641	46714707	66	mh21KK-324 / MHA-22	Kidd et al. 2017 [17] / Phillips et al. 2019 [18]	yes - shortened on 5'	46714549	46714707	rs2838868	rs8133697	60,7	rs2838868-rs7279250-rs8133697
43	22	19951207	19951271	64	mh22KK-060	Kidd et al. 2017 [17]	no	-	-	rs4818	rs4680	11	rs4818-rs8192488-rs4680
44	22	44763606	44763651	45	mh22KK-061	Kidd et al. 2017 [17]	yes - shortened on 3'	44763606	44763752	rs763040	rs5764924	54,8	rs763040-rs5764924

Supplementary Table S2. Summary of samples run in each laboratory and DNA inputs used for library preparation. NTC: non-template control

		Input DNA (ng)		
		Ion S5	MiSeq	GeneReader
Reference set	NA12878	10	40	1
	NA24143	1	40	1
	NA24149	1	40	1
	NA24385	1	40	1
	NA06994	1	40	1
	NA07000	1	40	1
	NA07029	1	40	1
	2391d-A*	1	1.6	1
	2391d-C*	1	1.6	1
	2800M*	1	1	1
Dilution series	2800M*	1	1	1
	2800M	0.500	0.500	0.500
	2800M	0.250	0.250	0.250
	2800M	0.125	0.125	0.125
	2800M	0.062	0.062	0.062
	2800M	0.031	0.031	0.031
	2800M	0.015	0.015	0.015
	2800M	0.007	0.007	0.007
Mixture analysis	2391d-A*	1	1.6	1
	2391d-C*	1	1.6	1
	2391d-D	1	1.5	1
Negative control	NTC	-	-	-

*The 2800M control DNA (1 ng), 2391d-A and 2191d-C were analysed once but used for different analyses

Supplementary Table S3. Mean, standard deviation and median percentage values of base misincorporation obtained for the reference set above a 2% rate.

Platform	rs number	mean	sd	median
Ion S5	rs9933604	11,76	4,66	9,94
	rs2184289	11,09	5,99	11,02
	rs636967	9,39	4,84	8,82
	rs7979745	7,83	2,13	7,79
	rs250132	7,47	2,93	8,26
	rs4920562	7,40	3,35	6,47
	rs4585689	7,10	5,99	5,43
	rs2843150	6,97	7,96	5,35
	rs80266372	6,85	1,82	7,62
	rs11103107	6,62	7,06	3,74
	rs13436899	6,14	7,55	1,28
	rs4335841	5,58	2,48	6,45
	rs1220576	5,18	2,94	4,75
	rs13189636	5,13	6,06	2,24
	rs3001118	5,10	3,92	4,92
	rs10013748	5,05	2,50	4,31
	rs2078304	5,01	3,45	5,22
	rs6656830	5,00	4,33	2,38
	rs1393073426	4,97	1,96	4,92
	rs7722456	4,57	2,28	5,42
	rs11856402	4,49	3,38	4,47
	rs2293883	4,39	4,36	4,24
	rs7457014	4,18	2,28	3,61
	rs13172903	3,95	3,13	3,23
	rs10956402	3,91	3,12	3,08
	rs1033381	3,81	2,64	5,14

Platform	rs number	mean	sd	median
MiSeq	rs284842	6,09	1,40	5,56
	rs6764984	4,87	1,58	4,62
	rs4958482	4,44	2,34	4,73
	rs1097454	4,20	4,42	4,55
	rs4991456	3,86	6,82	0,43
	rs2018579	3,39	1,81	3,62
	rs60216576	3,38	1,60	2,79
	rs5754837	3,37	2,78	2,84
	rs13172903	2,79	2,72	1,59
	rs10795925	2,62	1,84	3,48
	rs13136917	2,17	2,87	0,81
	rs58318997	2,15	5,56	0,00
	rs370892	2,13	2,86	0,75
	rs735480	2,12	2,19	1,20
	rs2104668	2,05	2,24	1,26

Platform	rs number	mean	sd	median
GeneReader	rs2151695	8,29	6,08	5,32
	rs6764984	5,60	3,53	6,26
	rs515687	4,55	6,43	4,55
	rs66525445	3,81	2,13	3,57
	rs4991456	3,80	7,43	0,00
	rs7481235	3,68	3,06	4,55
	rs10795925	3,49	3,10	3,13
	rs58318997	3,34	3,48	3,17
	rs7623855	3,24	3,57	2,44
	rs6854973	3,01	3,91	0,00
	rs12045165	3,01	3,55	2,08
	rs1809429	2,80	4,38	0,00
	rs1762533	2,44	2,27	2,86
	rs1426654	2,38	2,24	1,46
	rs7593421	2,27	1,98	2,04
	rs1965430	2,27	1,25	2,10
	rs60216576	2,21	2,52	1,85
	rs822709	2,17	3,59	1,33

Platform	rs number	mean	sd	median
Ion S5	rs6835428	3,73	2,66	4,60
	rs7119007	3,68	1,67	3,45
	rs481258	3,66	4,81	0,66
	rs6493174	3,63	1,86	3,87
	rs370892	3,32	2,65	2,83
	rs9539946	3,24	3,66	1,79
	rs4257030	3,22	3,08	3,65
	rs2517863	3,16	1,76	3,44
	rs6950803	3,16	1,89	3,51
	rs58409270	3,15	3,06	1,97
	rs10852023	3,04	3,13	1,75
	rs634496	3,01	1,56	3,02
	rs117273376	2,99	2,03	2,54
	rs7520324	2,96	2,57	2,55
	rs57394812	2,89	4,10	1,18
	rs1462906	2,88	2,48	2,25
	rs2080362	2,82	5,90	0,34
	rs1338791	2,80	2,61	1,82
	rs417510	2,77	3,30	1,64
	rs4906213	2,76	2,93	2,33
	rs7929468	2,72	4,53	0,00
	rs9453456	2,71	2,36	2,82
	rs4991456	2,67	2,85	2,33
	rs4381739	2,66	2,49	2,58
	rs11658628	2,63	2,65	1,75
	rs6764984	2,63	5,97	0,66
	rs2188803	2,62	2,01	2,93
	rs9511467	2,62	2,65	1,61
	rs60216576	2,55	3,84	1,41

