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

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

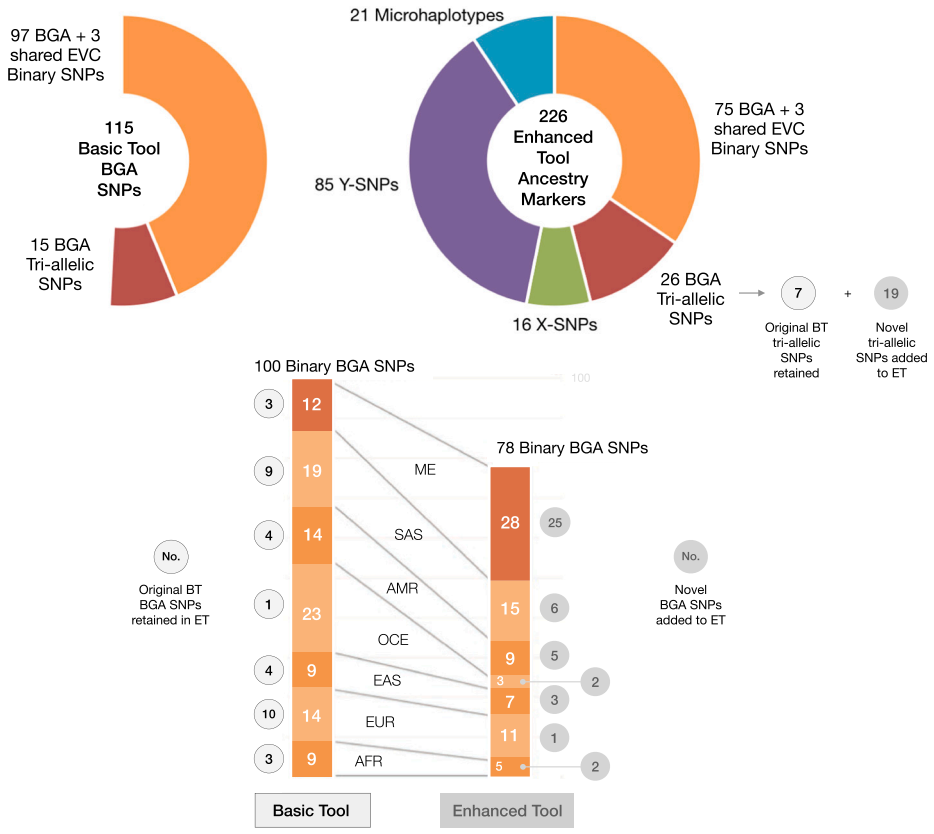

Fig. 1. Proportion of BGA SNPs and ancestry markers in the VISAGE Basic Tool (BT) and the VISAGE Enhanced Tool (ET). Amongst the binary autosomal BGA SNPs, all population-indicative sets were reduced in number, apart from Middle East (ME) informative SNPs, which were more than doubled in number. The 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*)

