Forensic Science International: Genetics 45 (2020) 102213

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems

M. de la Puente[a,b,*], C. Phillips[a,*], C. Xavier[b], J. Amigo[c], A. Carracedo[a,c], W. Parson[b,d], M.V. Lareu[a]

[a] *Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain*
[b] *Institute of Legal Medicine, Medical University of Innsbruck, Austria*
[c] *Fundación Pública Galega de Medicina Xenómica (FPGMX), Santiago de Compostela, Spain*
[d] *Forensic Science Program, Pennsylvania State University, PA, USA*

ARTICLE INFO

Keywords:
Microhaplotypes
SNPs
Massively parallel sequencing
MPS
MiSeq
Ion S5
Mixed DNA

ABSTRACT

A large number of new microhaplotype loci were identified in the human genome by applying a directed search with selection criteria emphasizing short haplotype length (< 120 nucleotides) and maximum levels of polymorphism in the composite SNPs. From these searches, 107 autosomal microhaplotypes and 11 X chromosome microhaplotypes were selected, with well-spaced autosomal positions to ensure their independence in relationship tests. The 118 microhaplotypes were assembled into a single multiplex assay for the analysis of forensic DNA with massively parallel sequencing (MPS). A single AmpliSeq-adapted primer set was made for Illumina MiSeq and Thermo Fisher Ion S5 MPS platforms and the performance of the assay was comprehensively evaluated in both systems. Five microhaplotypes showed critical sequencing failures in both MPS platforms and were removed, while a further 13 required manual checks and the application of sequence quality thresholds in one or both systems to ensure the successful analysis of low-level DNA in these loci. The targeting of short microhaplotype spans during marker selection, with an average length of 51 nucleotides in the 118 loci, led to a high level of sensitivity for the panel when sequencing the very degraded DNA typically encountered in forensic casework and the identification of missing persons.

1. Introduction

Microhaplotypes are small sets of closely-sited single nucleotide polymorphisms (SNPs) that show contrasting allele frequencies [1–3] and consequently have a range of haplotypes formed by the combined composite SNP alleles. The haplotypes have higher levels of polymorphism than any of the individual SNPs and such variation rises as the number of SNPs in the microhaplotype and the levels of contrast in their allele frequencies increases (although immediately adjacent SNPs often have identical frequencies, making all but one of them redundant [4]). Very few microhaplotypes with three or more composite SNPs have the full extent of haplotypes present (i.e., all 8 possible haplotypes in 3-SNP loci; 16 with 4-SNPs; 32 with 5-SNPs, etc.). The phase of the SNP alleles in the microhaplotype - their sequential combination on each DNA strand - must be detected in order to accurately describe each haplotype. For this reason, the emergence of massively parallel sequencing (MPS) techniques, where the phase of SNP alleles on each sequenced strand is detected, has brought microhaplotype analysis into mainstream forensic use. Microhaplotypes have a number of advantages for forensic identification compared to both SNPs typed as individual loci, and short tandem repeats (STRs) used in routine forensic profiling. Individual SNPs can be amplified in much shorter fragments than STRs, whose PCR must target often long repeat regions, making SNP panels a better choice for analysis of very degraded DNA. However, SNPs can only reach a maximum 50 % heterozygosity, in “perfect” loci that have the same 0.5 frequency for each allele. While this can extend to 62.5 % in tri-allelic SNPs and 75 % in tetra-allelic SNPs [5], such levels of variability are only seen in a fraction of human SNP variation, so are difficult to find in sufficient numbers for forensic use. STRs require longer amplicons than SNPs to ensure the repeat region is captured, and when STRs are analyzed by MPS, additional factors can influence the minimum size of the region to be amplified and sequenced, such as the need for robust alignment of low complexity sequence close to the repeat region and avoiding common SNPs in these flanking regions. More importantly, STR repeats create PCR slippage, which form stutter products detected in both capillary electrophoresis

[*] Corresponding authors at: Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain.
E-mail addresses: m.delapuente.vila@gmail.com (M. de la Puente), c.phillips@mac.com (C. Phillips).

https://doi.org/10.1016/j.fsigen.2019.102213
Received 1 August 2019; Received in revised form 31 October 2019; Accepted 27 November 2019
Available online 30 November 2019

and at comparable levels in MPS. The presence of stutter hinders the accurate identification of minor contributor alleles with the same size, especially when stutter products reach levels above 10 % of the parent repeat allele. Lastly, since the SNPs in microhaplotypes have much lower mutation rates than STRs, they are ideal loci for relationship testing as their levels of variation are higher than SNPs, but second order exclusions created by mutation are almost never seen. Therefore, microhaplotypes offer short amplicon analysis of challenging forensic samples, higher levels of polymorphism than individual SNP loci plus an absence of stutter products or high mutation rates typical of STR genotyping; making them compelling genetic loci for improved forensic DNA analysis in the MPS age.