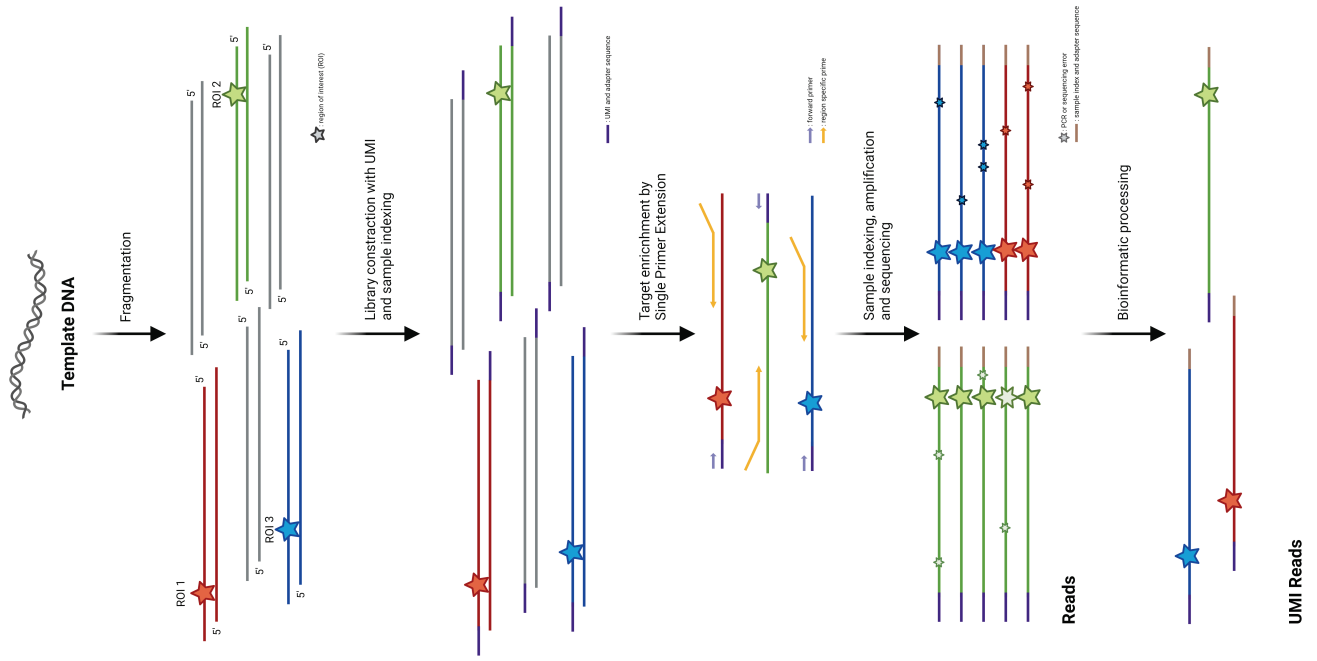
Platform	rs number	mean	sd	median
Ion S5	rs2052215	2,55	1,91	2,13
	rs6141024	2,53	2,36	1,82
	rs1370356	2,51	2,88	2,25
	rs1188423	2,49	5,16	0,00
	rs10176767	2,48	2,66	1,97
	rs2304485	2,46	3,27	0,00
	rs2456668	2,44	2,41	2,32
	rs1762533	2,38	6,30	0,00
	rs6923511	2,29	3,05	0,57
	rs5908228	2,24	2,40	1,50
	rs1739838	2,23	2,12	1,95
	rs11781355	2,22	1,95	1,96
	rs7034102	2,16	1,42	1,35
	rs425158	2,15	2,58	1,00
	rs5845678	2,08	2,41	1,20
	rs360676	2,01	2,77	0,00

Supplementary Table S4. Summary of underperforming markers.