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

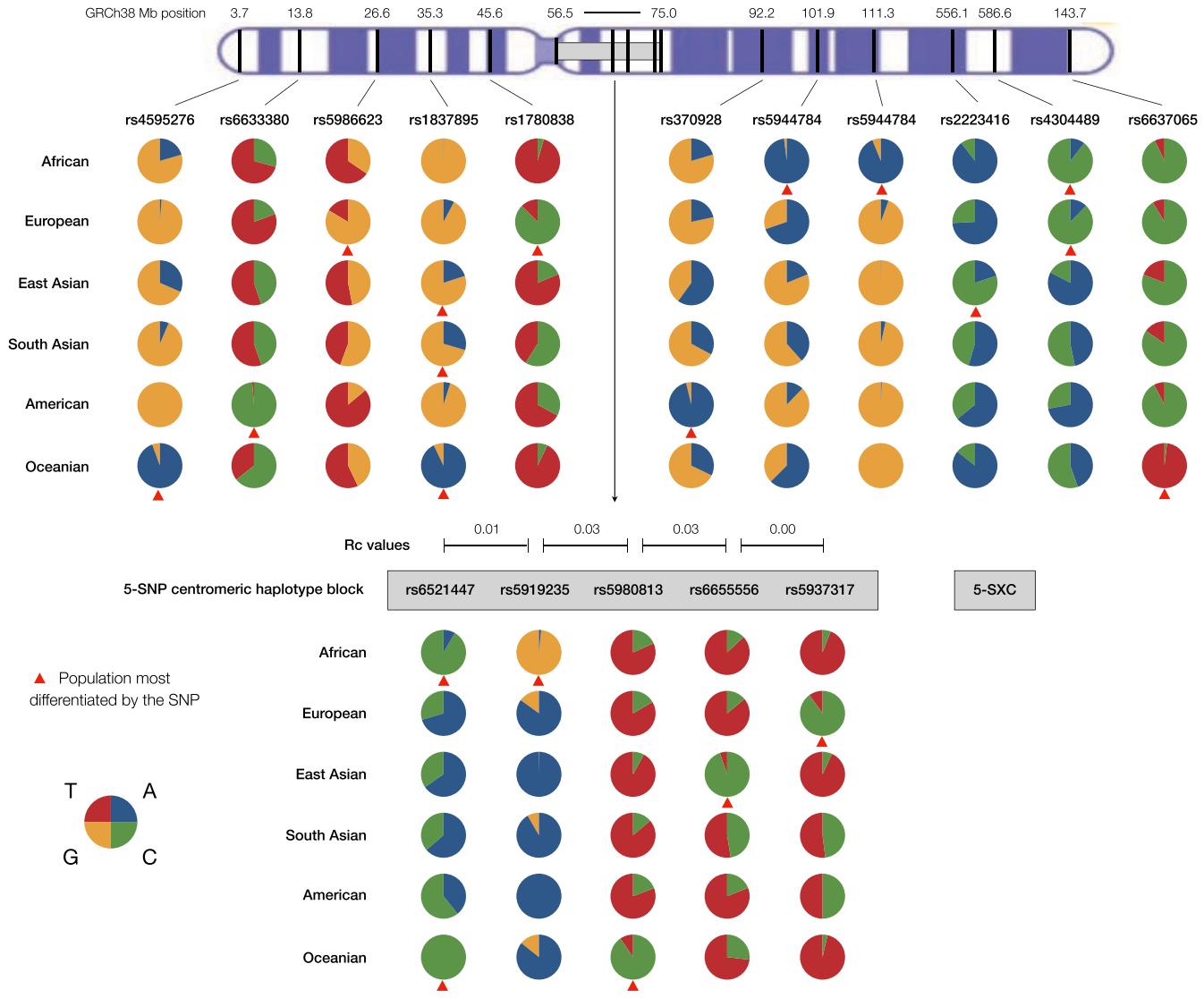

Fig. 2. X chromosome ideogram showing the positions of 16 ancestry-informative X-SNPs selected for ET. The grey bar on the 3′ side of the centromere defines the position of the 5-SNP X centromeric haplotype block (5-SXC) with estimated recombination fraction (Rc) values between the five component X-SNPs 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

