Forensic Science International: Genetics 46 (2020) 102232

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

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

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

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

ARTICLE INFO

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

ABSTRACT

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

1. Introduction

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

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

https://doi.org/10.1016/j.fsigen.2020.102232
Received 3 December 2019; Received in revised form 29 December 2019; Accepted 2 January 2020
Available online 17 January 2020

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

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

2. Materials and methods

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

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

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

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

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

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

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

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

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

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

2.3. Sequence quality checks of MPS Panel candidate SNPs

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

2.4. Kinship testing simulations

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

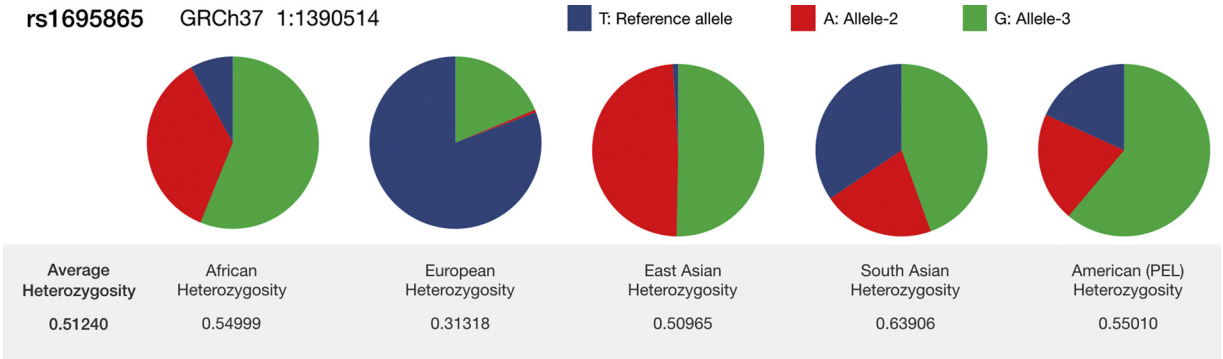

Fig. 1. The tri-allelic SNP rs1695865 has an overall heterozygosity of 0.5124, which is higher than the binary SNP maximum value of 0.5. Although listed as a potential candidate for the MPS Panel, it has a highly imbalanced allele frequency distribution and a much lower than average European heterozygosity,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

estimated using the method described by Kruijver [28].

