Forensic Science International: Genetics 42 (2019) 213–226

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

MAPlex - A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations

C. Phillips[a,*], D. McNevin[b,c], K.K. Kidd[d], R. Lagacé[e], S. Wootton[e], M. de la Puente[a,f], A. Freire-Aradas[a], A. Mosquera-Miguel[a], M. Eduardoff[f], T. Gross[g], L. Dagostino[h], D. Power[h], S. Olson[e], M. Hashiyada[i], C. Oz[j], W. Parson[f,k], P.M. Schneider[g], M.V. Lareu[a], R. Daniel[l,**]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
[b] Centre for Forensic Science, School of Mathematical and Physical Sciences (MaPS), Faculty of Science, University of Technology, Sydney, Australia
[c] National Centre for Forensic Studies, Faculty of Science & Technology, University of Canberra, ACT, Australia
[d] Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT, USA
[e] Human Identification Group, Thermo Fisher Scientific, 180 Oyster Point Blvd, South San Francisco, CA, USA
[f] Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria
[g] Institute of Legal Medicine, Faculty of Medicine and University Hospital, University of Cologne, Cologne, Germany
[h] Thermo Fisher Scientific, Victoria, Australia
[i] Kansai Medical University, Kansai, Japan
[j] Forensic DNA Laboratory, Division of Identification and Forensic Science (DIFS), Israel Police, National H.Q., Israel
[k] Forensic Science Program, The Pennsylvania State University, University Park, PA, USA
[l] Office of the Chief Forensic Scientist, Forensic Services Department, Victoria Police, Macleod, Australia

ARTICLE INFO

Keywords:
Forensic ancestry analysis
SNPs
Microhaplotypes
AIMs
Massively parallel sequencing
Population admixture

ABSTRACT

Current forensic ancestry-informative panels are limited in their ability to differentiate populations in the Asia-Pacific region. MAPlex (Multiplex for the Asia-Pacific), a massively parallel sequencing (MPS) assay, was developed to improve differentiation of East Asian, South Asian and Near Oceanian populations found in the extensive cross-continental Asian region that shows complex patterns of admixture at its margins. This study reports the development of MAPlex; the selection of SNPs in combination with microhaplotype markers; assay design considerations for reducing the lengths of microhaplotypes while preserving their ancestry-informativeness; adoption of new population-informative multiple-allele SNPs; compilation of South Asian-informative SNPs suitable for forensic AIMs panels; and the compilation of extensive reference and test population genotypes from online whole-genome-sequence data for MAPlex markers. STRUCTURE genetic clustering software was used to gauge the ability of MAPlex to differentiate a broad set of populations from South and East Asia, the West Pacific regions of Near Oceania, as well as the other globally distributed population groups. Preliminary assessment of MAPlex indicates enhanced South Asian differentiation with increased divergence between West Eurasian, South Asian and East Asian populations, compared to previous forensic SNP panels of comparable scale. In addition, MAPlex shows efficient differentiation of Middle Eastern individuals from Europeans. MAPlex is the first forensic AIM assay to combine binary and multiple-allele SNPs with microhaplotypes, adding the potential to detect and analyze mixed source forensic DNA.

1. Introduction

DNA Intelligence, also known as forensic DNA phenotyping, enables the prediction of biogeographical ancestry (BGA) and externally visible characteristics (EVCs) of the donor of forensic samples [1,2]. Such intelligence provides valuable investigative leads in the absence of database matches from STR profiling. The application of massively parallel sequencing (MPS) to forensic DNA analysis has enabled the simultaneous analysis of hundreds of polymorphic loci and multiple marker types, while providing sequence information not obtained using routine DNA profiling techniques. This has resulted in the ability to achieve greater resolution in ancestry prediction of global populations. The *Global AIMs* panel (herein gAIMs), was designed by the EUROFORGEN Consortium to provide a balanced differentiation of five

[*] Corresponding author at: Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain.
[**] Corresponding author at: Office of the Chief Forensic Scientist, Forensic Services Department, Victoria Police, Macleod, Australia.
E-mail addresses: c.phillips@mac.com (C. Phillips), runa.daniel@police.vic.gov.au (R. Daniel).

https://doi.org/10.1016/j.fsigen.2019.06.022
Received 23 February 2019; Received in revised form 4 June 2019; Accepted 26 June 2019
Available online 28 June 2019

continentally-defined population groups (Africa, Europe, East Asia, America, Oceania); i.e. each group is equally divergent for the 127 AIM-SNPs compiled in this set [3,4]. Two commercially available MPS panels were also developed as either a dedicated forensic ancestry test of 165 AIM-SNPs of the Thermo Fisher Scientific (TFS) Precision ID Ancestry Panel [5], or as part of a multi-purpose panel of STRs and SNPs for forensic identification, ancestry analysis and EVC prediction from Verogen; with 55 AIM-SNPs [6], also present in the TFS set, combined with 128 AIM-SNPs [7].

The existing forensic AIM-SNP panels have some limitations in their power to differentiate the populations in the Asia-Pacific region, an area with a complex distribution of genetic diversity and demographic histories across the Central, South and East Asian geographic continuum. This extensive region runs eastwards from the Caucasus and western Steppe across to Eastern China, the Korean peninsula and Japan, with the Indian sub-continent positioned midway between and south of these two extremes [8–12]. The far east of Asia also encompasses regions north to south that are almost as extensive in range; running from the sparsely populated northeastern regions of Siberia and Beringia, down to Island Southeast Asia which borders the region where Melanesian populations of Papua New Guinea mark the western edge of the Pacific region of Near Oceania. The limitations of existing forensic AIM panels to fully differentiate populations from the complex and widely distributed Asia-Pacific regions defined above has prompted the development of a new ancestry panel described in this study, named ‘MAPlex’ (Multiplex for the Asia-Pacific).

The marker composition of the MAPlex panel aimed to preserve the carefully equilibrated population divergence of the gAIMs SNP set [3,4], while improving the differentiation of Asia-Pacific populations and introducing, for the first time in a forensic MPS multiplex, a combination of binary SNPs, multiple-allele SNPs [13,14] and microhaplotypes [15,16]. Microhaplotypes consist of small sets of SNPs in short sequence spans (generally below 200 nucleotides in length) with contrasting haplotype frequencies amongst multiple populations [16]. AIM selection sought to enhance the MAPlex panel in four ways: i. selecting additional East Asian-informative SNPs from studies of population variation in this region [17–19]; ii. selecting a large number of South Asian-informative SNPs based on previous identification of loci with strong European-South Asian, or South Asian-East Asian differentiation [20]; iii. utilizing the advantage of sequence-based genotyping to allow the combination of microhaplotypes showing ancestry-informative variation with SNPs; iv. compiling multiple-allelic SNP data from the 1000 Genomes Phase III release (herein 1KG) [21] - leading to the inclusion of 30 tri- and 2 tetra-allelic AIM-SNPs. In combination with microhaplotypes, the multiple-allele SNPs enables the detection of mixed-source DNA. Furthermore, the accurate detection of sequence ratios in these loci with MPS offers the prospect of assigning ancestries to the contributors detected in simple mixtures.

We report the marker selection; forensic MPS assay development; and compilation of reference and preliminary test population data made to evaluate the ability of MAPlex to differentiate populations from the Asia-Pacific region and neighboring Middle East regions to the west of Central Asia. As there is a continuing paucity of population samples from Remote Oceania (Polynesian and Micronesian populations, including the populous archipelagos of New Zealand and Hawaii), the ability of MAPlex to differentiate this extensive region’s populations has yet to be fully evaluated.

2. Methods and materials

2.1. Selection of ancestry-informative markers

The 128 ancestry-informative markers previously compiled in the gAIMs panel [3] were all SNPs, of which six were tri-allelic loci with differentiated allele frequency distributions across multiple population groups. This small set of tri-allelic SNPs also provided limited mixture detection capacity. Both the overall ancestry-informativeness and mixture analysis potential of MAPlex were enhanced by increasing the number of tri-allelic SNPs, adding two tetra-allelic SNPs [14], and introducing a set of microhaplotypes that were carefully selected from 130 microhaplotypes identified and compiled by the Kiddlab and described elsewhere [22]. Therefore, AIM selection for MAPlex involved choosing the most informative binary SNPs for the targeted population differentiations, followed by addition of a significant proportion of multiple-allele loci at positions and microhaplotypes. To better match the expected amplicon sizes of the SNPs in MAPlex, over half of the original candidate microhaplotypes had their sequence spans reduced in length, at the cost of a degree of ancestry-informativeness.

2.1.1. Ancestry-informative microhaplotypes

The Kiddlab microhaplotypes show a wide range of variation characteristics, but not all loci identified are highly polymorphic or have contrasting haplotype frequencies amongst populations. Furthermore, many have haplotype variation mainly differentiating African from non-African populations, as would be expected from current knowledge of human population history [10,23,24]. Therefore, scrutiny and selection of suitable microhaplotypes for MAPlex centered on variant data from 1KG Phase III output [21] and assessed patterns of haplotype frequencies amongst four continental population groups; as defined by the 1KG project using population acronyms: ESN/GWD/LWK/MSL/YRI = 504 Africans (AFR); CEU/FIN/GBR/IBS/TSI = 503 Europeans (EUR); CDX/CHB/CHS/JPTL/KHV = 504 East Asians (EAS); BEB/GIH/ITU/PJL/STU = 489 South Asians (SAS), as summarized in Fig. 2 of [1]. To include assessments of Native American divergence for these loci, Peruvians from Lima, Peru (PEL) were used as the least admixed American-origin population in 1KG. In addition to 130 Kiddlab microhaplotypes, 7 of 27 loci discovered by Hiroaki et al. 2015 [25] were evaluated (labeled A–G, with -B already designated as mh01NK001 by Kiddlab). Compiling 1KG microhaplotype data provides three advantages: i. the haplotypes are accurately phased from sequence strand data rather than individual SNP genotypes; ii. there is comprehensive identification of short insertion-deletion polymorphisms (Indels) close to the target SNPs that can potentially interfere with sequence alignment and thus reliable recognition of SNP alleles; iii. the characterization of 2504 samples from varied populations of origin efficiently detects low frequency alleles that, when population-specific, may further improve ancestry analyses. Inclusion of every variant within the 5′ and 3′ bounding SNPs of a microhaplotype can lead to over-complicated patterns of haplotype variation, so low frequency variation, ranging from singleton SNPs (i.e. the variant nucleotide is observed in only one individual) up to SNPs with fewer than 20 allele observations (< 0.0002) were recorded but did not form part of the reconstructed haplotypes. For example, the selected microhaplotype mh15KK-095 has component SNPs rs2433354-rs534194624-rs2459391, but rs534194624 only has two observed G alleles in 1KG, so just rs2433354-rs2459391 haplotypes CA, CG, TA, TG were compiled.