A review of published microhaplotypes considered suitable for forensic analysis with MPS [3,6–9], indicates that many require amplicon sizes longer than ∼180 nucleotides (nt), when taking into account primer sequences. Proposed microhaplotype panels range from short haplotype spans ([6,7] average haplotype spans of 34 nt); to mid-range ([3] 113 nt and [8] 108 nt); to very long ([9] 263 nt). The longest microhaplotypes are less likely to perform well when sequencing very degraded DNA. While it is possible to reduce their size by trimming composite SNPs from either side, when this involves one or more highly polymorphic loci there can be a marked reduction in the microhaplotype's forensic informativeness. The present study reports a dedicated program to identify new microhaplotypes in the human genome, and their compilation and evaluation for forensic use, with emphasis on loci that could be short in length. Candidate microhaplotypes needed to be sufficiently polymorphic based on 1000 Genomes haplotype frequencies, and suitable for complementing other panels of forensic identification markers typed with MPS (including forensic assays in development that have small numbers of the best microhaplotypes published so far [3]). We brought together 118 novel candidate microhaplotypes from autosomal and X chromosome sites with an average haplotype span of ∼51 nt into one MPS multiplex suitable for FFPE (Formalin-Fixed Paraffin-Embedded) DNA, with amplicon lengths of 125–175 nt, including primer sequences. At the time of their identification and compilation, all the microhaplotypes we collected were new discoveries. The raised levels of informativeness that the selected microhaplotypes offered for relationship testing were assessed by simulating challenging pairwise kinship analysis scenarios, applying European haplotype frequency estimates from 1000 Genomes. The design and optimization of the MPS multiplex based on AmpliSeq™ primer chemistry is described for both Illumina MiSeq and Thermo Fisher Ion S5 sequencing platforms. We developed an in-house haplotype phasing pipeline based on publically available open-source software, which will aid the analysis of mixed DNA with the microhaplotypes compiled. Minor, but significant, differences were observed in the sequencing performance of each MPS platform and these impacted the final set of microhaplotypes that we established for enhanced forensic identification applications using this assay.

2. Material and methods

To avoid confusion, we differentiated the terminology used to describe mixture, panel and SNP components as: mixture *contributors*; the panel's microhaplotype *component* loci; and *composite* SNPs making up each microhaplotype.

2.1. Microhaplotype discovery, screening and primer design

The VCF genotypes from publicly available 1000 Genomes Project Phase III data (herein 1KG) [10] were queried for loci with two or more polymorphic SNPs showing minor allele frequencies (MAF) higher than 0.1, in segments shorter than 120 nt. Searches resulted in ∼35,000 candidate loci. Although internal codes were applied to the final choice of microhaplotypes during their development and are used in the text, we advocate applying the rs-numbers of the composite SNPs to define each candidate locus, which allows their straightforward identification from current genome variant databases (e.g. using rs28503881-rs4648788-rs72634811-rs28689700 to describe microhaplotype '1pA').

Chromosomal regions with at least 10 Mb separation were defined for autosomal candidate loci searches and 5 Mb for the X chromosome, with final numbers adjusted to the length of each chromosome to minimize levels of linkage disequilibrium between syntenic markers [11]. Candidate loci were placed into subsets based on their distributions, as defined above, and ranked by informativeness measured by the overall gene diversity value (GD) of each locus (i.e. a global GD value from combining all 1000 Genomes populations). Because 1KG haplotypes were counted to measure any one candidate's informativeness, we applied GD as a measure of microhaplotype variability, calculated as: $\mathbf{n(1 - \Sigma \boldsymbol{p_i}^2)/(n - 1)}$, following Nei [12], where n and p_i are the total number of samples and the relative frequency of the i-th allele, respectively. GD is directly related to the effective number of alleles (Ae) – a metric widely used to assess microhaplotypes for forensic identification or mixture analysis purposes [13].

Primer design was accomplished with Ion AmpliSeq Designer (TFS) using GRCh37/hg19 as the reference genome and selecting parameters for a single-pool Hotspot design targeting FFPE DNA (i.e. to obtain short amplicons of 125–175 nt).