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

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

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

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

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

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

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

3. Results and discussion

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

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

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

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

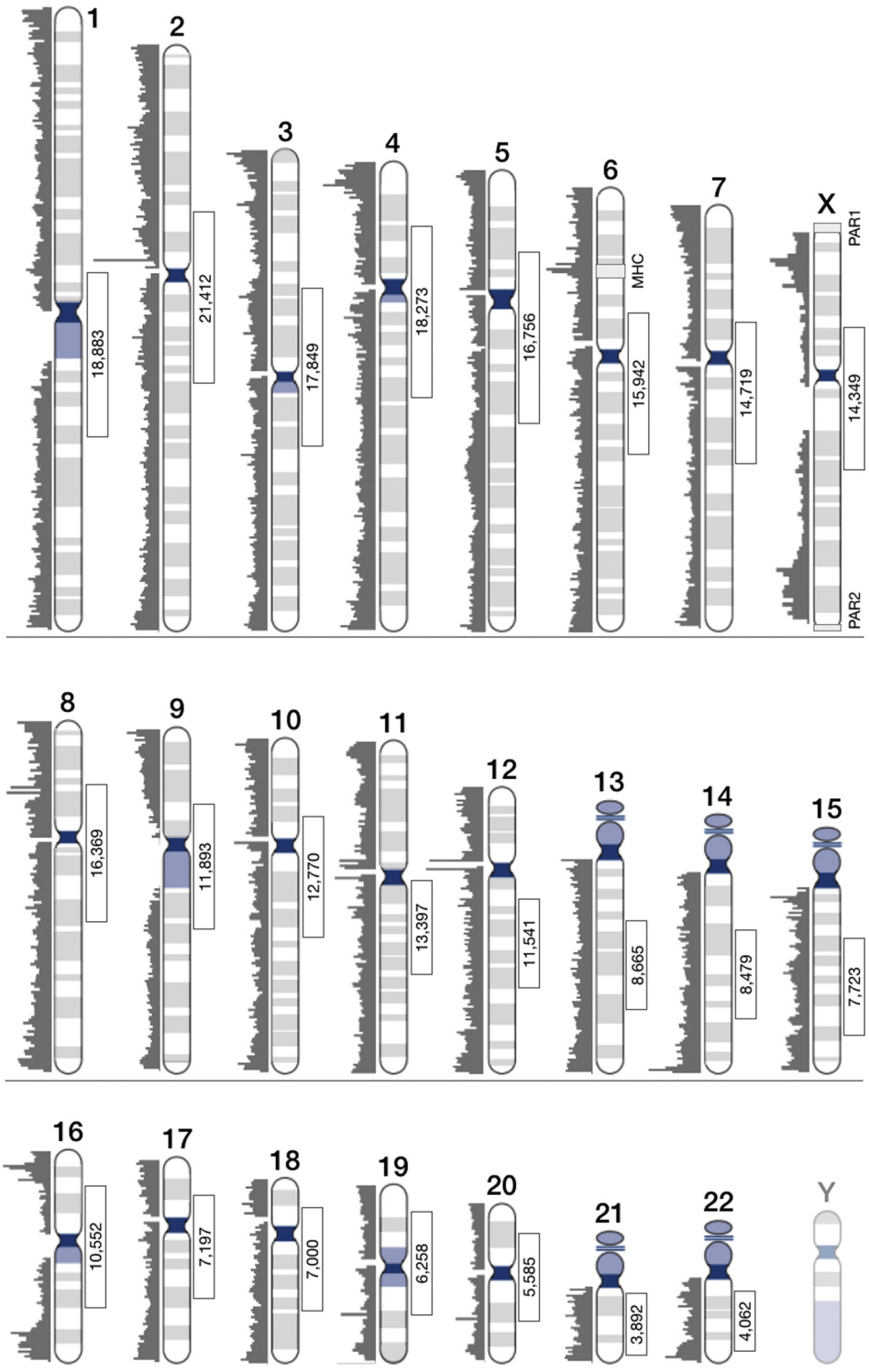

Fig. 2. Density plots of multiple-allele SNPs identified in the human genome. Each dark gray bar on the left represents total SNP counts per megabase of chromosome sequence. Y SNP data was not collected. Dark blue regions are facultative heterochromatin; light blue constitutive heterochromatin. Regions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