Details of the original Kiddlab and Hiroaki microhaplotypes are given in Supplementary File S1. The sequence lengths of these candidate microhaplotypes set by the 5′ and 3′ bounding SNPs, ranged from 291 nucleotides (nt) down to just 12 nt, with only nine longer than 200 nt. However, previous experience with building the SNP-only gAIMs multiplex [3] has indicated the expected size range of amplified fragments for MPS genotyping of polymorphisms is 80–180 nt. Therefore, 47 of 136 candidate microhaplotypes would be too long for a forensic PCR reaction; assuming 20 nt primers could be positioned immediately next to the bounding SNPs of the remaining shorter loci. Where possible, microhaplotypes were reduced in size and their 1KG haplotype frequencies re-estimated. In many cases this resulted in loss of ancestry-informativeness and these microhaplotypes were not selected, despite showing much higher levels of population differentiation at their full size. Microhaplotype mh13KK-218 (rs1927847-rs9536430), described in Supplementary Fig. S2, provides an example of the limitation of sequence length. This locus is 145 nt long, and comprises 4 high frequency SNPs and 3 with singleton observations, giving 16 haplotypes with a wide range of frequencies. Excluding the single 3′ SNP rs9536430 reduces the size to a more viable 65 nt, but Supplementary Fig. S2 shows the number of haplotypes halves to 8 and much of its ancestry-informativeness is lost.

A small proportion of microhaplotypes were not considered further, as they had restricted variability due to complete linkage disequilibrium (LD) between two or more component SNPs (i.e. identical allele frequencies in adjacent SNPs within the haplotype). The following loci had SNPs with complete LD: mh12KK-045 (rs1976893 = rs1568791); mh12KK-093 (rs11111391 = rs7970865 = rs7970874); mh10KK-087 (rs10884095 = rs1452267); plus, Hiroaki-A (rs338564 = rs338565) and Hiroaki-C (rs2293140 = rs2293141). Microhaplotype mh07KK-082 only had one polymorphic SNP with the other three (rs150209521-rs138869704-rs115966953) having very low levels of variation in 1KG populations.

2.1.2. Ancestry-informative SNPs

A multiplex size of 160–170 component sequences was targeted for the MAPlex MPS sequencing reaction, based on the successful development of the Precision ID Ancestry panel (TFS). All forensic ancestry-informative SNPs evaluated in two independent surveys of forensic ancestry analysis panels [1,26] were considered as candidate AIMs for MAPlex based on their ability to differentiate the five main population groups. Although the bulk of SNPs were already identified as loci with the most divergent allele frequency patterns globally [7,17,27–29], additional forensic SNP sets have been published and were included in the candidate pool [18–20,30–33]. For MAPlex, AIM selection proposed to combine ˜85 binary SNPs informative for the five main population groups (AFR; EUR; EAS; Native American and Oceanian) and ˜25 informative for South Asia. Approximately 30 multiple-allele SNPs were proposed for selection, therefore, the remainder of the panel would comprise ˜20-30 sequences with microhaplotypes.

The web-based *Snipper* likelihood calculator was used to estimate each SNP's Population-Specific Divergence value (PSD; alternatively-termed Locus Specific Branch Length or LSBL [29]) per population group, as previously described [3]. SNPs were combined to achieve closely matched cumulative PSD values for the five main continental population groups.

South Asian-informative SNPs were selected from the original 30 SNP candidate pool of the *Eurasiaplex* panel, which differentiates populations of the Indian sub-continent from those of Europe and East Asia [19]. A similar number of South Asian-informative SNPs were obtained from other published studies [18–20,31]. The much lower South Asian PSD values were not used to adjust marker composition in MAPplex.

A large candidate set of several thousand SNPs with three or four detectable alleles were compiled from 1KG (tetra-allelic SNPs [17]; tri-allelic SNPs, publication in preparation). As with binary SNPs, individual PSD values were calculated for all candidate multiple allele SNPs and the cumulative PSD values in each population were used to combine the most informative SNPs while maintaining balanced differentiation.

2.2. Massively parallel sequencing

Ion AmpliSeq™ MPS primer designs for the MAPlex assay were developed as a custom panel by Thermo Fisher Scientific, and a forensic SNP genotyping pipeline was subsequently optimized for the Ion S5 sequencing system (publication in preparation).

DNA libraries were prepared with an Ion Chef™, following manufacturer's guidelines (MAN0007450, Revision A.0). A total of 300 μL of the custom primer pool were added into positions A and B (150 μL each) of the Precision ID DL8 Reagents cartridge and 15 μL of each DNA sample (0.067 ng/μL, 1 ng) was pipetted into wells A1 to H1 of each of the Precision ID DL8 IonCode™ Barcode Adapter Plates (cycling through IonCode™ barcodes 0101-0132). All libraries were prepared using 22 target amplification cycles and a 4-minute anneal and extension time. DNA libraries were quantified using the Ion Library TaqMan™ Quantitation Kit. Libraries were pooled to a final concentration of 30 pM in a total volume of 25 μL. Template preparation and chip loading were performed on the Ion Chef™, following manufacturer's guidelines. Ion 520™ and 530™ Chef reagents and supplies were used together with Ion 530™ chips. For all libraries, two chips each containing 32 samples were prepared simultaneously. Sequencing was performed on the Ion S5 Semiconductor Sequencer using Ion S5™ sequencing reagents, according to manufacturer's guidelines. Raw sequence data were processed on the S5 Torrent Server VM. SNP genotyping was performed with the HID_SNP_Genotyper_v5_2_2 plugin.

To optimize the sequencing protocols for the developed MAPlex assay prior to population sample genotyping, standard control DNA samples were genotyped in an initial series of runs and individual marker sequence coverage relative to total coverage measured. Test DNAs were chosen from Coriell human cell-line repositories to cover as much allelic variation as possible and consisted of: NA06994 (1KG CEU, parent in trio 1340); NA07000 (CEU, other parent in trio); NA07029 (child in trio); HG00403 (1KG CHB); NA18498 (1KG YRI); NA10540 (Oceanian ancestry); NA11200 (Native American ancestry).

Reconstruction of phased SNP genotypes from MPS sequence output is a complex and unwieldy process requiring detailed scrutiny of the aligned sequence strands in genome data browsers such as IGV [34]. Therefore, we adapted the TVC_Microhaplotyper_v8.1 plugin (herein '*Microhaplotyper*') originally developed for the TFS forensic mixture analysis MPS panel. Custom targets and hotspots 'bed' files were constructed and *Microhaplotyper* parameters were modified to set a relative analytical threshold of 0.1 for single-source sample genotyping, mimicking the default 0.1 minimum_allele_frequency threshold routinely applied with HID_SNP_ Genotyper.

2.3. Reference population data

In addition to 1KG Phase III genotypes, reference population data were obtained from other online whole-genome sequencing data and in-house MPS sequence analysis of the CEPH Human Genome Diversity Panel (HGDP-CEPH) populations [23], as well as complementary sample sets amongst authors' laboratories. The main sources of additional online population genotypes were: Simons Foundation Genome Diversity Project (SGDP) sequence analysis of 263 individual genomes in 127 populations [35]); and Estonian Biocentre Genome Diversity Panel (EGDP) sequence analysis of 402 individual genomes in 126 populations [36]). Herein, the term 'samples' denotes 'individual genomes' analyzed in both projects.

SGDP samples include 24 individuals in common with 1KG (24) and 122 HGDP-CEPH panel samples have been sequenced. 1KG overlaps were removed and HGDP-CEPH data retained for combination with in-house genotyping of complementary population samples in the panel. The SGDP population sampling regime was designed to obtain an extensive worldwide distribution of geographical positions, with only 2–3 samples per region. Oceania is a notable exception, with 15 HGDP-CEPH samples from Papua New Guinea, which were used in this study as the main source of complete variant sets representing Oceania.

EGDP has no sample overlap with any other human genome diversity panel. Worldwide sampling coverage is particularly useful for Eastern Europe, Central South Asia, India and Mainland/Island Southeast Asia. One disadvantage of the use of EGDP data is the processing of project genotypes (with PLINK) to remove multiple-allele SNPs. An advantage of EGDP and SGDP data (E-SGDP for brevity) is the inference of SNP genotype phase - so individual haplotypes for microhaplotype loci of interest have been estimated in both databases using 1KG as the reference framework. Missing microhaplotype component SNPs in E-SGDP data were interpreted to be monomorphic for the reference nucleotide. It should be noted that E-SGDP SNP phase is inferred using probabilistic software (both used SHAPEIT), so there is a degree of error across chromosome distances that are usually much longer than the microhaplotypes of this study (such “switch error rates” were estimated to be 2–4% per individual in EGDP). Samples from populations not represented in 1KG also have a significant fraction of population-specific heterozygous positions (e.g. SGDP Onge 4%; Papuans 5%; KhoeSan 11% of homozygous positions in 1KG); and these SNPs would therefore be excluded from S-EGDP phase calculations, although no novel component SNPs (i.e. not in 1KG) were detected in the microhaplotype data we compiled. Detailed descriptions of the SNP phasing strategies used and estimates of their error are given in the supplementary data of the relevant papers (SGDP: section S-9.2 in [35]; EGDP section S-1.5 in [36]). Some caution is also needed in making use of EGDP data in general, as the SNP genotype call error rates from the sequencing techniques employed are difficult to assess in the absence of overlapping samples with 1KG (whereas SGDP genotype calls can be compared in 24 sample-overlaps).

Given the above limitations of EGDP, and to a less extent, SGDP data; HGDP-CEPH samples were genotyped as extensively as possible and where geographic coverage of population diversity could be extended. Therefore, 21 Algerian Mozabite (North African reference data); 23 Israeli Bedouin, 24 Israeli Druze, 24 Israeli Palestinian (Middle East); 15 Orcadian (Orkney Islands, UK); 24 Siberian Yakut (North East Asian) were genotyped. HGDP-CEPH Native American reference data comprised 14 Karitiana; 8 Surui; 6 Colombian Piapoco; 20 Mayan; 14 Pima. HGDP-CEPH Oceanians added to SGDP data comprised 14 Papuan and Bougainville Melanesians. European HGDP-CEPH Sardinians (26 samples) and Adygei from Caucasus (15) were treated as test populations and not used for reference purposes in STRUCTURE.

Human variation data compiled by the Genome Aggregation Database or gnomAD were also accessed [37]. It should be noted that only allele frequencies are listed by gnomAD, and population descriptions generally lack geographic precision, e.g. Latino and ‘Other’; African includes African continent and African American samples, etc. Therefore, only gnomAD SNP allele frequency data from Ashkenazi Jews (419 samples) were used for comparison with European frequencies from other genome projects. Nevertheless, the gnomAD variant database is compiled from sequence analysis of 4368 genomes and is particularly informative for rare variation, or to detect allele-3 or allele-4 variants in binary and tri-allelic SNPs, respectively. Allele frequency data from the TOPmed project (Trans-Omics for Precision Medicine [38]) was also used to assess rare allele-4 variation in tri-allelic SNPs. This project’s data is based on 9000 samples, consists of allele frequencies only, and has no population information.