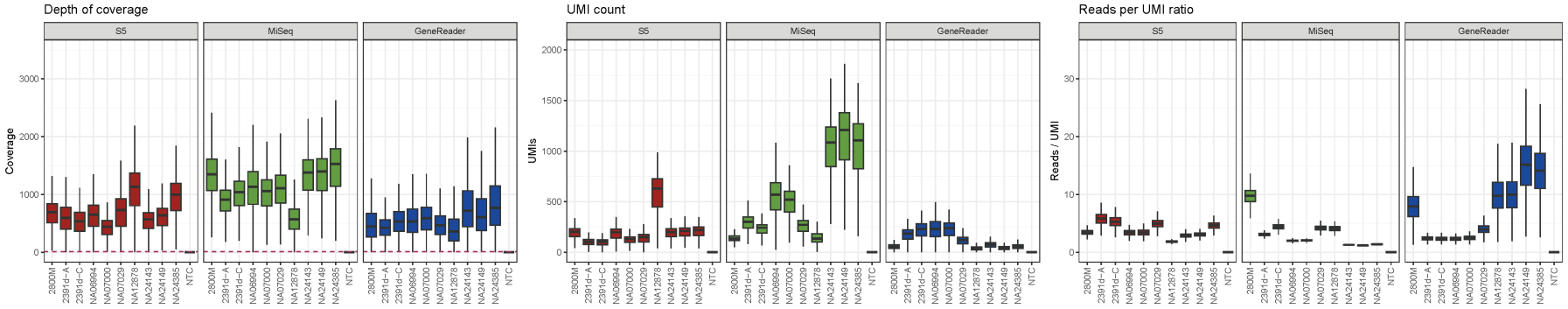
rs number	Chr	Position (hg19)	Marker type	Platform	Observations	Recommendation
rs58318997	7	38290763	Identity tri-allelic SNP	All	High misincorporation, imbalanced allele read frequency and discordances observed due to possible off-target amplification	Remove from analyses
rs80266372	1	149310808	Identity tri-allelic SNP	All	Discordances observed due to possible off-target amplification	Remove from analyses
rs2009433	22	23247082	Identity tri-allelic SNP	All	Mendelian incompabilities observed	Remove from analyses
rs399737	22	22747327	Identity tri-allelic SNP	All	Mendelian incompabilities observed	Remove from analyses
rs4991456	18	5972926	Identity tri-allelic SNP	All	High levels of misincorporation observed	Manual curation of genotype
rs9933604	16	25512659	Identity tri-allelic SNP	Ion S5	High levels of misincorporation observed	Manual revision of the alignment
rs2184289	14	83267990	Identity tri-allelic SNP	Ion S5	High levels of misincorporation observed	Manual revision of the alignment
rs2843150	1	2247744	Identity tri-allelic SNP	Ion S5	High levels of misincorporation observed	Manual revision of the alignment
rs319305	7	95602845	Identity tri-allelic SNP	Ion S5	Discordances observed due to incorrect alignment	Manual revision of the alignment
rs6764984	3	84523763	Identity tri-allelic SNP	Ion S5	Discordances observed due to incorrect alignment	Manual revision of the alignment
rs7623855	3	66856941	Identity tri-allelic SNP	All	Imbalanced allele read frequence. Discordances observed in GeneReader due to allele drop-out	Manual curation of genotype
rs1809429	4	3942139	Identity tri-allelic SNP	MiSeq and GeneReader mainly	Discordances observed due to off-target read alignment	Remove from analyses
rs7657799	4	105375423	Ancestry SNP	MiSeq	High levels of coverage observed for non-template control	Manual curation of genotype
rs784685	9	109904752	Identity tri-allelic SNP	MiSeq	High levels of coverage observed for non-template control	Manual curation of genotype
rs11531789	9	111520933	Identity tri-allelic SNP	MiSeq	High levels of coverage observed for non-template control	Manual curation of genotype
rs10147017	14	80872905	Identity tri-allelic SNP	MiSeq	High levels of coverage observed for non-template control	Manual curation of genotype
rs7629097	3	13195993	Identity tri-allelic SNP	All	Discordances observed with 1000 Genomes data	None
rs4461915	14	106925184	Identity tri-allelic SNP	All	Discordances observed with 1000 Genomes data	None
rs6478720	9	128789376	Identity tri-allelic SNP	Ion S5	Mendelian incompabilities observed due to incorrect alignment	Manual revision of the alignment
rs66525445	21	9417961	Identity tri-allelic SNP	All	Imbalance allele read frequency in heterozygotes	Manual curation of genotype
rs9617584	22	16955079	Identity tri-allelic SNP	All	Imbalance allele read frequency in heterozygotes	Manual curation of genotype
rs1097454	3	120617060	Identity tri-allelic SNP	All	Imbalance allele read frequency in heterozygotes	Manual curation of genotype
rs370892	21	18264580	Identity tri-allelic SNP	All	High levels of misincorporation (~2-3%) and imbalance allele read frequency in heterozygotes	Manual curation of genotype
rs10151146	14	106106549	Identity tri-allelic SNP	All	Imbalance allele read frequency in heterozygotes	Manual curation of genotype
rs4906213	14	102991021	Identity tri-allelic SNP	All	Imbalance allele read frequency in heterozygotes	Manual curation of genotype
rs8029886	15	99829867	Identity tri-allelic SNP	All	Imbalance allele read frequency in heterozygotes	Manual curation of genotype

rs10795925	10	12098836	Identity tri-allelic SNP	All	High levels of misincorporation (~2-3%) and imbalance allele read frequency in heterozygotes	Manual curation of genotype
rs11103107	9	138520914	Identity tri-allelic SNP	All	Imbalance allele read frequency in heterozygotes	Manual curation of genotype