From each subset, the most informative locus fulfilling prescribed quality requirements was incorporated into the panel. Quality screening included: (i) a visual inspection of the sequence with the Ensembl genome browser [14] to discard candidates showing features that could hinder alignments, such as long poly-tracts, repetitive regions, Indels or structural variants spanning the microhaplotype position; (ii) MH sequences that aligned with multiple genomic positions in nucleotide BLAST [15] were discarded, to avoid duplicated regions; (iii) the primer pair designs obtained with AmpliSeq Designer were submitted to the In-silico PCR tool (UCSC Genome browser [16]) using both GRCh37/hg19 and GRCh38/hg18 genome builds and primer designs that generated unexpected amplicons were discarded, to reduce the possibility of non-specific amplification; and (iv) remaining primer pairs were submitted to SNPCheck v3 [17] to scan for variants sited in the primer binding regions that could hinder annealing.

2.2. MH panel informativeness metrics and evaluation

For each potential MH marker, composite SNP variation was screened to ensure only variants with a MAF ≥ 0.05 were defined. Allele frequency, gene diversity and random match probability estimates were calculated from 1KG data for African, East Asian, European, South Asian and admixed American population groups (herein AFR, EAS, EUR, SAS, AMR, respectively) [10]. In order to assess the informativeness of the whole panel under diverse kinship analysis scenarios, simulations of different pedigrees for: i) full-siblings; ii) half-siblings; iii) first-cousins; and iv) second-cousins, were performed in *Familias* v3.1.9.6 [18,19] for the compiled autosomal MH markers using 1KG European haplotype frequency estimates. The number of simulations for each pedigree was set to one million and downstream processing of data and plotting of the resulting likelihood ratios was carried out with R v3.5.0 [20].

2.3. DNA samples and forensic validation

Five Coriell cell line DNAs NA18498, NA06994, NA07000, NA07029 and NA11200; and three forensic control DNAs 2800 M, 9947A (Promega) and 007 (Applied Biosystems) were analyzed in order to implement the panel on MiSeq and Ion S5 massively parallel sequencing (MPS) platforms. This data evaluated the quality of the sequences obtained, including parameters for coverage, levels of misincorporation, allele balance and strand bias. The use of Coriell cell line DNAs allowed assessments of concordance of MPS genotypes with publically available data from 1KG [10] for NA18498, NA06994 and

NA07000 and Simons Foundation Human Genome Diversity Project (herein SGDP) [21] for NA11200. Both projects include in their public releases phased data for SNP sets in short sequence segments and these were directly compared to the haplotypes we obtained through a custom MH calling pipeline. Moreover, samples NA06994 (father), NA07000 (mother) and NA07029 (son) are a confirmed family trio allowing a simplified initial assessment of Mendelian inheritance.

Evaluation of forensic sensitivity of both platforms was based on a set of artificial samples which were established to assess: (i) sensitivity to low-level DNA, using dilutions of 2800 M for DNA input concentrations (in duplicated libraries) of 0.5 ng, 0.25 ng, 125 pg, 62.5 pg and 31.25 pg; (ii) sensitivity to degraded DNA, using 007 DNA artificially degraded by sonicating at 40 kHz for 0, 90, 180 and 240 min. Degradation status of the sonicated samples was gauged by typing core STRs with PowerPlex® ESI 17 (standard protocols). Artificial DNA mixtures were prepared from ~1 ng/μL of 2800 M and 9947A at volume ratios of 1:1, 1:3, 1:7 and 1:15 and run on the Ion S5™ system alone.

2.4. MPS library construction, template preparation and sequencing

Libraries for the Ion S5™ MPS system (herein Ion S5) were constructed from 1 ng of input DNA (except sensitivity samples) using the Precision ID Library Kit (TFS) and Ion Xpress™ Barcode Adapters (TFS), following manufacturer's protocols. Libraries were quantified with the Ion Library TaqMan™ Quantitation Kit (TFS) following manufacturer's protocols, diluted to 30 pM and pooled in equimolar proportions for template preparation (a maximum 32 libraries pooled per chip). Templates were constructed with the Ion S5™ Precision ID Chef & Sequencing Kit and loaded into Ion 530™ Chips using the Ion Chef™, following the manufacturer's protocols. Sequencing with the Ion S5 detector was set for a read length of 200, with 500 flows.