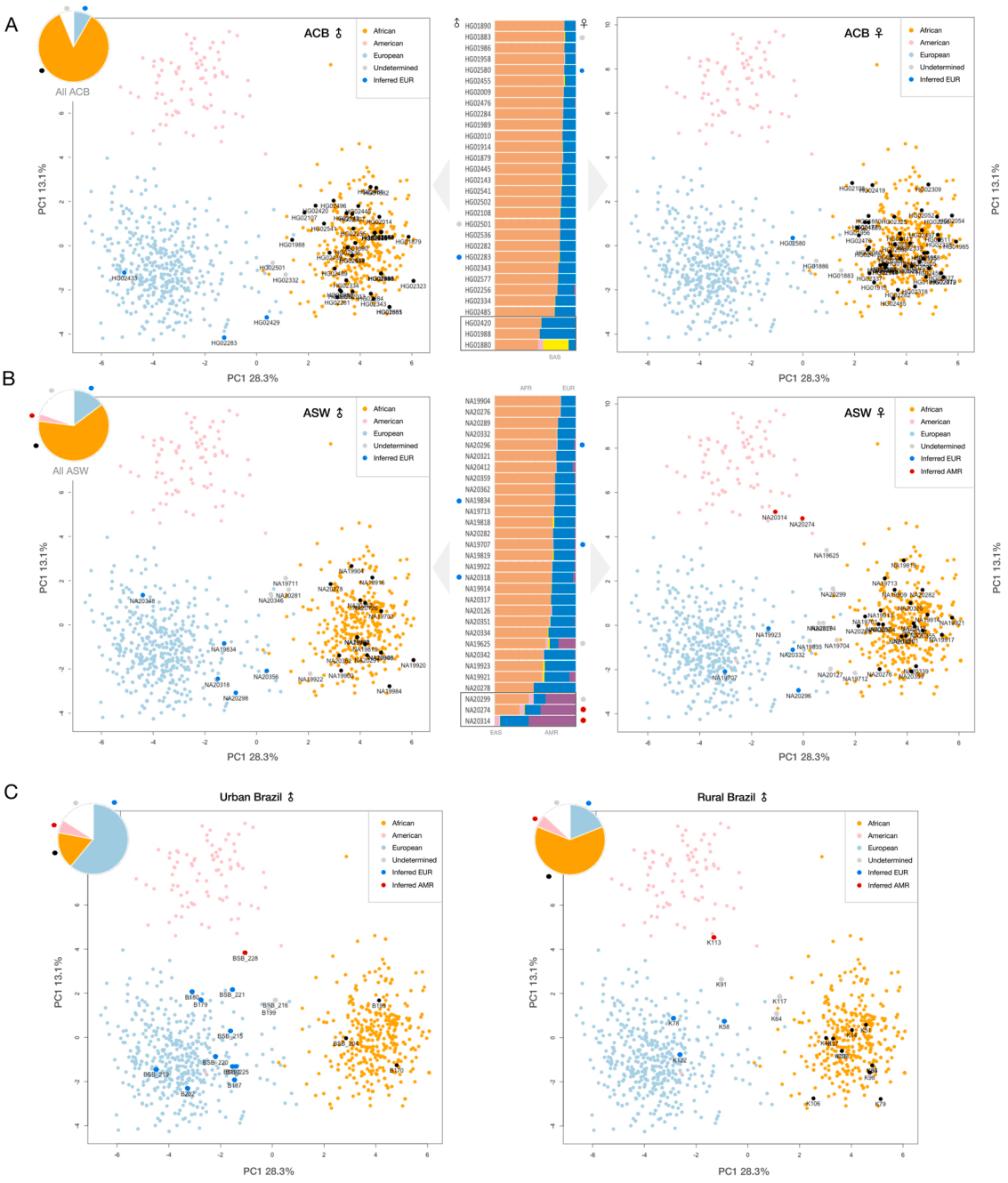

Fig. 3. PCA plots of AFR, EUR, AMR reference populations and admixed African populations from 1000 Genomes, using 16 X-SNP genotype data. **3 A**: African Caribbeans from Barbados (ACB); **3B**: African Americans from SW USA (ASW), each with male and female separately analysed; **3 C**: VISAGE Study 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.