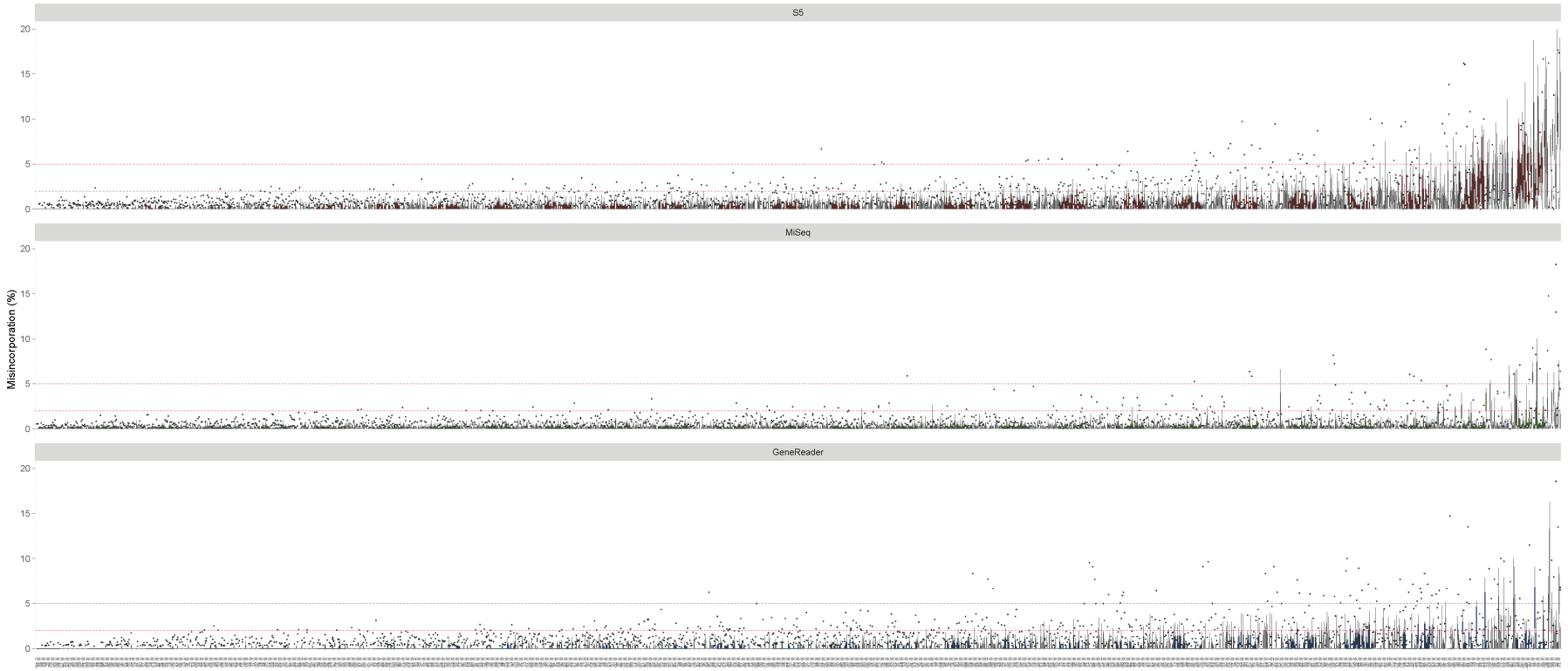
ANNEX II. SUPPLEMENTARY FIGURES FOR ARTICLE 6



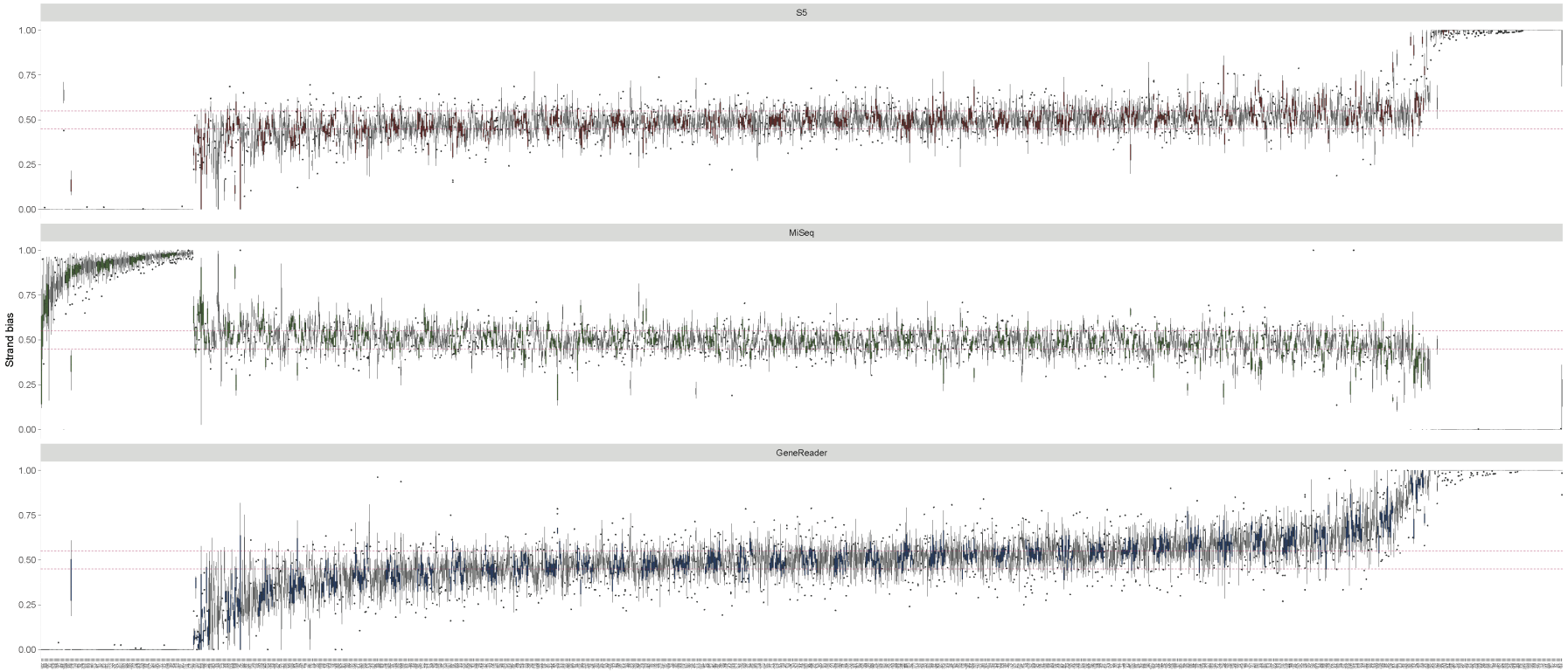
Supplementary Fig. S1. Simplified and schematic representation depicting the UMI principle. Adapted from the QIAseq™ Targeted DNA Panel Handbook with BioRender.com.