In-house population samples comprised 50 Iraqis (from the Kurdistan region of Northern Iraq) added to HGDP-CEPH Middle East reference population data; plus 30 Eritreans and 15 Somalis (from Eritrea and Somalia, resident in Germany) used as East African reference population data. Therefore, population analyses with MAPlex genotypes apportioned global variability in the reference data into six population groupings (acronyms AFR, EUR, SAS, EAS, AME, OCE); and added Middle East; North African; East African; and North East Asian groups (ME, NAF, EAF, NAS).

Apart from the HGDP-CEPH Oceanians and a small number of other HGDP-CEPH samples in SGDP, all SGDP and EGDP profiles were used as test data. Admixed 1KG population samples (ACB, ASW, PUR, CLM, MXL) were compiled in this study but their data are not included for brevity. In contrast, 1KG PEL were divided into 18 samples with no detectable admixture added to HGDP-CEPH Native American samples as reference data; and 67 that had varied, mainly AME-EUR admixture, included in test samples (based on high density SNP array analyses, publication in preparation). In addition, samples from three populations were analyzed as test populations to assess the efficiency of the compiled reference population MAPlex genotypes: 31 Turkish (resident in Germany); 50 Israelis (resident in Israel); and 38 Japanese (resident in Japan). Ethical approval was obtained for these (University of Canberra Committee for Ethics in Human Research Project number 11–119 and extension 15–64); plus the Iraqi, Eritrean and Somali samples (approval no. 17-416, by the Ethics Commission of the Medical Faculty of the University of Cologne).

2.4. Population data analyses

Three population analyses are routinely applied to forensic samples of unknown origin: genetic cluster detection with STRUCTURE software; Bayes analysis from the likelihood ratios of each population group assignment; and principal component analysis (PCA). Bayes and PCA tests can be made in a joint analysis (i.e. working from one input file) by using the ‘Classification of multiple profiles with a custom Excel file of populations’ option in *Snipper* [39]. However, this system only analyses SNPs (i.e. up to four alleles) and can only apply PCA tests to binary SNPs. Therefore, evaluation of the population differentiation capabilities of MAPlex focused on STRUCTURE (v.2.3.4) to analyze the reference population genotypes, as it accepts and co-analyses both SNP alleles combined with phased haplotype data. STRUCTURE runs followed standard parameter settings [40], comprising 10,000 burnin steps and 10,000 MCMC iterations, applying the admixture model with prior information on sample origins for reference populations (i.e. using *POPFLAG* = 1), with 10 simulations per K inferred genetic clusters. Individual (sample) and population Q-matrix data were analyzed in CLUMPAK, which merges the results of multiple simulations per K-value [41].

3. Results and discussion

3.1. Ion AmpliSeq™ assay conversion rate and optimization of the MPS multiplex

The Ion AmpliSeq™ assay successfully combined 144 SNPs and 20 microhaplotypes from the amplification of 164 target sequences. A small proportion of candidate markers were rejected at the design stage or under-performed in MPS sequencing.

For binary SNPs, all 108 candidate SNPs were incorporated successfully; an assay conversion rate of 100%. Issues associated with either design of amplification primers, sequencing, or alignment were largely avoided by reference to assessments of previous panel development [3,4]. For multiple-allele SNPs, five of the first 30 candidate loci of 28 tri-allelic SNPs, 2 tetra-allelic SNPs, were rejected (˜88% assay conversion rate) and replaced with 13 alternative tri-allelic SNPs. Two of these subsequently failed MPS quality thresholds and were removed, leading to 36 multiple-allele SNPs in MAPlex.

Primer designs were made for the 22 candidate microhaplotypes submitted. Locus mh14KK-101 (MHA-14; 5′-rs28529526-rs10134526-3′) repeatedly failed to produce good quality sequences and was removed. Another, mh16KK-255-sub 1 (MHA-16; 5′-rs9937467-rs17670111-3′) gave low sequence coverage values and was likely compromised in the MPS amplification steps by its very close proximity to another successfully incorporated microhaplotype (mh16KK-255-sub 2: MHA-17; 5′-rs12929083-rs9926495-3′); representing an MPS assay conversion rate for microhaplotypes of ˜91%.

A series of PCR primer pools was created and evaluated before the final combination of primer designs for 164 markers was established, based on consistent and balanced sequence coverage from Ion S5 MPS analyses. The seven Coriell human cell-line DNAs were run as common quality controls and their sequence coverage per marker was compared between prototype primer pools and the final optimized pool. Fig. 1 outlines the average sequence coverage obtained (average % proportions of total coverage per locus from 7 Coriell DNAs). Individual coverage is shown for the 17 loci that had sub-optimal MPS sequencing performance during evaluation of the prototype pools; and this is compared to the average coverage across the 148 unmodified primer pairs at the top of the plot. Primers were redesigned in 11 loci resulting

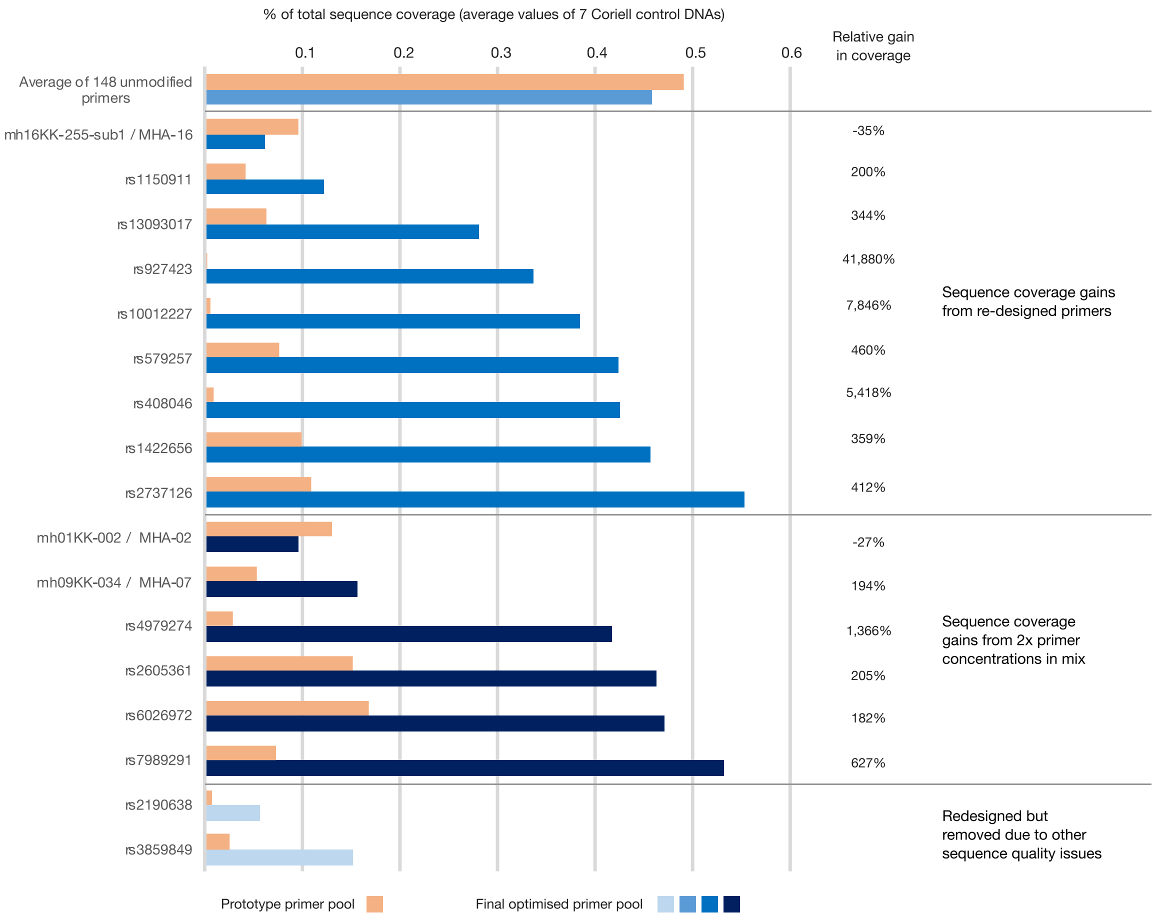

Fig. 1. Average sequence coverage (from 7 Coriell control DNAs) observed for MAPlex markers comparing the first prototype primer pool with the final pool. Loci with re-designed primers and doubled primer concentrations grouped separately and arranged by increasing final pool coverage levels. The average sequence coverage from unmodified primers in the prototype and final pools was 0.49% and 0.46%, respectively.

in improved sequence output in 10, with substantial gains in sequence coverage observed in rs927423, rs10012227 and rs408046. SNPs rs2190638 and rs3859849 were removed due to sequence quality issues not related to coverage (base of Fig. 1). Increased coverage was observed in 5 of 6 other loci where primer concentrations were doubled. Average coverage in the 148 unmodified primer pairs dropped slightly from 0.49% using the prototype pool to 0.46% for the final pool composition.

For nearly all under-performing loci, primer re-designs or increased relative concentrations brought their sequence coverage to levels matching the average % proportion of total sequence coverage (˜0.45%). Microhaplotype mh16KK-255-sub1 (MHA-16) failed to produce sufficient sequence coverage and, as described above, this was likely due to primer cross-reaction caused by the division of mh16KK-255 into two smaller microhaplotypes only separated by 46 N T. Locus mh16KK-255-sub2 (MHA-17) performed well in all cases and MHA-16 was not removed from the final primer pool, but the marker's sequence data were excluded from analysis. Supplementary Fig. S3 plots sequence coverage for the whole SNP set obtained from each primer pool (ranked by final primer pool % coverage) and indicates some marked contrasts in many other loci possibly due to several cycles of adjustment of primer ratios in the development of the final pool and altered relative primer binding specificities.

3.2. Microhaplotype adjustments and incorporation into MAPlex

Although a large proportion of the Kiddlab microhaplotypes that were evaluated for inclusion in MAPlex [22] could be shortened, the majority of these reduced-size sequences consequently lost ancestry-informativeness and were not considered further. Of the 22 candidate microhaplotypes chosen, nine were the same length as loci originally identified by Kiddlab. The 13 with reduced sizes brought their expected total amplified sequence lengths close to, or below, those of the SNPs in MAPlex. A maximum shortened microhaplotype length of 77 nt was achieved and this ensured primers could be designed to produce amplified sequences targeting these sites with equivalent lengths to SNPs. Fig. 2A shows the distribution of total amplified sequence lengths for the 144 SNPs, which reached a maximum 178 nt with an average length of 143.6 nt. Fig. 2B shows the amplified sequence lengths for microhaplotypes successfully incorporated in MAPlex. Length reductions led to a well-matched set of sequences with an average size of 147.6 nt and ensured the MAPlex assay had good balance in terms of the relative amplification efficiency of 164 component loci.

All component SNPs identified by 1KG in the full set of candidate microhaplotypes evaluated for MAPlex are outlined in Supplementary File S4. This file also gives summary haplotype frequencies for all microhaplotype loci considered for MAPlex (1KG population groups plus PEL: Peruvians from Lima, Peru).