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

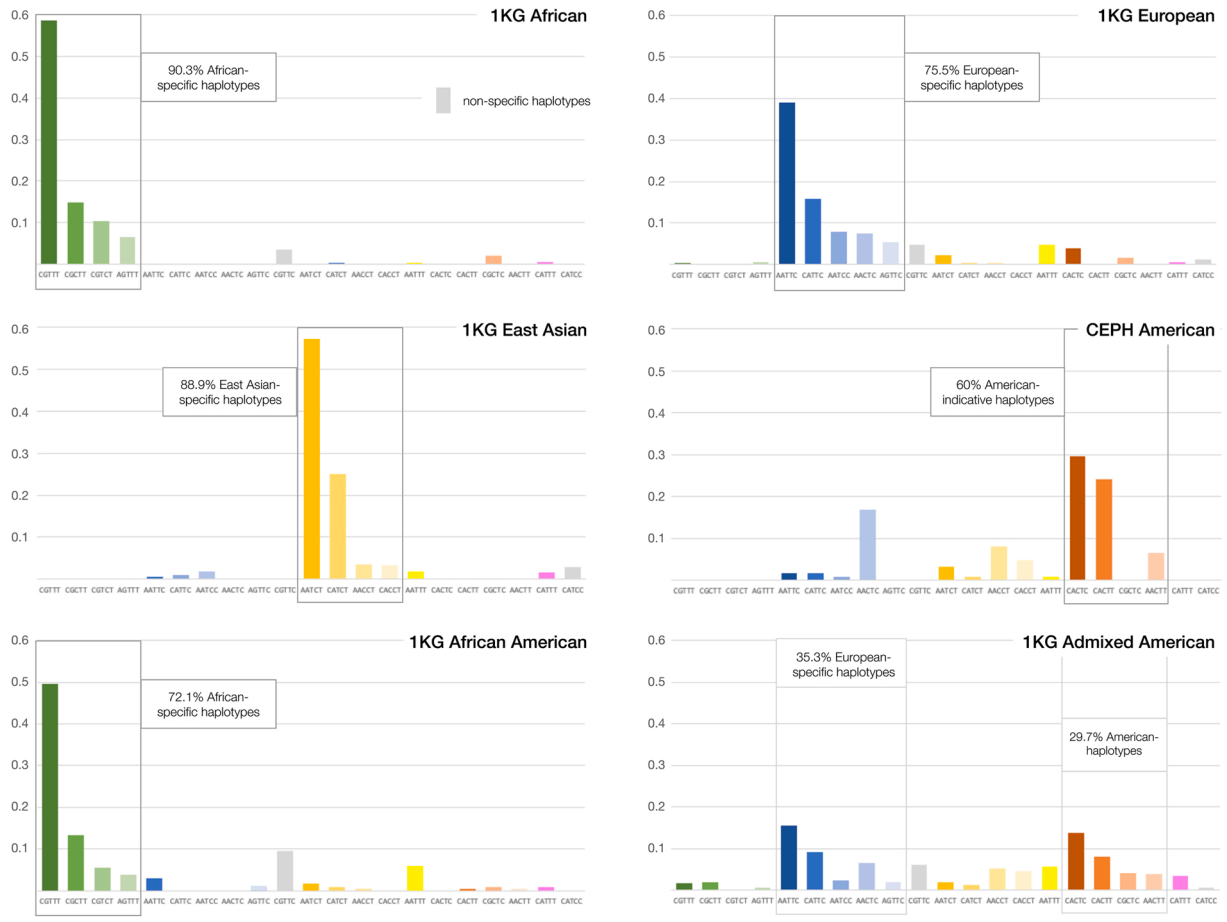

Fig. 4. Haplotype frequency estimates for the 5-SNP X centromeric haplotype block (5-SXC) in 1000 Genomes AFR, EUR, SAS, EAS, admixed African and admixed American populations, plus CEPH AMR (includes 18 1KG PEL), CEPH OCE and Study Fijians (to allow comparison with the relatively small sample size for 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