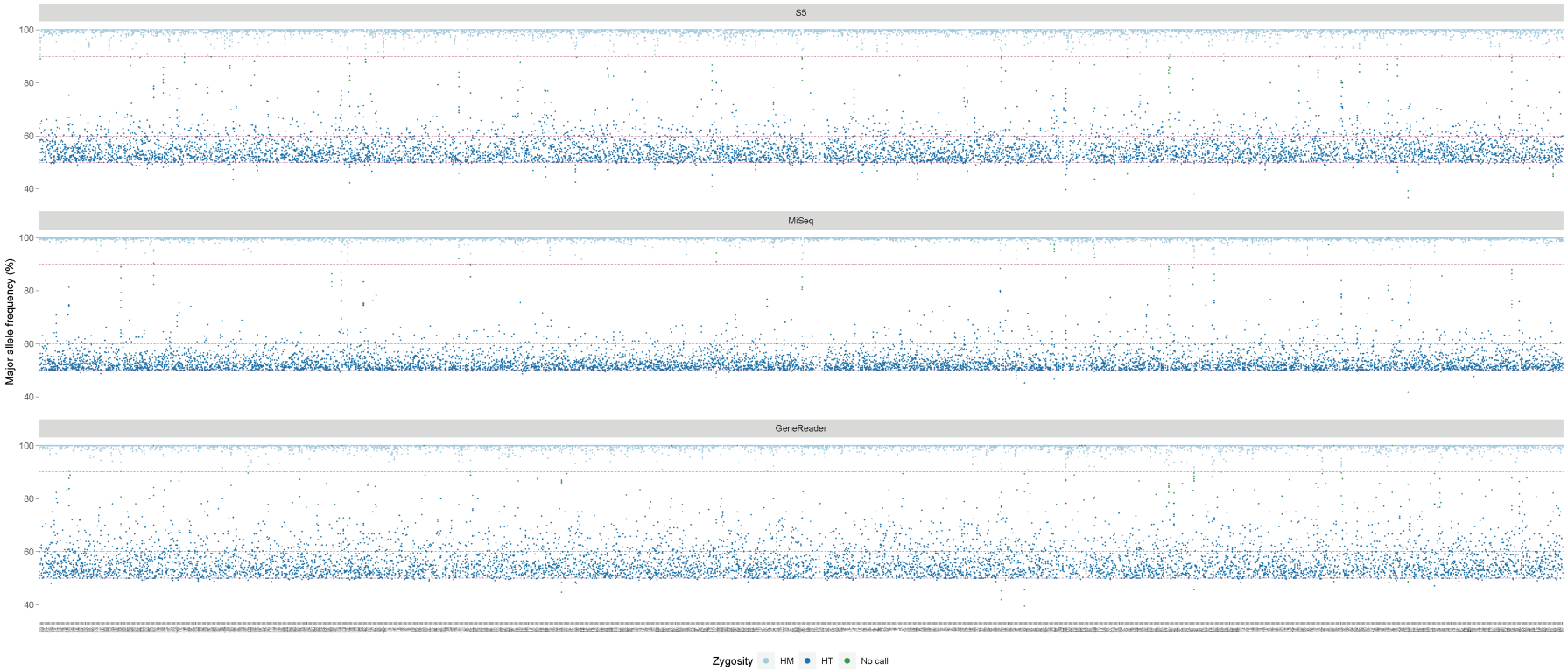
Supplementary Fig. S2. Depth of coverage, UMI count and reads per UMI rate distributions for the reference sample set.
Maroon dashed line indicates the 10 UMI calling threshold.



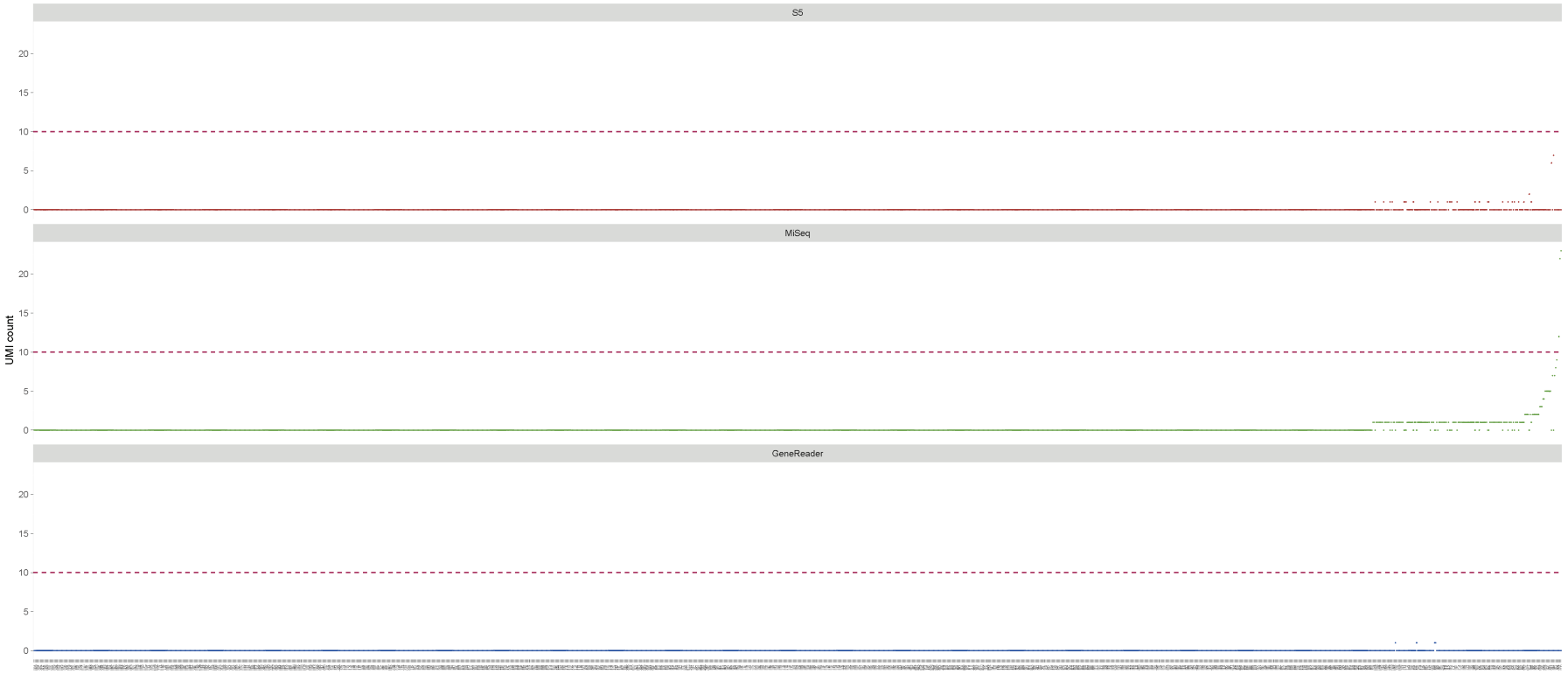
Supplementary Fig. S3. Allele misincorporation rate distribution per platform.
Maroon dashed lines represents the 2 and 5% misincorporation threshold.