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

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

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

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

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

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

3.2. Relationship test simulations

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

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

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

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

3.2.2. Binary SNPs vs tri-allelic SNPs

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

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

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

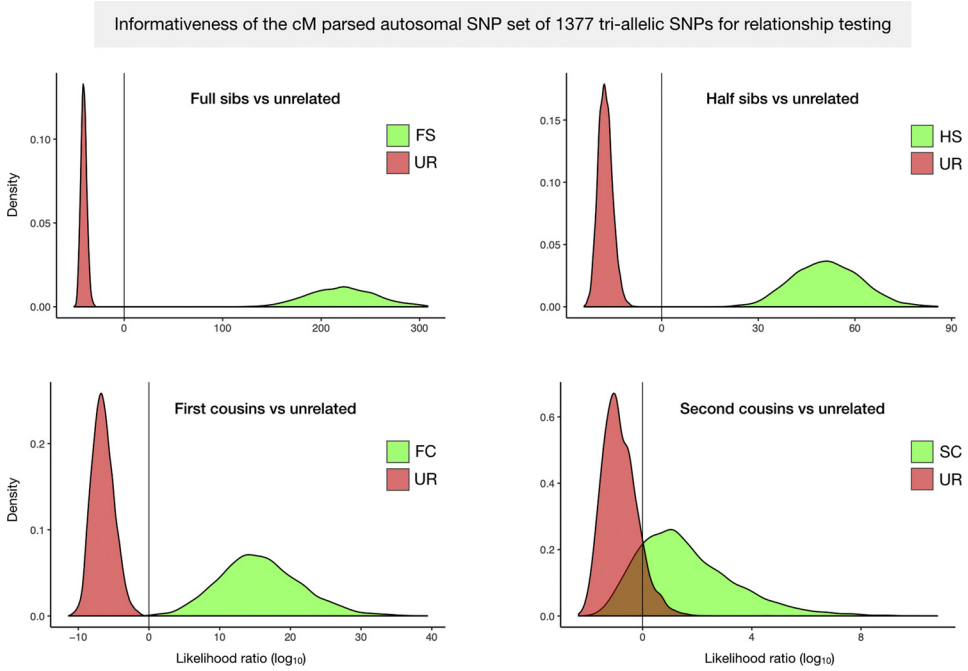

Fig. 3. Log LR distribution plots obtained from 5,000 simulations of relationship tests using European allele frequencies of 1,411 autosomal tri-allelic SNPs compiled for the MPS panel (i.e. after cM parsing but including SNPs discarded for sequence quality reasons). Relationship tests comprised full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

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

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

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

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

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

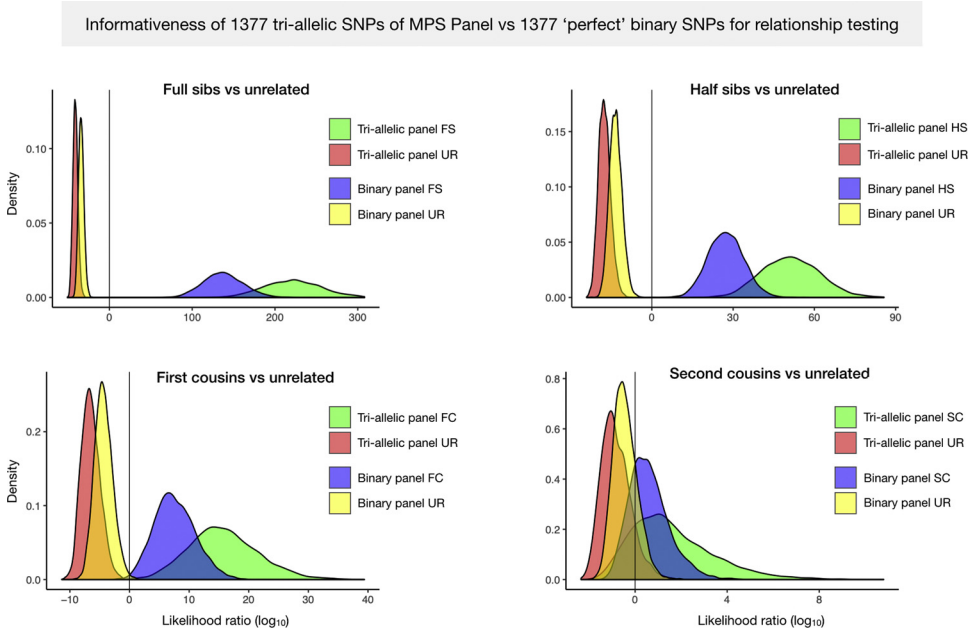

Fig. 4. The same Log LR plots as Fig. 3 shown with distributions from an equivalent number of perfect binary (0.5:0.5 bi-allelic) SNPs for each relationship testing scenario. While the binary SNP distributions are less separated, only second cousins have significant proportions of LRs that cross the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