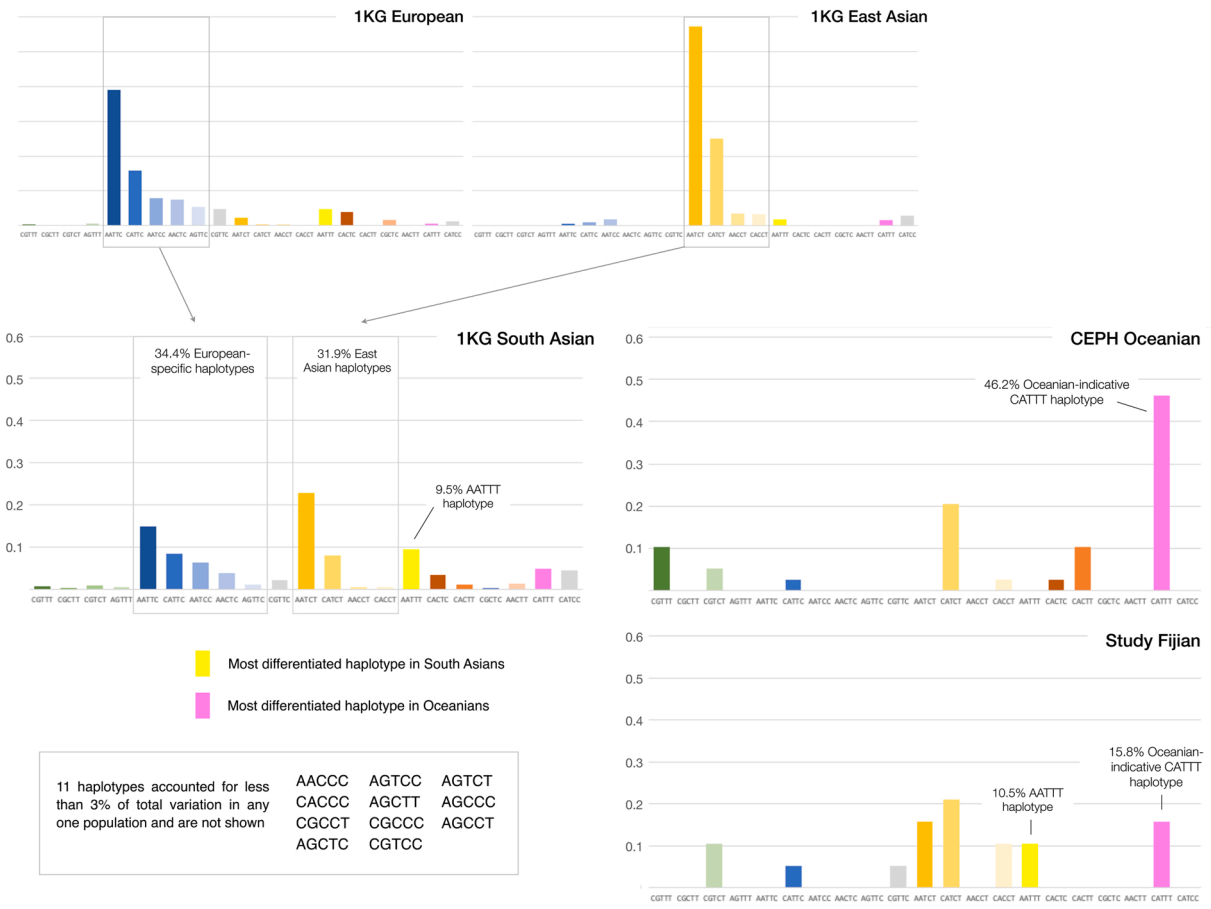

Fig. 4. (*continued*).