Detailed genomic maps of the selected microhaplotypes are outlined in Fig. 3A–C, showing positions of all 1KG SNPs or Indels within the 5′ and 3′ bounding SNPs. These maps are divided into Fig. 3A showing identical size microhaplotypes; Fig. 3B loci reduced from the 5′ bounding SNP; and Fig. 3C loci reduced from the 3′ bounding SNP. Fig. 3A–C indicate (in red) ten novel polymorphic SNPs that were identified by 1KG within the established size limits, mostly amongst the reduced length loci. These variants complement the listed component

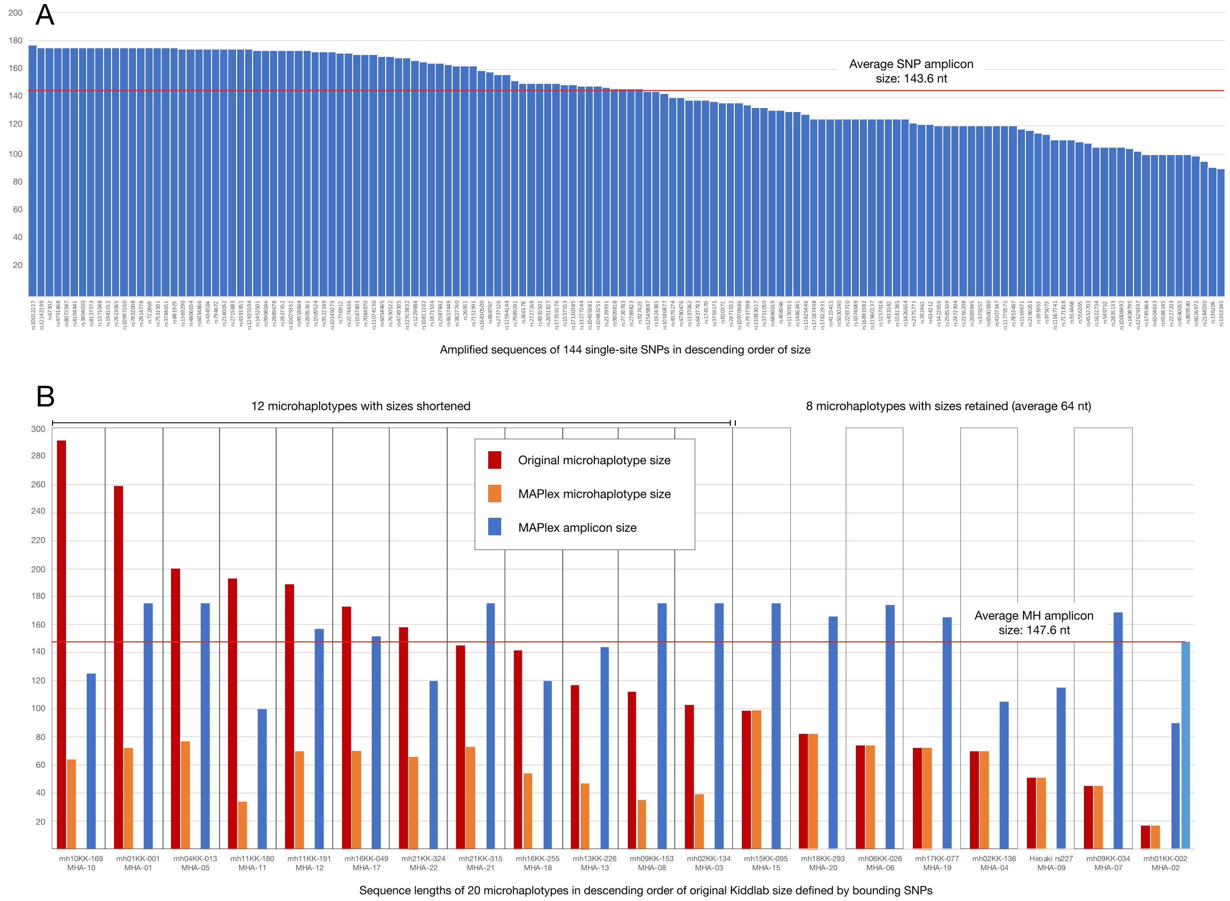

Fig. 2. A. Amplicon sizes of SNP components. Fig. 2**B.** Sizes of original microhaplotype spans compared with shortened versions (12 of 20 loci) and the final amplicon sizes. nt: nucleotides.

SNPs compiled by Kiddlab, first identified by targeted Taqman SNP genotyping [15,16,22]. Two microhaplotypes contained Indels that could potentially interfere with alignment of their sequences in MPS analysis. The Indel rs557418490 in mh09KK-153 is a rare variant in 1KG (3 observations), while the Indel-SNP-Indel combination of rs138895664-rs6517970-rs202132081 has variants at high frequency and in successive nucleotides in mh21KK-315. These three sites produce four distinct sequence motifs positioned at the 5′ end of the shortened

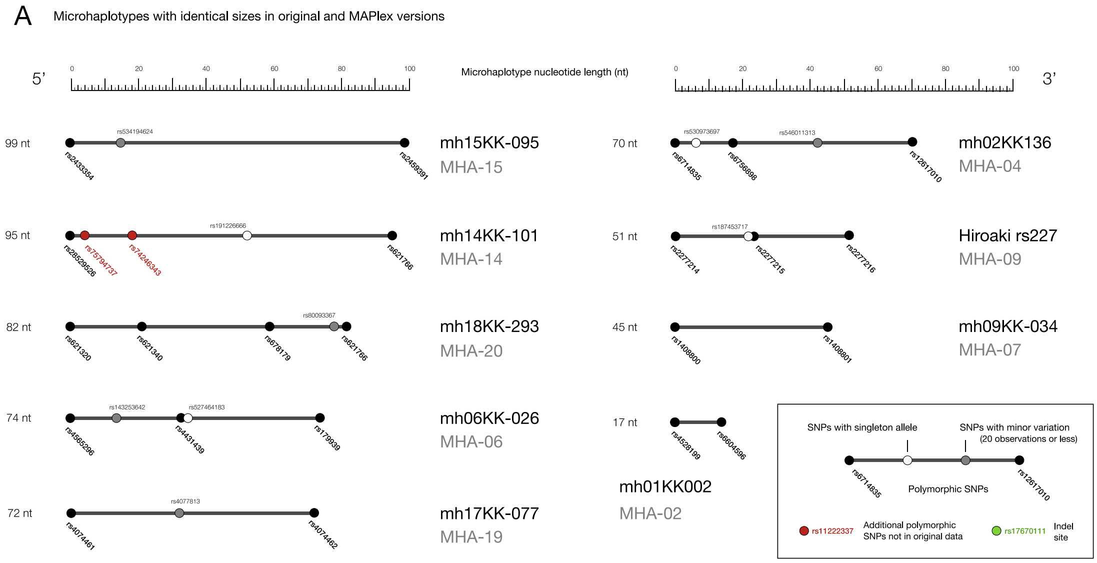

Fig. 3. Genomic maps of 22 candidate microhaplotypes showing original versions at full length and loci with reduced sizes. **3A**: Identical sizes in original and MAPlex versions. **3B**: Four microhaplotypes reduced in size by exclusion of 3′ component SNPs (plus MHA-05, a mid-segment design). **3C**: Six microhaplotypes reduced in size by exclusion of 5′ component SNPs (plus MHA-16/MHA-17, two designs at each end of the original mh16KK-049 locus). Component SNPs are marked with black, gray or white spots based on 1KG allele frequencies > 0.004, ≤0.004, singletons, respectively. Red dots signify polymorphic SNPs discovered in 1KG data (> 0.004) and green dots Indel sites. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

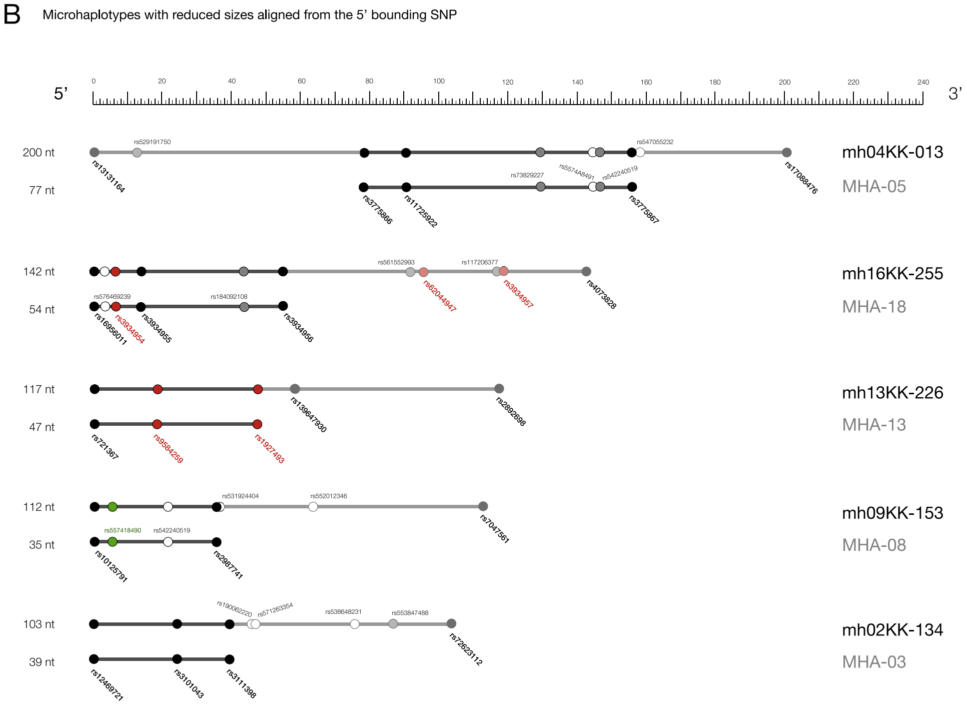

Fig. 3. (*continued*)

MHA-21 microhaplotype 5′-rs6517970-rs6517971-3′ (inset diagram in Fig. 3C). In all sequence analyses of new population samples made so far, the rs6517970 position has genotyped as a tri-allelic SNP with the frequency of G-allele calls matching those of 1KG. Whether these genotypes are created by an insertion displacing the sequence alignments or they represent true tri-allelic variation at one position will require the sequencing of larger numbers of samples and detailed scrutiny in IGV.