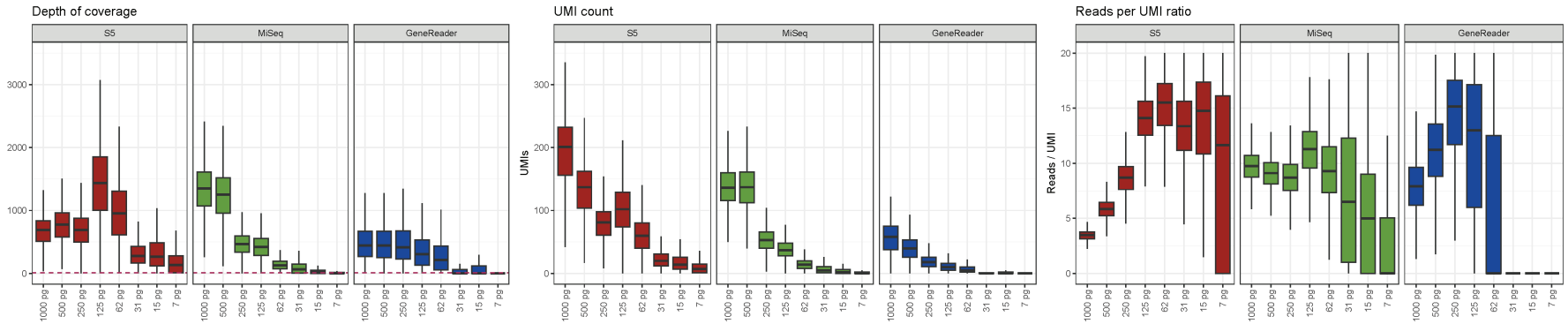
Supplementary Fig. S4. Strand bias boxplots per marker for each platform (UMI reads on the forward strand / reverse strand).
Maroon dashed lines represent the 45−55% interval



Supplementary Fig. S5. Major allele frequency distribution (reads for the most frequent allele divided by total coverage) for the reference samples. Maroon dashed lines represent the ideal 50−60% and >90% intervals for homozygous (HM) and heterozygous (HT) calls, respectively.



Supplementary Fig. S6. UMI count per marker for each platform in the non−template controls.
Maroon dashed line represents the 10 UMI calling threshold.



Supplementary Fig. S7. Depth of coverage, UMI count and reads per UMI rate distributions for the dilution series.
Maroon dashed line indicates the 10 UMI calling threshold.

ANNEX III. SUPPLEMENTARY FILE 1 FOR ARTICLE 6

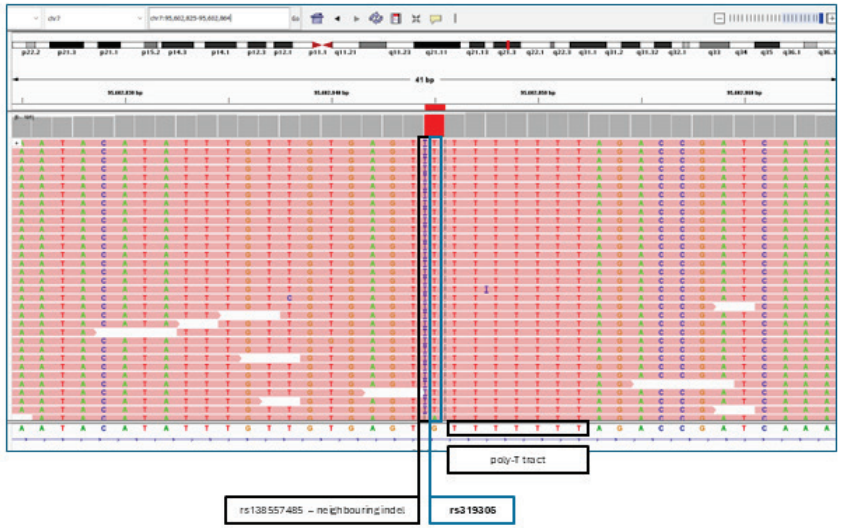
Supplementary File S1.

IGV captures of markers displaying probable alignment artifacts

rs319305

A poly-T tract and an A-insertion on both sides of the SNP prevents the correct interpretation on the Ion S5 reads.

GeneReader alignment



Ion S5 alignment



rs6764984

A poly-A tract causes misalignments in the Ion S5 reads. In addition, the presence of other variants in the nearby region hinders the correct interpretation

GeneReader alignment



Ion S5 alignment



rs1809429

Allele drop-outs found in GeneReader and MiSeq, sequence reads were found aligning to off-target sequence.

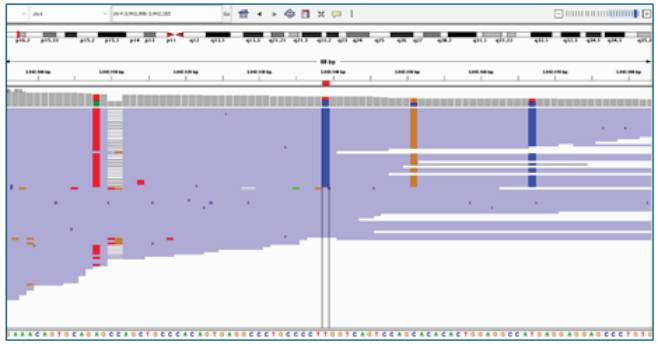
GeneReader alignment



rs183110408 (reads in reverse)