Libraries for the MiSeq FGx™ MPS system (herein MiSeq) were constructed from 1 ng of input DNA (except sensitivity samples) using the AmpliSeq™ Library PLUS (96 reactions) for Illumina® and AmpliSeq™ CD Index Set A for Illumina®, following manufacturer's protocols, but modifying the number of library amplification cycles from 7 to 10. Libraries were quantified using the Agilent High Sensitivity DNA Kit (Agilent Technologies, AG) and the 2100 Bioanalyzer Instrument (AG). Libraries were then diluted to 2 nM and pooled in equimolar proportions (a maximum of 32 libraries pooled per flow cell). Template reactions and sequencing were performed with the MiSeq detector using the MiSeq Reagent Kit v2 (300-cycles), following manufacturer's protocols (including the suggested final loading concentration of ~7 pM).

2.5. Sequence analysis and compilation of phased haplotypes

Characterization of composite SNP genotypes in each MH was performed using the SNP calling pipelines offered by TFS and Illumina. Ion S5 data was analyzed using Torrent Suite v5.2.2 (TFS) and reference genome GRCh37/hg19. Genotype calls and other information of composite SNPs (coverage, total of A/C/T/G and forward/reverse reads, allele read frequencies, etc.) were obtained using the HID SNP Genotyper v5.2.2 plugin (TFS). Data generated with the MiSeq was analyzed using the Local Run Manager DNA Amplicon Analysis Module v2.1.0 (Illumina) on the Local Run Manager Off-Instrument v2.0 (Illumina), with reads aligned to GRCh37/hg19 using BWA-MEM algorithm. Output BAM/BAI files were analyzed with bam-readcount [22] to recover information about coverage, total of A/C/T/G and forward/reverse reads. The composite SNP genotypes were manually called by applying the same default parameters as Genotyper, i.e. a minimum coverage of six reads and minimum allele read frequency of 0.1 for heterozygotes. Downstream processing and plotting of data were performed using R v3.5.0 [20].

The haplotypes of each MH locus (as phased composite SNP genotypes) and their levels of sequence coverage cannot be obtained with TFS or Illumina software, so we developed and optimized the following custom MH calling pipeline. First, a synthetic partial reference genome was constructed by assembling 100 kb sequence segments extracted from GRCh37/hg19 comprising each MH amplicon. Next, raw reads in FASTQ format from both platforms were aligned to the partial reference genome using Burrows-Wheeler aligner (BWA) [23]. Alignments were further processed with SAMtools [24] to create the required input files for running the microhaplot R package [25], these included a VCF file of the composite SNPs for each MH and alignments in SAM format, sorted and filtered out of short reads (< 100 bp) and low-quality alignments (mapping quality < 30). Microhaplot output included a raw table of allele strings and depth per MH, that were further filtered by minimum coverage per allele (min_cov), set at 15 reads and minimum allele read frequency (min_allele_frequency) set to 0.1 for single-donor samples and 0.02 for mixtures. Scripts and guidelines for processing raw reads in order to obtain phased MH alleles are available upon request.

The visualization of alignments produced by Torrent Suite, Local Run Manager and BWA was made using IGV v2.3.40 [26].

3. Results and discussion

3.1. MH panel characterization and informativeness

Supplementary Table S1 contains details of 107 autosomal and 11 X chromosome microhaplotypes (MHs) incorporated into the single pool Ampliseq design (details available upon request). Fig. 1 shows the distribution of these 118 MH loci across the human genome. The spacing rules applied during selection produced an evenly distributed set of markers, showing at least 10 Mb distance for autosomal loci and 5 Mb for X chromosome loci. During their development MHs were given

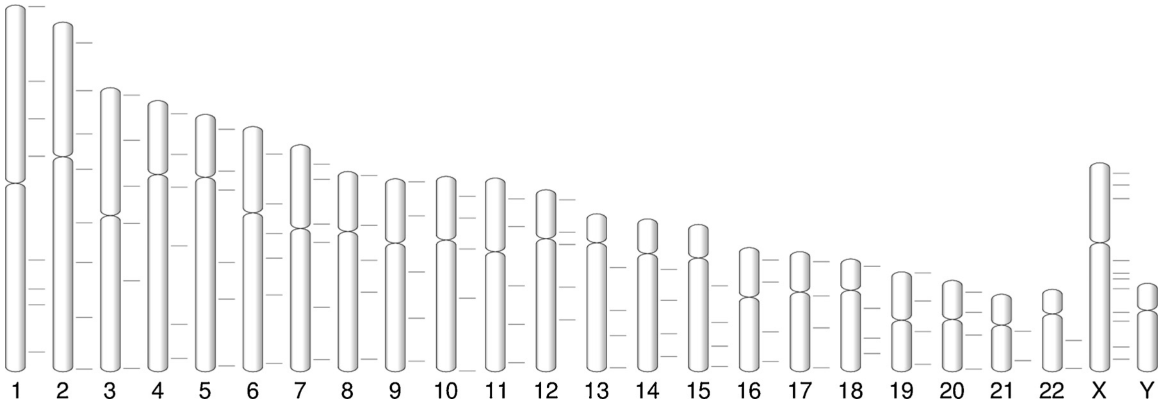

Fig. 1. Ideogram representing chromosome positions of the 107 autosomal and 11 X chromosome MHs included in the panel.