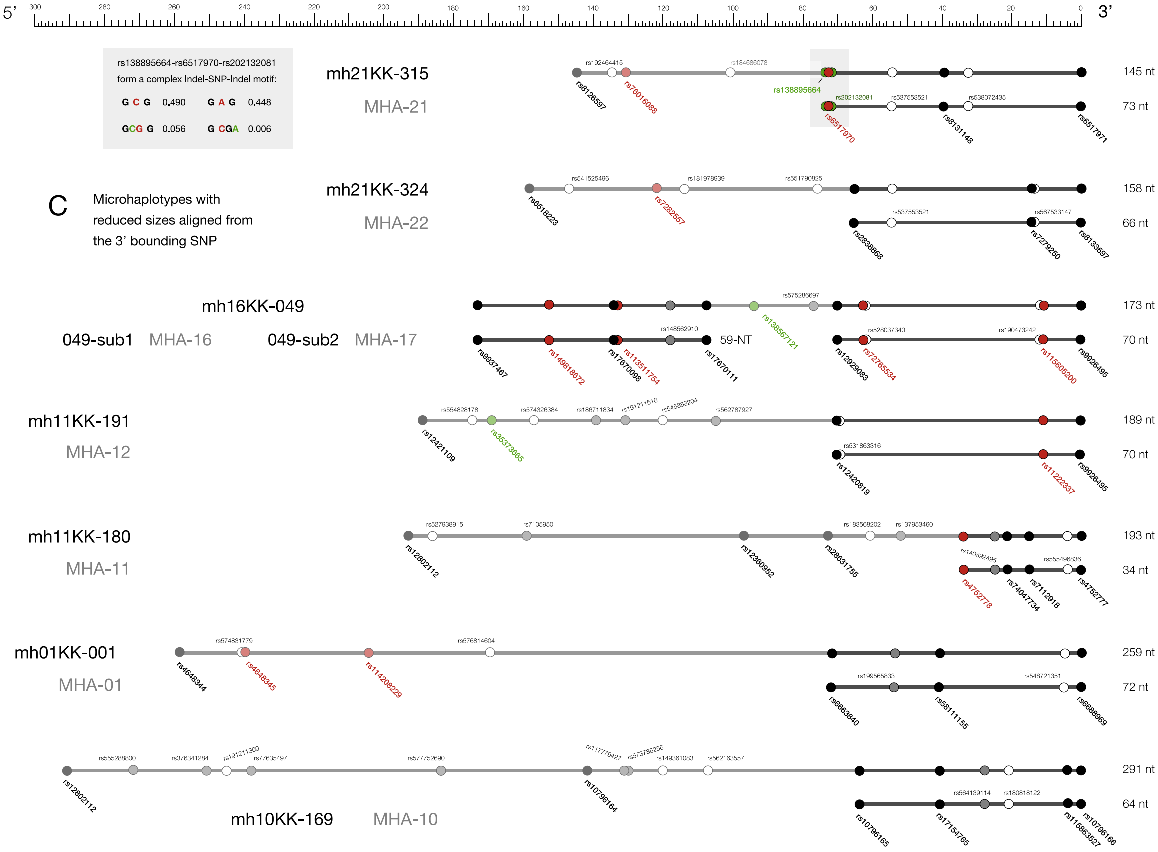

Fig. 3. (*continued*)

3.3. SNP sets incorporated into MAPlex

Of the 144 SNPs successfully incorporated into MAPlex, 89 (83 binary, 6 tri-allelic SNPs) had been previously genotyped without major issues in the gAIMs MPS assay [4], reflecting the emphasis placed on known MPS performance of core forensic AIMs. In contrast, sequence alignment problems had been observed previously for gAIMs rs2080161, rs595961 and rs6875659 due to flanking homopolymeric tracts [4] and as a result these SNPs were not considered for MAPlex.

Only one additional binary SNP rs6500380 was added to improve East Asian differentiation [32]. Including this SNP compensated for reduction of the original gAIMs 5-group differentiation SNP set to accommodate microhaplotypes, South Asian-informative SNPs, and a larger proportion of tri-allelic SNPs. As previously discussed by Soundararajan et al. [26], autosomal SNP-based forensic ancestry sets share only a handful of their most differentiating SNPs. Just eleven SNPs overlap between MAPlex and the TFS Precision ID Ancestry Panel, including the most common overlapping AIMs of: rs12913832; rs1426654; rs16891982; rs2814778; and rs3827760. Therefore, up to 318 AIMs can be analyzed with two parallel forensic MPS runs incorporating these two panels, although the value of this approach for improved population differentiation was not formally assessed in this study.

3.3.1. Balance of population informativeness and population-specific Divergence values

The division of population informativeness amongst the binary SNPs in MAPlex was: 20 Oceanian-informative; 16 Native American-informative; 6 African-informative; 22 European-informative; 20 East Asian-informative; and 24 South Asian-informative markers. The five-group Population-Specific Divergence (PSD) values were calculated for each SNP by comparing each population group's variant data with the other four populations combined and are listed in Supplementary File S5. The cumulative PSD values for each population group indicate their degrees of divergence when the accumulated SNP genotypes are used to compare each region as a possible ancestry for unknown forensic samples. Fig. 4 plots the cumulative PSD values, rising at different rates as each of the 144 SNPs is added. Plot lines progress to a point of convergence with a well-equilibrated set of cumulative PSD values when calculated for the 84 SNPs: ranging from 8.92 for Americans to 10.1 for Africans. The balance in cumulative PSD continues as South Asian-informative SNPs and multiple-allele SNPs are added, with a somewhat reduced level of American divergence compared to the other populations. It was difficult to compile a sufficient number of AIMs to differentiate American populations, as previously observed with the original gAIMs panel [3,4]. In addition, a wide selection of AIMs were available to choose to differentiate Africans but were excluded to preserve PSD balance in the other populations.

South Asians are differentiated reasonably well from European and East Asian ancestries, but it was not possible to reach comparable levels of differentiation of South Asians from Middle East populations while keeping the SNP panel a compact size. The cumulative PSD values of South Asia are lower amongst SNPs than the other population groups values and only rise significantly for EUR-SAS pairwise Divergence values, underlying selection of this subset of SNPs for MAPlex primarily to differentiate South Asian from closely related European populations. The most differentiated EUR-SAS SNP rs10008492, shown in Fig. 4 produces the biggest rise in Divergence.

3.3.2. Multiple-allele SNPs

Table 1 gives the locus details of the 34 tri-allelic and 2 tetra-allelic SNPs incorporated in MAPlex; also listing 5 loci rejected from primer designs and two that failed MPS quality thresholds during primer development. All six tri-allelic SNPs previously added to gAIMs were retained, but several tetra-allelic SNPs with strong population differentiation characteristics, identified since the gAIMs panel was made, were not considered. It can be expected that a significant proportion of multiple-allele SNPs are actually divergent nucleotides in replicated sequences resulting from segmental duplications and therefore the variants and their surrounding sequences are not unique parts of the genome. This has proved to be the case for many tetra-allelic SNPs, with 17 of the 41 loci identified to be potentially useful for forensic analysis [14] showing multiple chromosome hits (personal communication, Peter de Knijff, Leiden). Recognized multiple-hit SNPs include: rs12481195 (SAS/EAS-informative alleles); rs539004 (AFR); and arguably the best tetra-allelic AIM-SNP, rs4839116 (AME-specific T allele at high frequency).

Similar detailed checks of tri-allelic SNP candidates ensured their genomic locations were in unique sequences. Furthermore, all SNPs listed in Table 1 had three or four alleles identified by gnomAD, and most had the same pattern of variation in TOPmed data - databases that have enhanced power to detect low frequency allele-3 or allele-4 variants from larger sample sizes. Table 1 provides the population-wide allele frequency estimates from both databases for multiple-allele SNPs described in 1KG to have four alleles. Three tri-allelic SNPs had rare allele-4 variants in gnomAD, and a further four in TOPmed only, extending variation in these loci, albeit at relatively low genotype frequencies.

3.3.3. Adventitious microhaplotypes in SNP sequences

As sequence analysis provides scope for genotyping additional variation close to the targeted SNP site, closely positioned flanking sequence variants with divergent allelic variation from the target SNP create adventitious microhaplotypes that potentially add extra information. Flanking sequence variants within ± 20 nt of MAPlex SNPs and with allele frequencies greater than 0.001 were collated from 1KG, gnomAD and TOPmed sources and are listed in Supplementary Table S6. 30 SNPs have one flanking variant, 5 have two, and rs10186877 has three flanking variants. Although flanking variants generally have very low frequencies, a noticeable polymorphism gain is found in rs6504633, where the four alleles of the target SNP are expanded to a 3-SNP microhaplotype by including rs144594022 (overall 1KG variant frequency: 0.0052) and rs6504634 (0.8924). The increase in heterozygosity from SNP to microhaplotype is 64% to 66%, but more importantly, informative variation is found in rs6504634 in Africans and South Asians. Adjustment of sequence analysis details in the TFS *Microhaplotyper* plug-in can readily accommodate the best extra variants listed, providing additional differentiation power from MAPlex sequence data.

3.4. Compilation of reference population data and comparisons with test populations

A total of 3638 SNP genotype and haplotype profiles are listed in Supplementary Table S7. Haplotypes are listed in two separate adjoining columns for each microhaplotype locus. Because of the absence of multiple-allele SNP genotypes in EGDP and no phased haplotypes in SGDP for MHA-08, MHA-11 and MHA-20, the overall data completeness is 97.4%. This grid was used as the input for STRUCTURE analyses, with data from the admixed 1KG populations of ACB, ASW, PUR, CLM, and MXL at the bottom of Supplementary Table S7 not analyzed further. Note that populations were subsequently re-ordered according to their geographic distribution in the STRUCTURE cluster membership proportions data (Supplementary File S10) in order to highlight relationships between reference and test populations.

3.4.1. Worldwide patterns of variation in MAPlex AIMs

Supplementary Fig S8 gives individual summary maps of the global distribution of variation in the 144 SNPs and 20 microhaplotypes of MAPlex. SNPs are in the same order as those in Supplementary File S5. SNP summary allele frequencies are shown as pie charts representing combined data from each of the four major 1KG population groups and

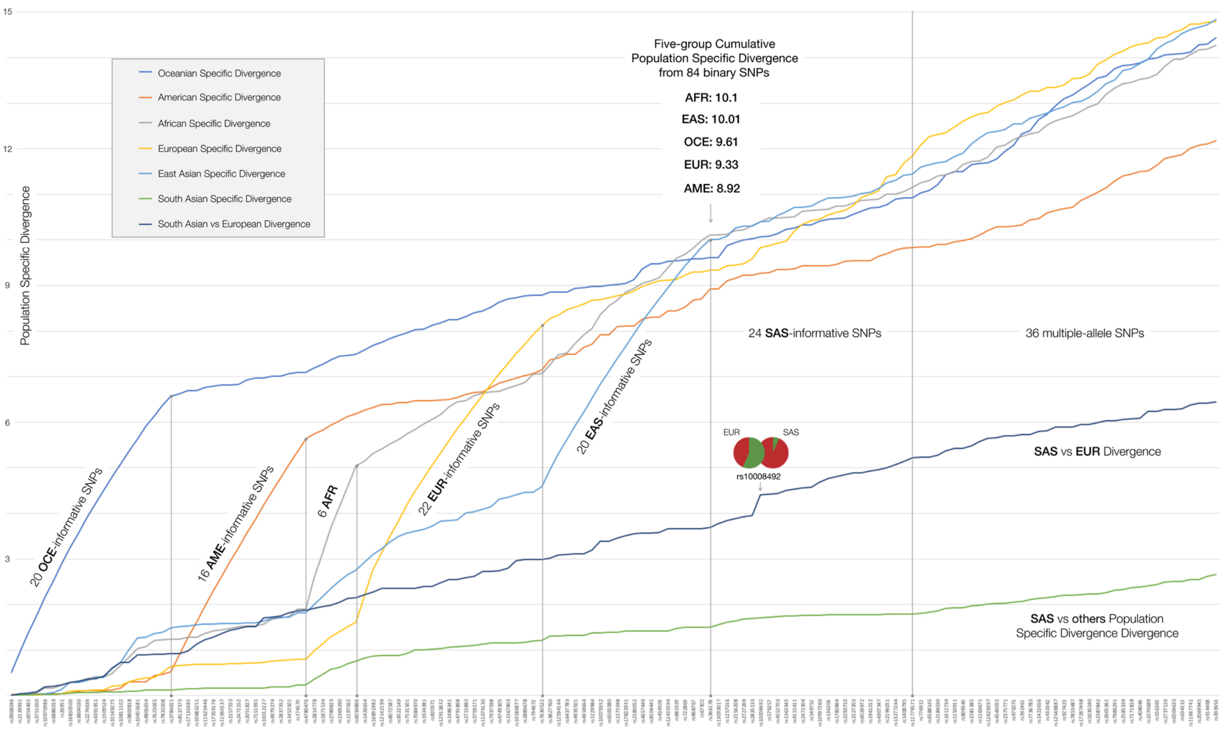

Fig. 4. Plots of cumulative Population-specific Divergence values for each of the five population groups (i.e. values for each SNP comparing OCE vs. non – OCE genotypes, etc.). Separate plotlines are shown for SAS PSD (SAS genotypes vs. other groups) and for pairwise Divergence between SAS and EUR.