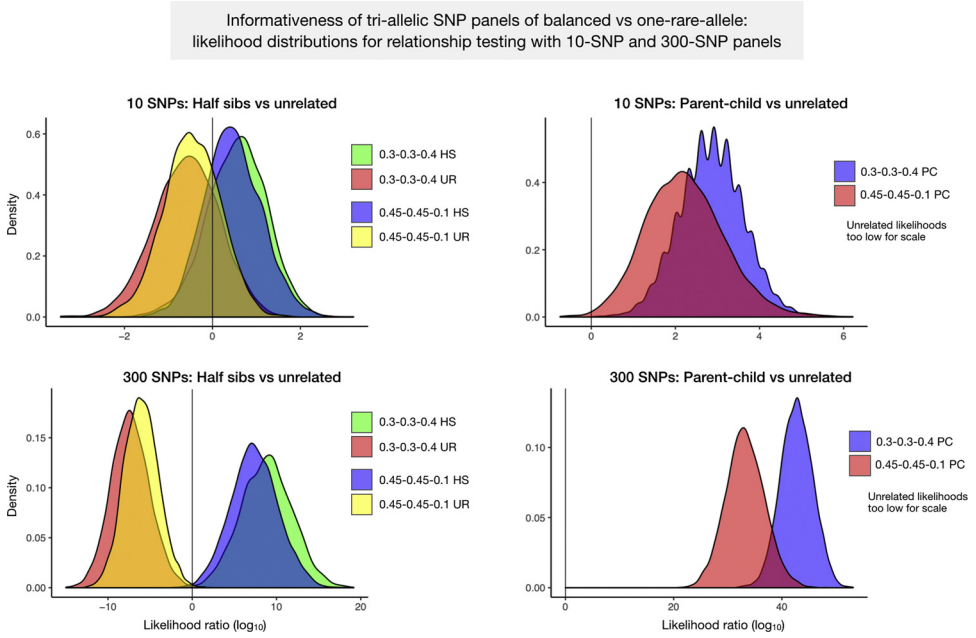

Fig. 5. Log LR plots from relationship test simulations exploring two contrasting allele frequency distributions in tri-allelic SNPs. A 10-SNP panel (upper) and 300-SNP panel (lower) were analyzed each with balanced allelic distributions of 0.34, 0.33, 0.33 frequencies, and one rarer allele of 0.5, 0.45,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