Ion S5 alignment

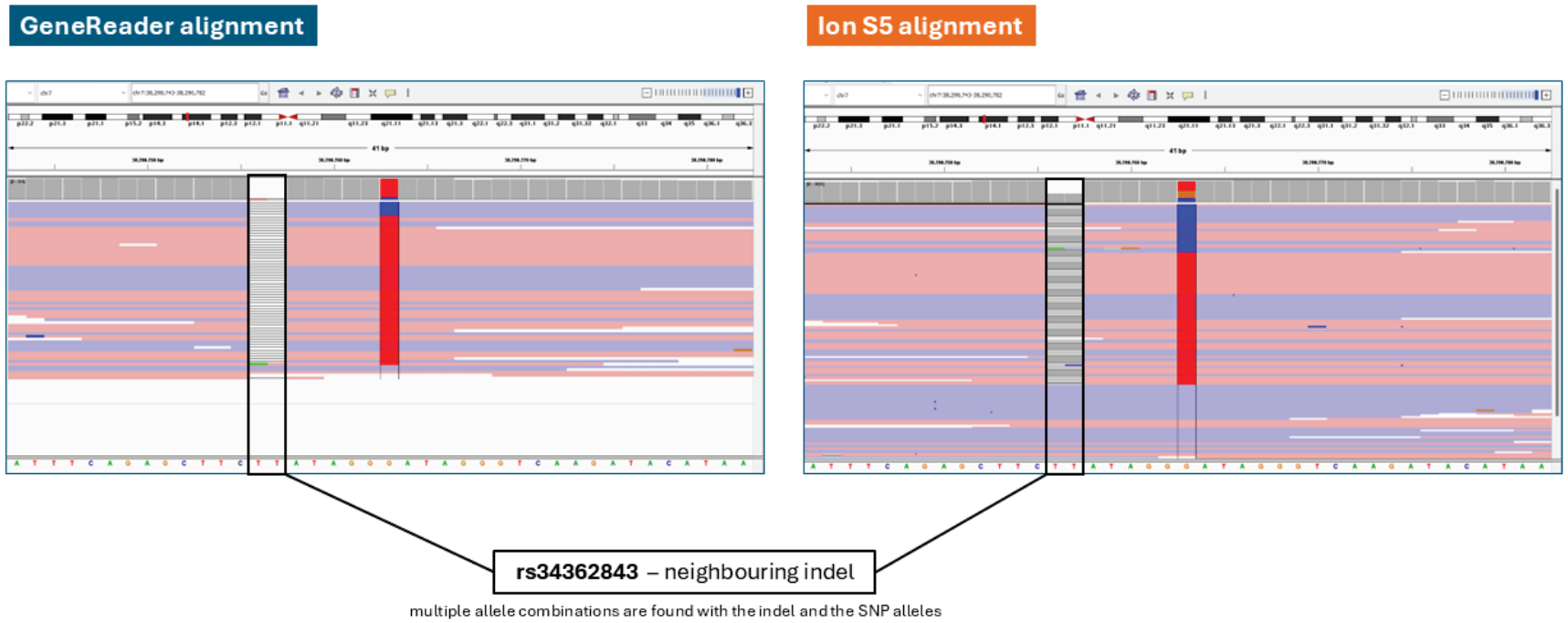


rs183110408 (reads in reverse)



rs58318997

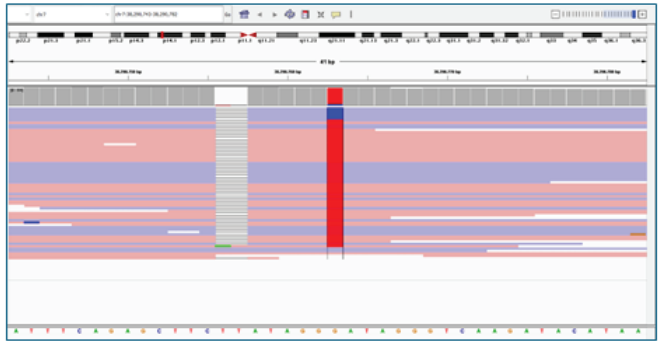
Multiple alleles were found for single-source samples. In addition, reads were observed aligning off-target.



rs58318997 (cont.)

Multiple alleles were found for single-source samples. In addition, reads were observed aligning off-target.