individually from the PEL population, plus HGDP-CEPH North Africans (21); Oceanians (28); and Middle East (71). The single well-defined population from gnomAD of Ashkenazi Jews (300) is also included. The rs-number and closest or actual gene where each SNP is sited is given in the top left, and population informativeness is briefly summarized top right. When TOPmed or gnomAD report extra alleles (e.g. an allele-3 in binary SNPs, or allele-4 in tri-allelic SNPs), these are listed above the key of common alleles. Flanking sequence (± 25 nt) is given at the base of each map. When this sequence has been identified as potentially low complexity sequence it is given as lower-case nucleotides. The 11 SNPs overlapping with the Precision ID Ancestry Panel are highlighted.

Certain features of the distribution of variability observed in the SNP maps are worth highlighting. First, both PEL and OCE populations show consistently high levels of homozygosity, as can be expected from their distance from the hypothetical center of human population expansion, commonly given as East Africa (see Fig. 3B in [24]). In many such SNPs finding the allele in a forensic sample that is absent from these populations contributes disproportionately to the ancestry inference likelihoods, justifying the marker's inclusion even when there is a lack of divergence elsewhere (e.g. rs10149275-G). This characteristic is sometimes accentuated in tri-allelic SNPs (e.g. rs927423).

Second, 8 SNPs have observable divergence, i.e. an allele frequency differential > 0.1, between Ashkenazi Jews and European populations (these patterns marked for relevant SNPs in Supplementary Fig. S8). SNPs rs10008492, and rs1229984 have higher than average levels of divergence, with the rs1229984-T allele that is diagnostic of East Asia and almost absent elsewhere except at 0.2-0.25 frequencies in Middle East and Jewish samples. Studies indicate a high frequency in several Middle East populations and evidence for positive selection at rs10008492, and rs1229984 in these populations [42,43]. Third, 9–10 SNPs have above-average divergence between EUR and SAS populations, underlining enrichment for this differentiation during MAPlex SNP selection. Binary SNPs with high EUR-SAS differentiation include rs12142199, rs136206, and rs4115411; while three tri-allelic SNPs have highly contrasted allele-3 frequencies with EUR: rs408046-T; rs12629397-C; rs2585339-C (this allele is almost absent from EUR).

Lastly, almost all the tri-allelic SNPs in MAPlex are highly informative for at least one population group, with rs2585339 possibly the most informative of all; having a G-allele absent from EAS and OCE, and a C allele at a frequency of 0.27 in SAS which is almost absent from EUR. In comparison, the long-established tri-allelic SNPs of rs4540055 and rs5030240 are now amongst the least informative of these loci, despite showing a range of allele frequency distributions amongst global populations, which we continue to compile [44]. The tri-allelic SNP rs914468 shows the lowest levels of divergence across most populations, but has the singular characteristic of a T-allele in SAS and NAF at informative frequencies (0.25, 0.1) but not present in EUR.

Maps of the global distribution of microhaplotype variation are arranged in Supplementary Figs S8 as bar charts of 1KG data only, but these show individual populations in each group to illustrate the extent of within-group, between-population variation. Where loci have been shortened, the original length is given at the top right, and very rare haplotypes (1 or 2 observations) not visible in the bar charts are listed separately. Sequences placed at the bottom show all SNPs compiled for the haplotypes in red and the nucleotides between the 5′ and 3′ bounding SNPs.

Generally, microhaplotypes have little within-group, between-population variation. Microhaplotypes with the highest within-group variation appear to be mainly loci showing the highest haplotype diversity, notably: MHA-4; MHA-10; MHA-18; MHA-20; and MHA-21. Although MHA-02, MHA-07 and MHA-08 only have 2 common-variation SNPs, they have four haplotypes at moderate frequencies, making them more informative than tetra-allelic SNPs and confirming that most well-chosen microhaplotypes will be more informative than SNPs 'per sequence' in large-scale forensic MPS multiplexes. The least polymorphic microhaplotype MHA-15 would not be a choice for forensic identification purposes, but the almost complete lack of variation in EUR makes it informative for the other populations. In contrast, the most polymorphic microhaplotype MHA-21 only has 3 SNPs but benefits from tri-allelic variation in rs6517970.

Patterns of microhaplotype variation in EAS populations suggests the selection of several loci with the most contrasting haplotype

Table 1
Characteristics of multiple allele single-site SNPs in MAPplex, failing MPS quality checks, or rejected. GnomAD and TOPmed alleles indicated when 1000 Genomes lists four variants.

dbSNP description	SNP	Chr.	GRCh37 Position	Gene	Reference Allele	Other Alleles	gnomAD Frequencies (all populations)				TOPMed Frequencies			
Tri-allelic	rs776912	1	10847784	CASZ1	T	A,C								
Tri-allelic	rs6588145	1	65859784	DNAJC6	T	A,G								
Tri-allelic	rs2184030	1	206667441	–	G	A,C								
Tetra-allelic	rs1612734	1	208441488	–	G	A,C,T	G: 0.693	A: 0.281	C: 0.022	T: 0.005	G: 0.684	A: 0.279	C: 0.030	T: 0.007
Tri-allelic	rs1150911	1	228494382	OBSCN	T	A,C,G	T: 0.306	A: 0.000005	C: 0.260	G: 0.435		No data		
Tri-allelic	rs809540	2	7879001	–	G	C,T								
Tri-allelic	rs1581385	2	213036346	ERBB4	C	A,T								
Tri-allelic	rs1169671	3	10222955	IRAK2	C	G,T								
Tri-allelic	rs12629397	3	65814779	MAGI1	G	C,T								
Tri-allelic	rs4540055	4	38803255	TLR1	A	C,T								
Tri-allelic	rs2375771	4	187371930	–	C	A,G,T	(C,G,T only)	C: 0.389	G: 0.490	T: 0.120	C: 0.265	A: 0.000016	G: 0.546	T: 0.189
Tri-allelic	rs975073	5	23239056	–	T	A,C								
Tri-allelic	rs392461	5	81720271	–	C	A,G,T	C: 0.415	A: 0.124	G: 0.460	T: 0.001	C: 0.281	A: 0.213	G: 0.504	T: 0.002
Tri-allelic	rs7736783	5	82402869	XRCC4	A	G,T								
Tri-allelic	rs1422656	5	172893826	–	T	A,C								
Tri-allelic	rs433342	8	17747876	FGL1	A	C,G								
Tri-allelic	rs12549887	8	69139831	PREX2	C	A,G,T	(A,C,T only)	C: 0.403	A: 0.248	T: 0.350	C: 0.193	A: 0.366	G: 0.000024	T: 0.440
Tri-allelic	rs927423	9	38331864	–	A	C,T								
Tri-allelic	rs7853487	9	107368603	OR13C2	T	A,C								
Tri-allelic	rs17287498	10	54530788	MBL2	C	A,T								
Tri-allelic	rs5030240	11	32424389	WT1	C	A,G								
Tri-allelic	rs2387842	12	38736442	–	T	C,G								
Tri-allelic	rs2605361	12	74903531	–	T	A,G								
Tri-allelic	rs7989291	13	57572989	–	A	C,G								
Tri-allelic	rs2585339	14	49134978	–	T	C,G								
Tri-allelic	rs7171818	15	58855169	LIPC	G	A,T								
Tri-allelic	rs408046	15	80031510	–	A	G,T								
Tri-allelic	rs1074689	16	52216074	–	A	C,T								
Tri-allelic	rs556365	16	65927802	–	T	A,G								
Tri-allelic	rs2737126	17	3618815	ITGAE	C	G,T,A	C: 0.479	A: 0.00003	G: 0.294	T: 0.227	C: 0.343	A: 0.000024	G: 0.376	T: 0.281
Tetra-allelic	rs6504633	17	48112927	–	G	A,C,T	G: 0.160	A: 0.016	C: 0.236	T: 0.588		No data		
Tri-allelic	rs634212	18	6033023	–	G	A,C,T	(C,G,T only)	G: 0.444	C: 0.189	T: 0.367	G: 0.160	A: 0.001	C: 0.325	T: 0.513
Tri-allelic	rs11667741	19	3085244	–	T	A,C								
Tri-allelic	rs2069945	20	33761837	PROCR	C	A,G								
Tri-allelic	rs914468	20	62100463	KCNQ2	C	G,T								
Tri-allelic	rs393953	21	43389036	–	G	A,T								
Failed MPS QC	rs2190638	7	148378745	–	T	A,C								
Failed MPS QC	rs3859849	22	39817017	TAB1	C	A,G	(A,C,G only)	C: 0.320	A: 0.196	G: 0.484	C: 0.085	A: 0.274	G: 0.603	T: 0.038
Rejected from design	rs1695865	1	1390514	–	T	A,G								
Rejected from design	rs7587417	2	202156054	ALS2CR12	T	C,G								
Rejected from design	rs353216	5	148842165	–	T	C,G								
Rejected from design	rs6496739	15	91494430	UNC45A	C	A,T								
Rejected from design	rs145521066	18	7302006	–	A	C,T								

frequencies between East Asia and other groups has enhanced this differentiation. This extends to loci with little or no haplotype variation in EAS, such as MHA-06 (almost exclusively GCCG), and MHA-19 (mainly GGG); where the detection of alternative haplotypes common in other population groups markedly increases their ancestry inference likelihoods. Several loci have the converse phenomenon of EAS-specific haplotypes for the same reason, notably: MHA-01; MHA09; MHA-11; MHA-18 at low frequencies; the more differentiated MHA-08, where the indicative TG haplotype is at a relatively high frequency; and MHA-13, where TCC is the most common EAS haplotype, but found at lower frequencies in PEL, SAS and AFR. MAPlex microhaplotypes show a general lack of divergence between EUR and SAS, although the MHA-01 AGCCC haplotype is largely confined to SAS but at low frequency, and the MHA-09 GGT haplotype is absent from EUR but at a very low frequency in SAS (and more frequent in AFR).