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

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

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

3.4. Ancestry informativeness of the MPS panel

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

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

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

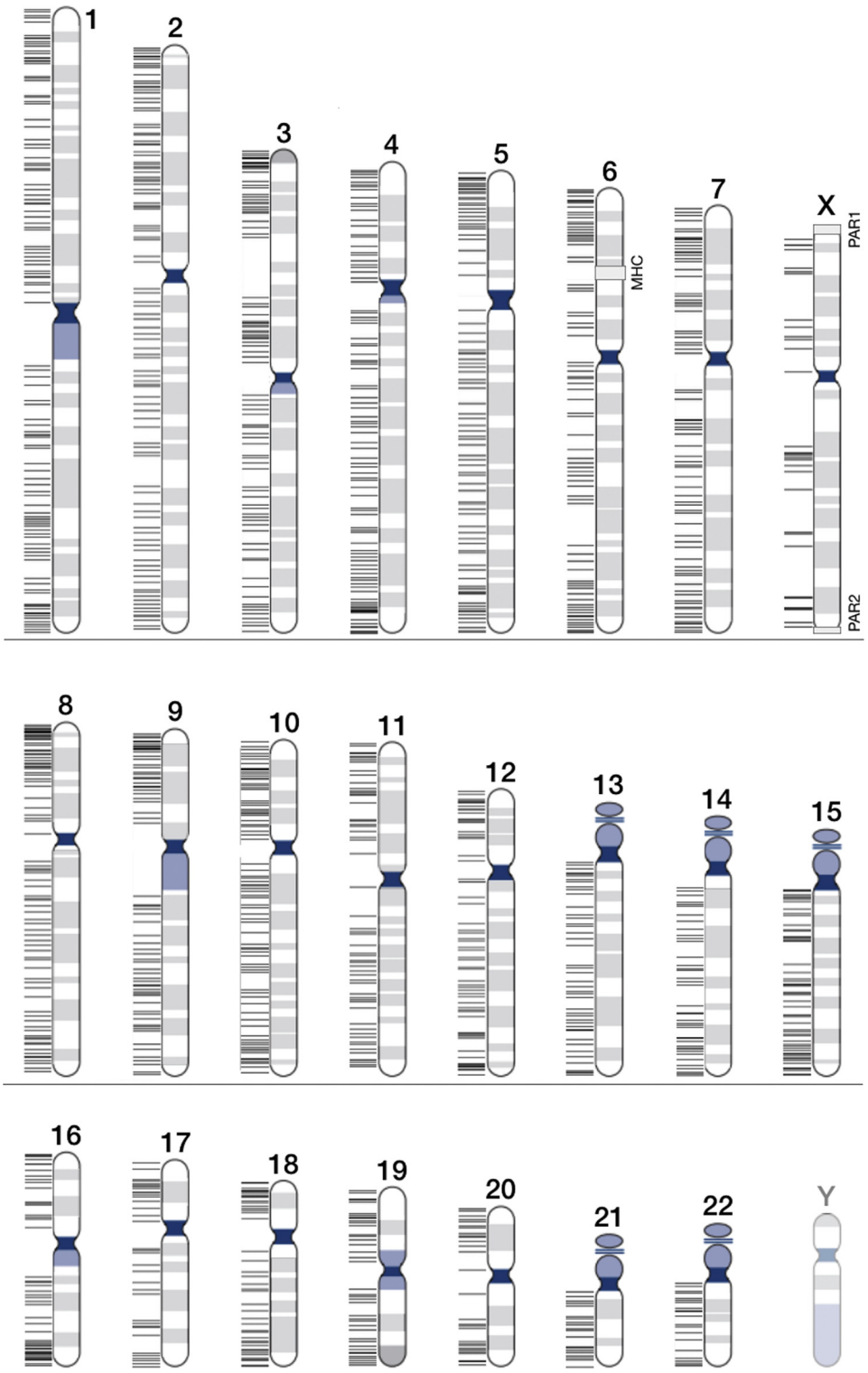

Fig. 6. Genomic positions of 1,411 tri-allelic SNPs selected for the MPS panel, scaled to the same chromosome ideograms as Fig. 2.

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

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

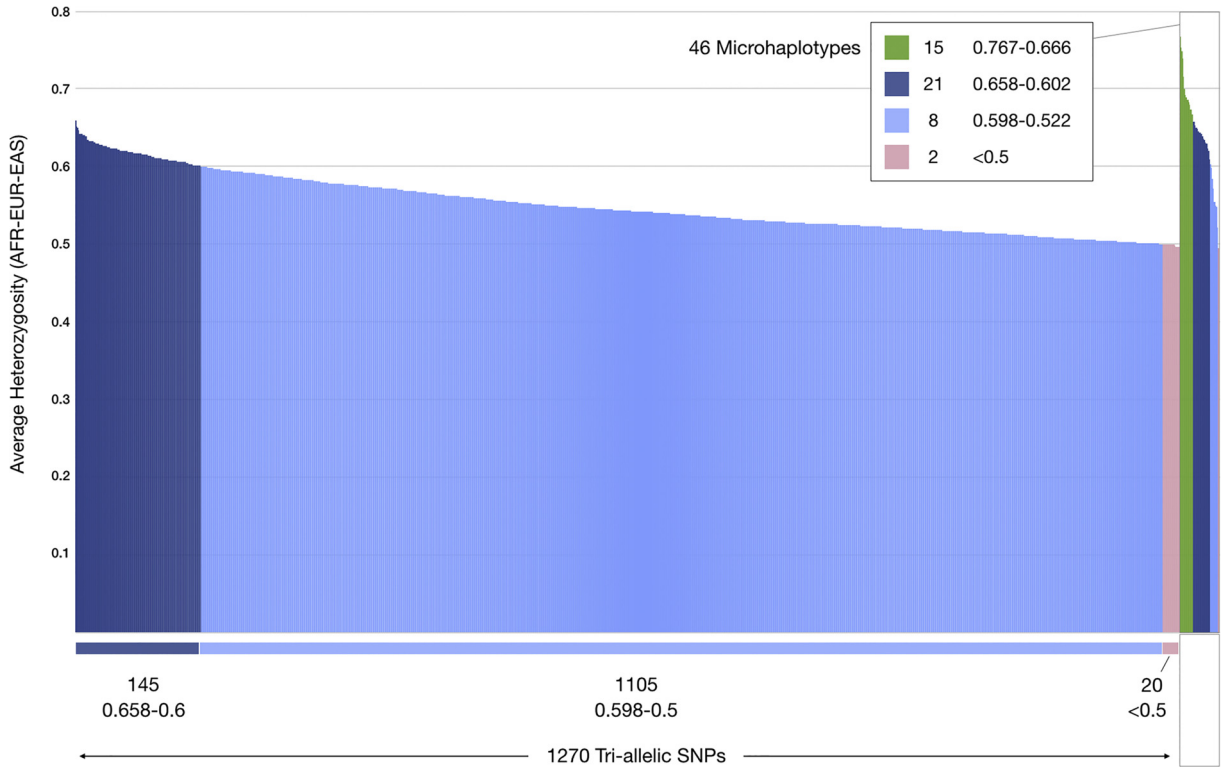

Fig. 7. Distribution of average heterozygosity values amongst 1,270 tri-allelic SNPs (bars left to center) and 46 microhaplotypes (bars on the far right), selected for the final MPS Panel. Of the tri-allelic SNPs, 145 (11.4 %) had the highest values exceeding 0.6 (mid blue bars); 1,105 (87 %) had values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