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

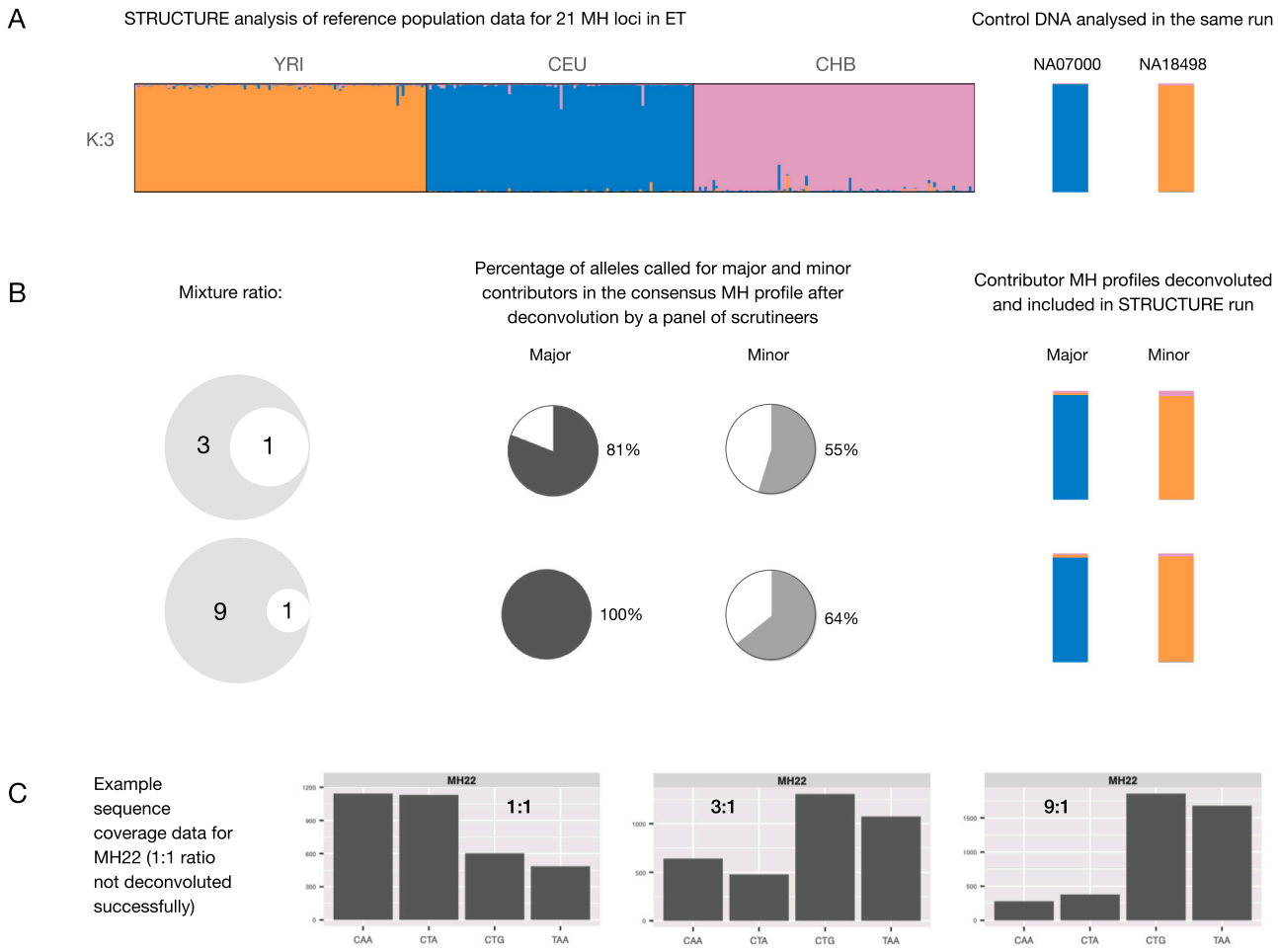

Fig. 5. Pilot study to evaluate the ability of the 21 Microhaplotypes of ET to detect mixed DNA and identify the ancestry of contributors in simple 2-way mixtures. **5 A:** STRUCTURE analysis of AFR, EUR, EAS reference populations indicates the 21 MHs differentiate these populations efficiently and MH 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-