Our experience of combining SNPs with microhaplotypes in an ancestry panel suggests the identification of many more novel microhaplotypes will be worthwhile and will lead to improved population differentiation, but not necessarily with sufficient divergence within-groups, between-populations, to improve this level of geographic resolution.

3.4.2. Genetic cluster analysis of population samples using STRUCTURE

From a series of STRUCTURE analyses, K:7 genetic clusters were consistently identified from CLUMPAK output and accompanying ΔK estimates [40,45] as the optimum clustering regime for the population data listed in Supplementary File S7. At K:8, two genetic clusters were commonly recognized in SAS reference populations, or less frequently in Africans (example CLUMPAK population Q-matrices for merged K:8 data in Supplementary Fig. S9).

In all STRUCTURE analyses made, reference population samples were assigned numbers 1–10 (population group acronyms listed in Supplementary File S7, col. B) corresponding to *POPFLAG* = 1 in the STRUCTURE input data. All test population samples were assigned 0, corresponding to *POPFLAG* = 0, so allele frequency estimates were not made from this data. Fig. 5 shows a cluster plot of the reference data set of 2333 samples with well-established populations of origin, representing K:7 clusters identified from their allele and haplotype frequencies (clusters colored as in Rosenberg et al. 2002 [23]).

K:7 represented the highest value of K that resulted in further differentiation of a population (of the 10 indicated in bold in Fig. 5) from the preceding value of K. This justified seven genetic clusters identified by STRUCTURE, which could be associated with geographic locations. The three reference populations with joint or shared cluster memberships gave patterns that correspond with each population's geographic location, discernible from their position in the cluster plot and the joint membership proportions they show. First, East Africans from Eritrea and Somalia have joint African and Middle East cluster memberships, with comparable ranges between ˜30-50% African cluster proportions. Second, North Africans from Algeria have the same joint cluster proportions but ˜10-15% African proportions. Third, Northeast Asian samples from Siberia (HGDP-CEPH Yakut) are predominantly East Asian (mostly 80–90% proportions) but with discernible Middle East cluster membership at ˜5% proportions.

The full listing of K:7 cluster membership proportions for all 2333 reference population samples and 836 test population samples are given in Supplementary File S10. Reference samples from population codes 1–10 are in the same order as the cluster plot of Fig. 5. To help track samples that have unusual cluster membership patterns, those with second or multiple cluster proportions > 0.2 are marked with boxes, while outlier samples (proportions > 0.5 in an alternative cluster) are shown in red at the base of each population set, which are ranked from highest to lowest majority cluster membership proportions. Just 11 reference samples had > 0.5 memberships in alternative clusters to those expected, comprising: 1 Spanish and 4 Tuscan Italians with ME majority clusters; 1 Druze ME (EUR); 1 Bedouin ME (AFR); 3 N AFR Algerians (1 EUR, 2 AFR); and 1 Gujarati Indian (EUR). These results indicate that a small number of individuals from reference populations located close to continental margins have shared memberships with populations from neighboring regions, which is to be expected, given the levels of gene-flow known to have occurred historically in these regions. All East African samples had AFR and ME clusters in high proportions (all but 1/45 > 0.2). In comparison, North Africans had much lower AFR cluster proportions.

Patterns of cluster membership in test samples are shown in Fig. 6 with the order of samples matched to those in Supplementary File S10 (proportions > 0.2 in alternative clusters boxed and marked in green). Test sample genetic cluster patterns are complex and difficult to summarize, particularly in Central Asian EGDP and Eastern European SGDP individuals, but as a whole, largely show expected distributions of membership proportions when compared to their sampling locations. Most West, North, Central and Southern Europeans show clear majority EUR cluster memberships, while individuals from Eastern Europe, Caucasus and West Central Asia have EUR and ME clusters in varied proportions, also seen in Sardinians. The pale blue ME cluster is present in varied proportions in almost all Eurasian populations outside the above parts of Europe and South Asia. Therefore, this genetic cluster can be interpreted to represent 'non-European, West Eurasian' genetic

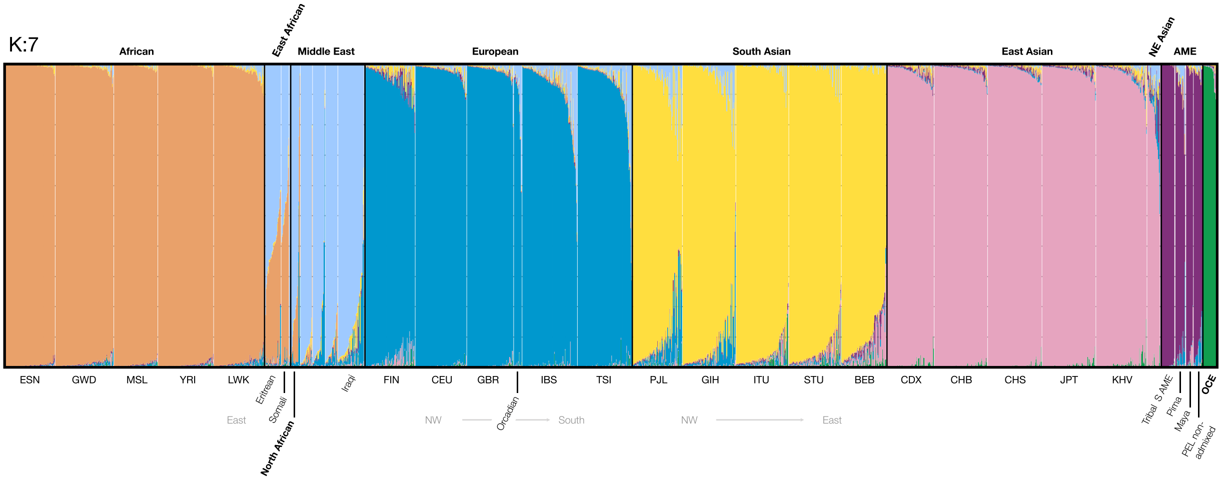

Fig. 5. STRUCTURE cluster plot of the reference data set of 2333 samples from 1KG, SGDP and MAPlex genotyping with established populations of origin. The plot describes merged data from multiple runs using CLUMPAK and represents K:7 genetic clusters identified from distributions of allele and haplotype frequencies amongst the samples. Clusters colored as in Rosenberg et al. 2002 [23], population groups labeled in bold.

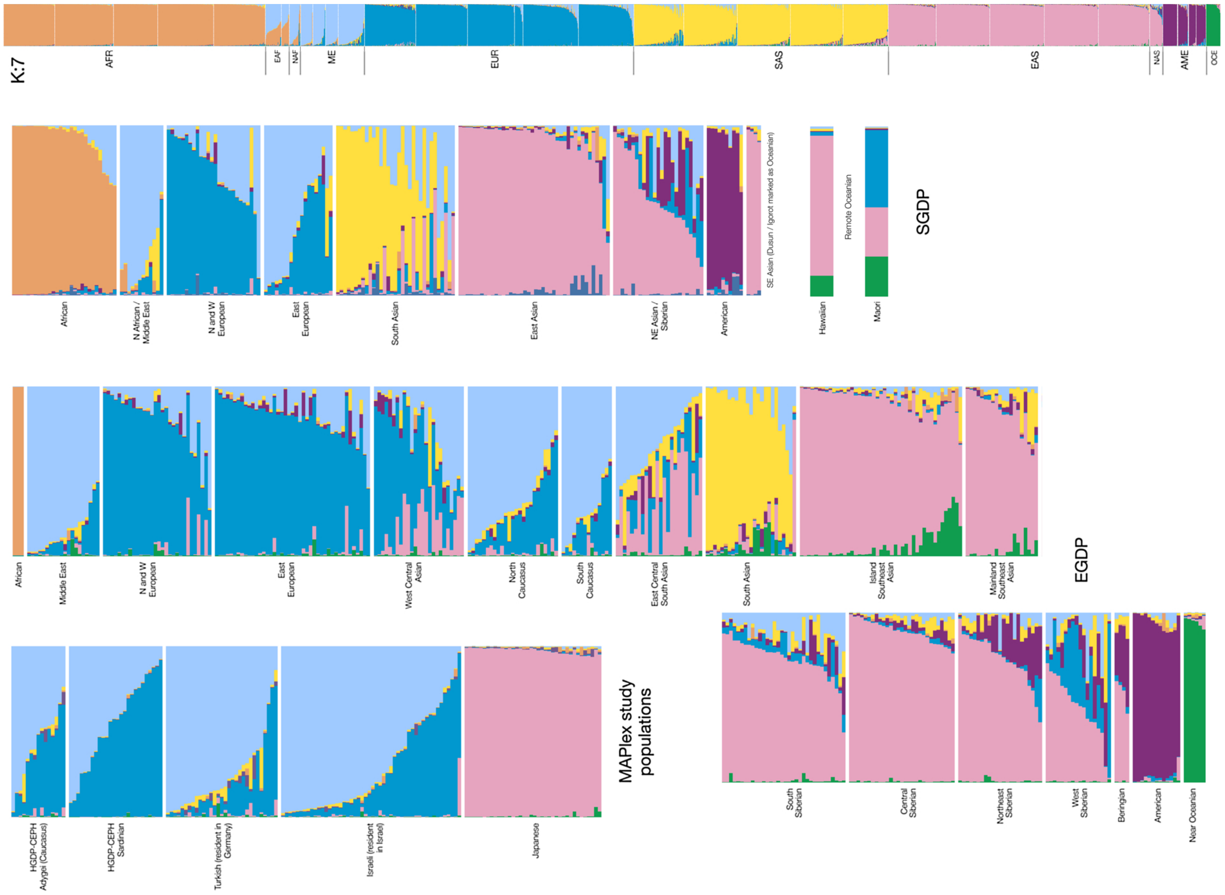

Fig. 6. Enlarged STRUCTURE cluster plots of test samples from SGDP, EGDP and MAPlex study populations produced from the same analytical run identifying the genetic clusters shown in Fig. 5. Samples are arranged into population groups positioned geographically and sequentially (except EGDP West Siberians and MAPlex CEPH Adygei). Enlarged Hawaiian and Maori plots from single individuals indicate atypical cluster membership patterns compared to reference Oceanians and test Near Oceanians from EGDP.

variability. Central South Asian and Siberian populations exhibit complex patterns too, largely dictated by their regional position and most closely neighboring population groups, but these population samples range across very large distances and consequently cluster patterns cannot always be systematically interpreted.