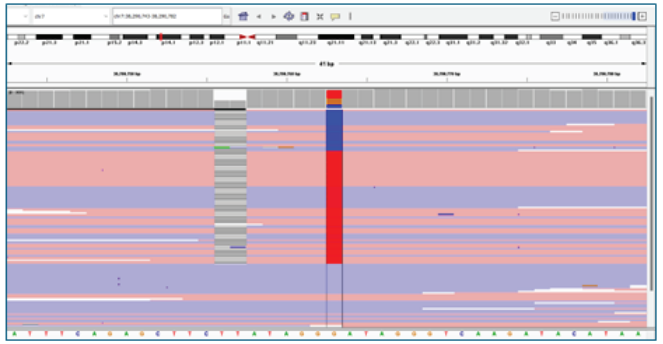
GeneReader alignment



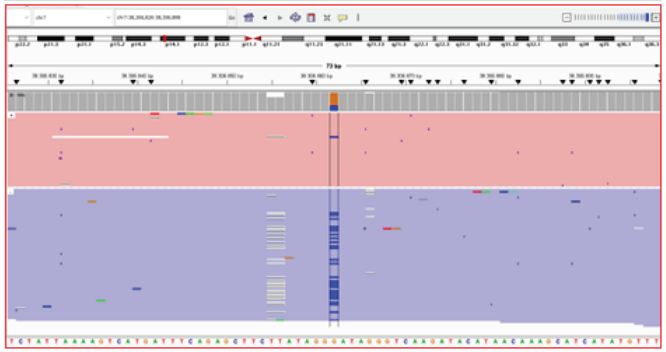
rs201552694



Ion S5 alignment



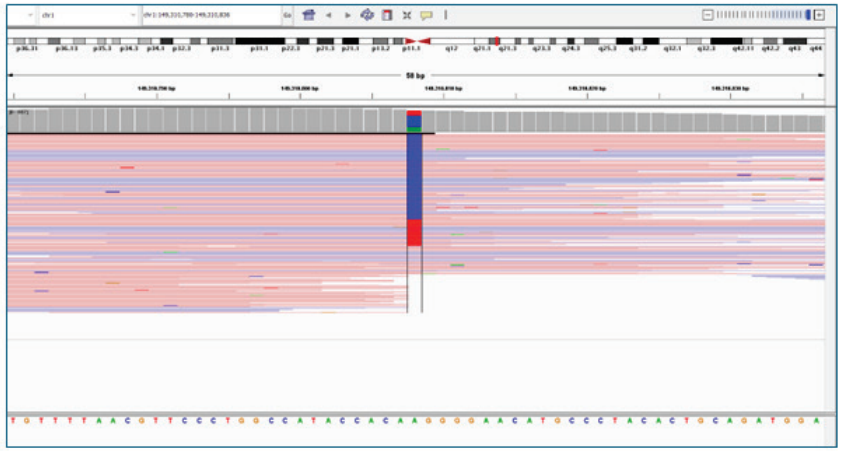
rs201552694



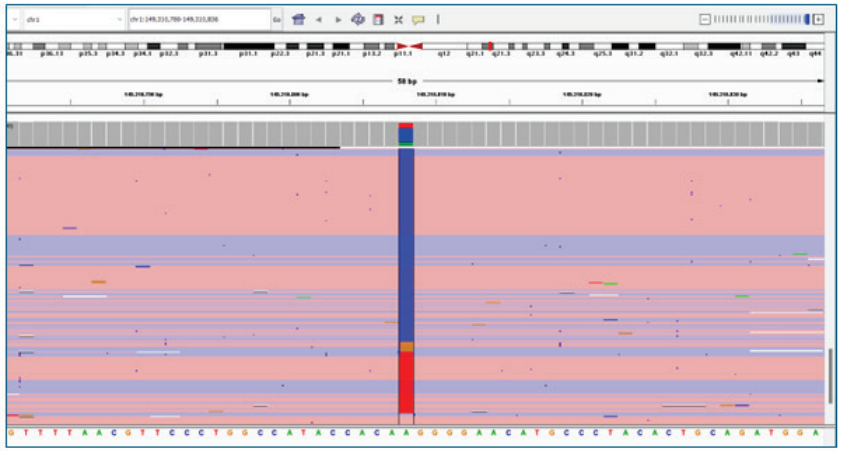
rs80266372

Multiple alleles were found for single-source samples. In addition, reads were observed aligning off-target and G insertions on Ion S5 likely due to a poly-G tract.

GeneReader alignment



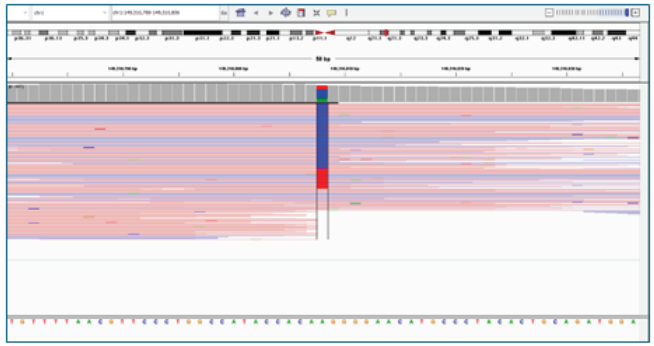
Ion S5 alignment



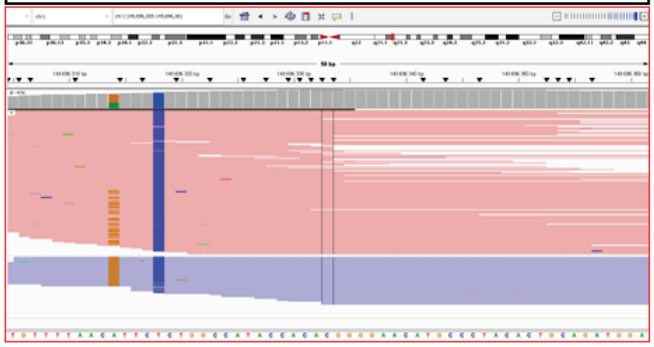
rs80266372 (cont.)

Multiple alleles were found for single-source samples. In addition, reads were observed aligning off-target and G insertions on Ion S5 likely due to a poly-G tract.

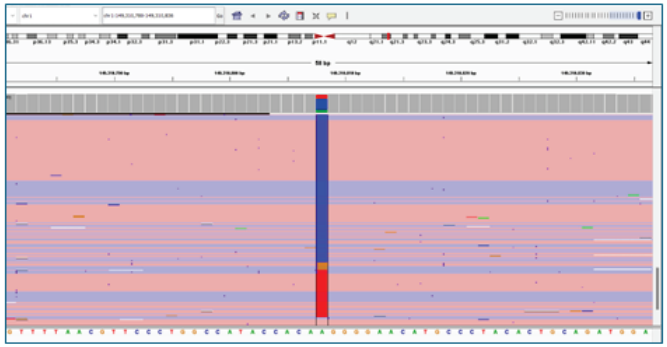
GeneReader alignment



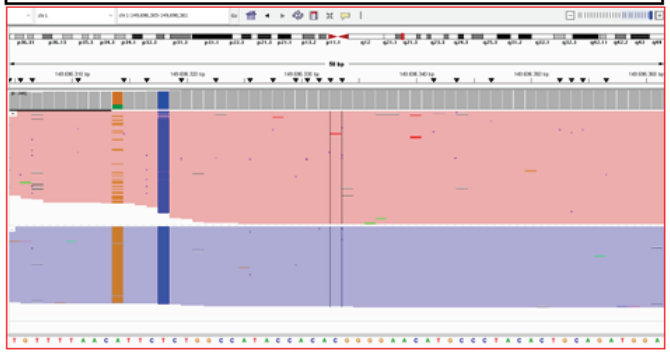
rs1253935158



Ion S5 alignment



rs1253935158







Forensic genetics analyses usually imply the use of reduced sets of markers to maintain the sensibility. The shift towards the use of Massively Parallel Sequencing methods in the forensic field entails the expansion of the number of markers that can be analyzed. Within this Doctoral Tesis, the development and validation of a total of five new marker sets that expand the range of tools available for forensic analyses is described.