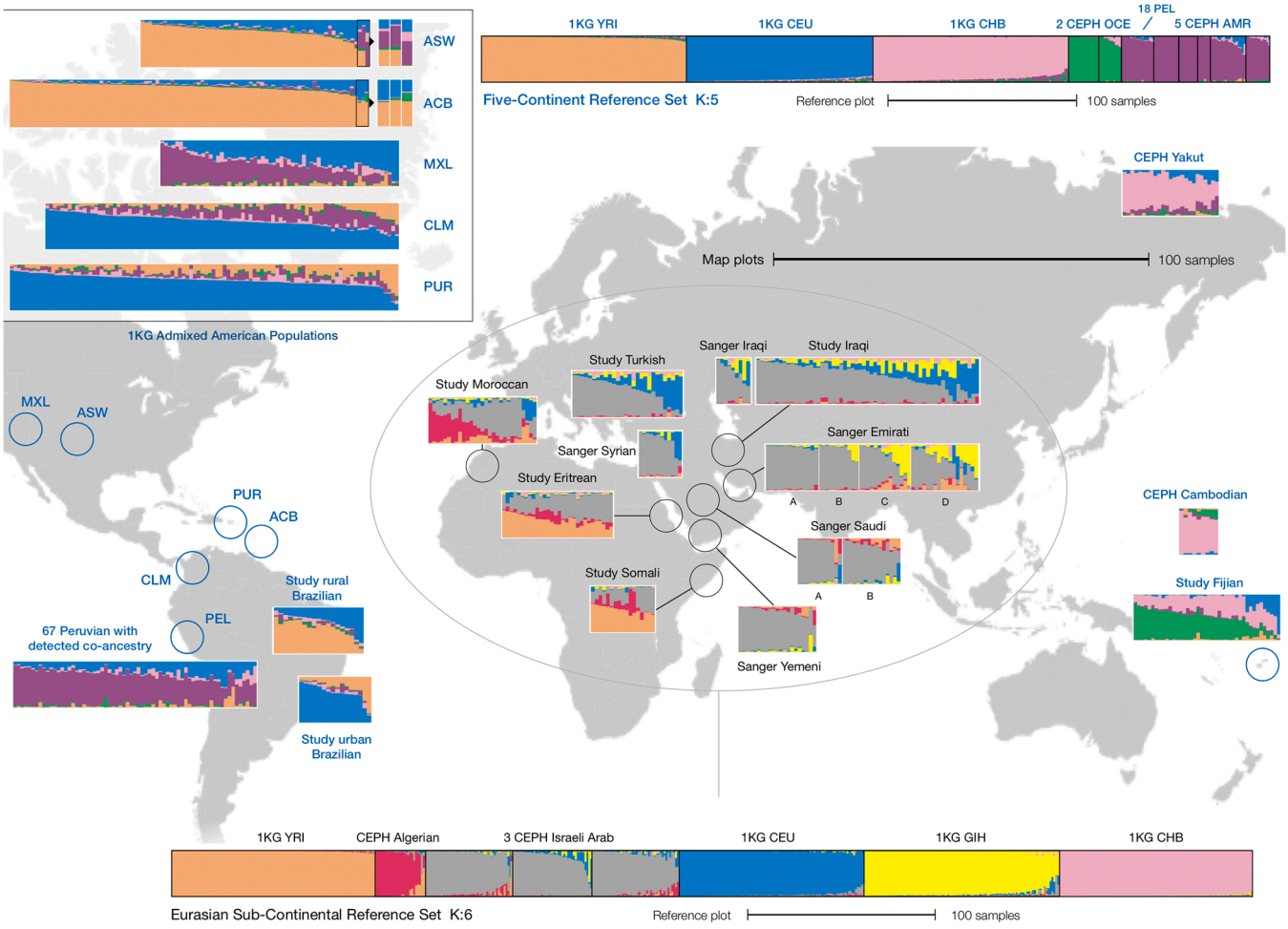

Fig. 6. Cluster plots of STRUCTURE analyses of selected 1KG, CEPH, Sanger ME and Study populations. Nested STRUCTURE analyses consisted of first stage K:5 runs using Five-Continental reference population datasets (POPFLAG=1) comprising 1KG AFR (YRI); EUR (CEU); EAS (CHB); 2 CEPH OCE populations; 5 CEPH 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 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). 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.