4. Concluding remarks

MAPlex represents the first forensic MPS assay that combines microhaplotypes and SNPs in one multiplex. The relatively balanced distribution of amplicon sizes amongst sequences with polymorphisms and those containing small sets of SNPs in short spans is reflected in well-equilibrated levels of sequence coverage in the 164 markers. This, in turn, indicates that the size reduction of several long but informative candidate microhaplotypes was worthwhile, and consequently MAPlex should have the necessary sensitivity to analyze challenging or low-level DNA typically found in forensic samples. The evaluation of the performance of MAPlex in forensic analyses is the subject of the next phase of development focused on operational validation of the panel.

The preliminary testing of a limited number of population samples in this study indicates enhanced South Asian differentiation by including multiple-allele SNPs and South Asian-informative SNPs. Adding these AIMs increased the divergence between West Eurasian, South Asian and East Asian populations, which has also led to efficient differentiation of Middle Eastern individuals and Europeans. In most cases, microhaplotype loci selected for MAPlex exhibited more variation than multiple-allele SNPs but did not necessarily have more contrasted haplotype frequency distributions. Therefore, despite the prevailing view that the currently described microhaplotypes provide efficient population differentiations [15,16,22,46], discovery of new microhaplotypes should proceed further, and we advocate continued assembly of combined panels of SNPs and microhaplotypes for forensic ancestry analysis. Future forensic ancestry panel designs could take the same approach as that adopted for MAPlex of shrinking the core binary AIM-SNP set and increasing markers with multiple alleles that show informative population differentiations. A current initiative we have started is to improve the measurement of population divergence in multiple allele SNPs and microhaplotypes, with the goal of detecting the best AIMs in each class of marker [47]. The added value of population-specificity for a particular allele in a marker was evident when comparing MAPlex genotypes amongst the populations analyzed. When such alleles have restricted geographical distributions and are absent from most or all other populations, they disproportionately contribute to the differentiation power of the panel compared to loci with contrasting allele frequencies, but which are shared.

The potential to differentiate East Asian sub-populations was an important target in the development of MAPlex, as well as the ability to fully delineate South Asians from Europeans and East Asians, or Near Oceanians from East Asians. The last phase of the development of MAPlex will perform more extensive testing of population samples from the Asia-Pacific region, to achieve finer resolution of Asian and Pacific populations with expanded reference data.

Although this study did not explore the effect of doubling the total number of forensic autosomal AIMs by combining MAPlex and Precision ID MPS ancestry panels, the minimal overlap between these sets and consequent broad spread of divergent markers provides the most powerful approach to forensic ancestry analysis, as well as more efficient mixed DNA detection from multiple-allele loci and a higher total number of binary SNPs. Combination of these autosomal panels with Y and mitochondrial variation provides the potential to further differentiate sub-populations found within the main continental population groups and to address the challenges associated with population admixture detection and analysis.

Acknowledgements

CP, AFA, AMM, MdlP, MVL and the work of this study 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. AFA 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 (Modalidade B, ED481B 2018/010). 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). The authors wish to thank Shop Mallick and Iain Mathieson of the Harvard Medical School, for their guidance on the phasing of SGDP SNP data; and Andrew Pakstis of the Department of Genetics, Yale University School of Medicine, for a critical review of the manuscript.

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.2019.06.022.

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] 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.

[4] M. Eduardoff, T.E. Gross, C. Santos, M. de la Puente, D. Ballard, C. Strobl, C. Børsting, N. Morling, L. Fusco, C. Hussing, et al., Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM, Forensic Sci. Int. Genet. 23 (2016) 178–189.

[5] M. Al-Asfi, D. McNevin, B. Mehta, D. Power, M.E. Gahan, R. Daniel, Assessment of the precision ID ancestry panel, Int. J. Legal Med. 132 (2018) 1581–1594.

[6] K.K. Kidd, W.C. Speed, A.J. Pakstis, M.R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F.R. Friedlaender, J.R. Kidd, Progress toward an efficient panel of SNPs for ancestry inference, Forensic Sci. Int. Genet. 10 (2014) 23–32.

[7] R. Kosoy, R. Nassir, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, P.K. Gregersen, J.W. Belmont, et al., Ancestry-informative marker sets for determining continental origin and admixture proportions in common populations in America, Hum. Mutat. 30 (2009) 69–78.

[8] M.L. Quintana-Murci, R. Chaix, R.S. Wells, D.M. Behar, H. Sayar, R. Scozzari, C. Rengo, N. Al-Zahery, O. Semino, et al., Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor, Am. J. Hum. Genet. 74 (2004) 827–845.

[9] Y. Xue, T. Zerjal, W. Bao, S. Zhu, Q. Shu, J. Xu, R. Du, S. Fu, P. Li, M.E. Hurles, H. Yang, C. Tyler-Smith, Male demography in East Asia: a north-south contrast in human population expansion times, Genetics 172 (2006) 2431–2439.

[10] G. Hellenthal, G.B. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush, S.A. Myers, Genetic atlas of human admixture history, Science 343 (2009) 747–751.

[11] X. Cai, Z. Qin, B. Wen, S. Xu, Y. Wang, Y. Lu, L. Wei, C. Wang, S. Li, X. Huang, L. Jin, H. Li, Genographic Consortium; human migration through bottlenecks from Southeast Asia into East Asia during last glacial maximum revealed by Y chromosomes, PLoS One 6 (2011) e24282.

[12] K.C. Suo, H. Xu, C.C. Khor, R.T. Ong, X. Sim, J. Chen, W.T. Tay, K.S. Sim, Y.X. Zeng, X. Zhang, et al., Natural positive selection and north-south genetic diversity in East Asia, Eur. J. Hum. Genet. 20 (2012) 102–110.

[13] 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.

[14] C. Phillips, J. Amigo, Á. 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.

[15] K.K. Kidd, A.J. Pakstis, W.C. Speed, R. Lagacé, J. Chang, S. Wootton, E. Haigh, J.R. Kidd, Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics, Forensic Sci. Int. Genet. 12 (2014) 215–224.

[16] F. Oldoni, K.K. Kidd, D. Podini, Microhaplotypes in forensic genetics, Forensic Sci. Int. Genet. 38 (2019) 54–69.

[17] C. Tian, R. Kosoy, A. Lee, M. Ransom, J.W. Belmont, P.K. Gregersen, M.F. Seldin, Analysis of East Asia genetic substructure using genome-wide SNP arrays, PLoS One 3 (2008) e3862.

[18] C.X. Li, A.J. Pakstis, L. Jiang, Y.L. Wei, Q.F. Sun, H. Wu, O. Bulbul, P. Wang, L.L. Kang, J.R. Kidd, K.K. Kidd, A panel of 74 AISNPs: improved ancestry inference within Eastern Asia, Forensic Sci. Int. Genet. 23 (2016) 101–110.

[19] O. Bulbul, L. Cherni, H. Khodjet-el-khil, H. Rajeevan, K.K. Kidd, Evaluating a subset of ancestry-informative SNPs for discriminating among Southwest Asian and circum-Mediterranean populations, Forensic Sci. Int. Genet. 23 (2016) 153–158.

[20] C. Phillips, A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M.D. Perez Carceles, Á. Carracedo, P.M. Schneider, M.V. Lareu, Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int. Genet. 7 (2013) 359–366.

[21] 1000 Genomes Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, et al., A global reference for human genetic variation, Nature 526 (2015) 68–74.

[22] 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. 29 (2017) 29–37.

[23] 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.

[24] 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.

[25] N. Hiroaki, F. Koji, K. Tetsushi, S. Kazumasa, N. Hiroaki, S. Kazuyuki, Approaches for identifying multiple-SNP haplotype blocks for use in human identification, Leg. Med. 17 (2015) 415–420.

[26] U. Soundararajan, L. Yun, M. Shi, K.K. Kidd, Minimal SNP overlap among multiple panels of ancestry-informative markers argues for more international collaboration, Forensic Sci. Int. Genet. 23 (2016) 25–32.

[27] I. Halder, M. Shriver, M. Thomas, J.R. Fernandez, T. Frudakis, A panel of ancestry-informative markers for estimating individual bio-geographical ancestry and admixture from four continents: utility and applications, Hum. Mutat. 29 (2008) 648–658.

[28] P. Paschou, J. Lewis, A. Javed, P. Drineas, Ancestry-informative markers for fine-scale individual assignment to worldwide populations, J. Med. Genet. 47 (2010) 835–847.

[29] 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. Uribe Figueroa, et al., Development of a panel of genome-wide ancestry-informative markers to study admixture throughout the Americas, PLoS Genet. 8 (2012) e1002554.

[30] K.B. Gettings, R. Lai, J.L. Johnson, M.A. Peck, J.A. Hart, H. Gordish-Dressman, M.S. Schanfield, D.S. Podini, A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population, Forensic Sci. Int. Genet. 8 (2014) 101–108.

[31] L.M. Huckins, V. Boraska, C.S. Franklin, J.A. Floyd, L. Southam, P.F. Sullivan, Using ancestry-informative markers to identify fine structure across 15 populations of European origin, Eur. J. Hum. Genet. 22 (2014) 1190–1200.

[32] X. Zeng, R. Chakraborty, J.L. King, B. LaRue, R.S. Moura-Neto, B. Budowle, Selection of highly informative SNP markers for population affiliation of major US populations, Int. J. Legal Med. 130 (2016) 341–352.

[33] C. Santos, C. Phillips, M. Fondevila, R. Daniel, R.A.H. van Oorschot, E.G. Burchard, M.S. Schanfield, L. 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.

[34] H. Thorvaldsdóttir, J.T. Robinson, J.P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, Brief. Bioinform. 14 (2012) 178–192.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] The *Snipper* Page for Executing Combined Bayes-PCA Analysis of Multiple SNP Profiles, (2019) Accessed February http://mathgene.usc.es/snipper/analysismultipleprofiles.html.

[40] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo, M.V. Lareu, An overview of STRUCTURE: applications, parameter settings, and supporting software, Front. Genet. 4 (2013) 98.

[41] 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.

[42] H. Li, N. Mukherjee, U. Soundararajan, Z. Tarnok, C. Barta, S. Khaliq, A. Mohyuddin, S.L. Kajuna, S.Q. Mehdi, J.R. Kidd, K.K. Kidd, Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia, Am. J. Hum. Genet. 81 (2007) 842–846.

[43] S. Gu, H. Li, A. Pakstis, W.C. Speed, D. Gurwitz, J.R. Kidd, K.K. Kidd, Recent selection on a class I ADH locus distinguishes southwest asian populations including ashkenazi jews, Genes 9 (2018) E452.

[44] 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 SNP*for*ID project, Int. J. Legal Med. 132 (2018) 435–440.

[45] 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.

[46] O. Bulbul, A.J. Pakstis, U. Soundararajan, C. Gurkan, J.E. Brissenden, J.M. Roscoe, B. Evsanaa, A. Togtokh, P. Paschou, E.L. Grigorenko, et al., Ancestry inference of 96 population samples using microhaplotypes, Int. J. Legal Med. 132 (2018) 703–711.

[47] E.Y.Y. Cheung, C. Phillips, M. Eduardoff, M.V. Lareu, D. McNevin, Performance of ancestry informative SNP and microhaplotype markers, paper submitted, April (2019).