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

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

4. Concluding remarks

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

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

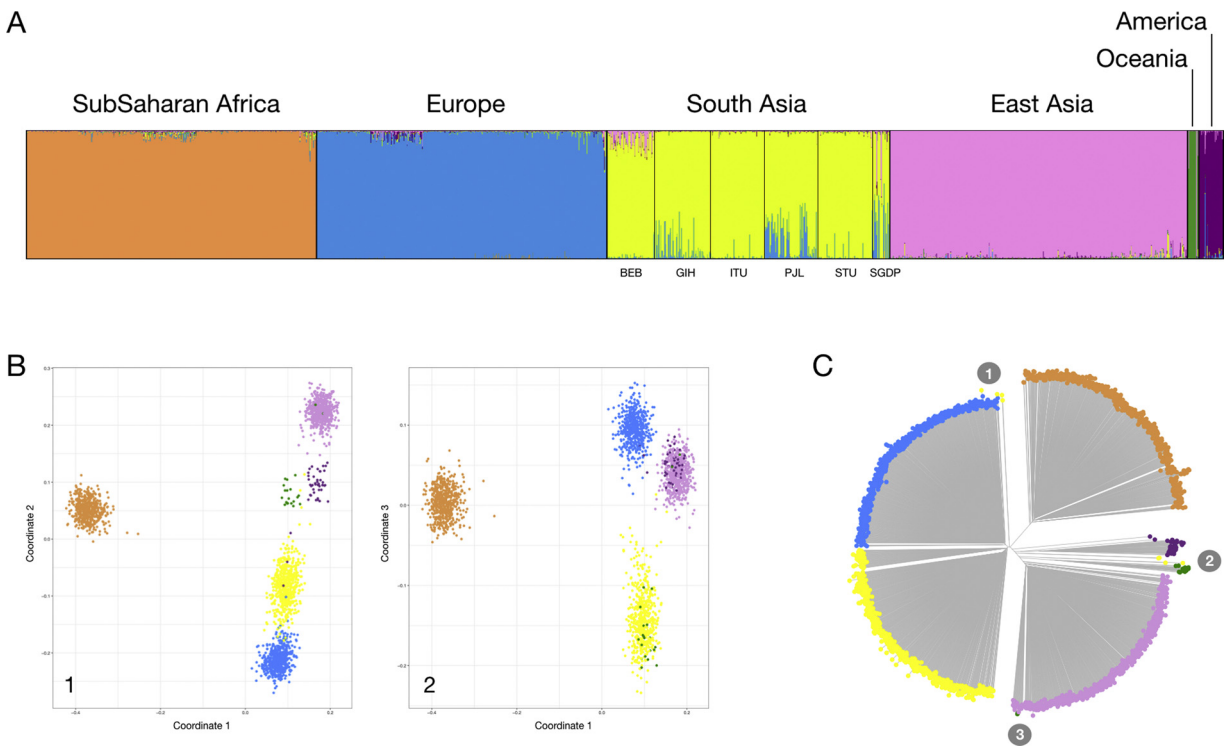

Fig. 8. An ancestry informativeness 'snapshot' of the 1,241 autosomal tri-allelic SNPs of the MPS Panel analyzed with STRUCTURE, MDS and Neighbor-joining tree tests. Samples from the six main population groups were taken from 1000 Genomes and Simons Foundation genome diversity project (SGDP). South Asian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

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

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

Acknowledgements

This study is supported by MAPA, Multiple Allele Polymorphism Analysis (BIO2016-78525-R), a research project funded by the Spanish Research State Agency (AEI), and co-financed with ERDF funds. MdlP is supported by a postdoctoral fellowship awarded by the Consellería de Cultura, Educación e Ordenación Universitaria and the Consellería de Economía, Emprego e Industria of the Xunta de Galicia (ED481B 2017/088).

Appendix A. Supplementary data

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

References

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

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