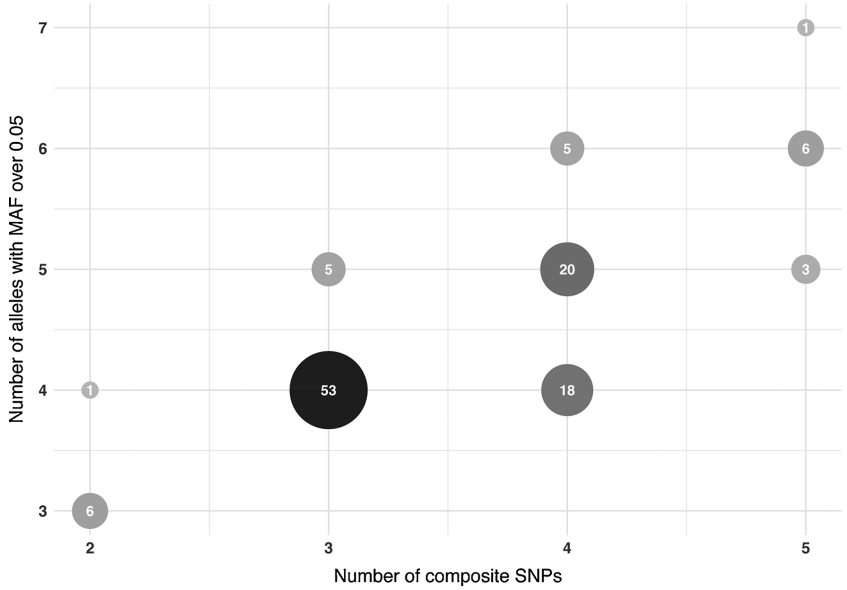

Fig. 2. Bubble graph of numbers of MHs with the range of total alleles and composite SNPs. MAF: minor allele frequency.

internal codes comprising the chromosome number or X, p or q-arm, and position (letters in alphabetical order according to chromosome coordinates).

As shown in Fig. 2, most MHs in the panel consist of 3 SNPs and four common haplotypes when compiling only composite SNPs with MAF ≥ 0.05. Fig. 3 shows values of average GD in the whole 1KG dataset (26 populations) and average values within “superpopulations” as defined by 1 KG, to give a closer insight into the actual informativeness of each locus. The average autosomal MH GD value was 0.655 while X chromosome MHs gave a lower average value of 0.574. Differences between the GD measures of 1KG AFR, EAS, EUR, SAS, AMR population groups are due to haplotype frequency differences among these populations. Bar charts of haplotype frequency data for these 1KG population groups are compiled in Supplementary File S1.

Fig. 4 shows cumulative random match probabilities across the 107 autosomal MHs. Cumulative values when including all populations reach a minimum of 3.72E-104. Even with the expected differences amongst the five population groups, the panel is highly informative in each group. These values, in decreasing discrimination power are AMR: 2.78492E-95; AFR: 5.16273E-93; SAS: 1.11732E-93; EUR: 4.02124E-91; EAS: 8.41441E-84. Expected theoretical cumulative values for a panel composed of 107 perfectly balanced two-haplotype markers (i.e. a frequency of 0.5 in each haplotype) reach values of 2.6E-46; while panels of three, four and five haplotype markers would reach values of 4.3E-79, 1.5E-103 and 5.4E-123, respectively. Since Fig. 4 indicates a consistent slope, the removal of any markers during the evaluation process would have a comparable and equally marginal effect on the overall discrimination power of the panel.

The results of the kinship analysis simulations for full sibs, half sibs, first cousins and second cousins are represented in Fig. 5. The distribution of likelihood ratios (LRs) of related-as-claimed and unrelated individuals does not overlap for half and full siblings, with average LRs of related pairs of 1E18.7 and unrelated pairs of 1E4.7. The degree of overlap in the LR distributions of first cousins suggests the need for additional markers in such cases; while analysis of second cousins requires a much higher number of markers.

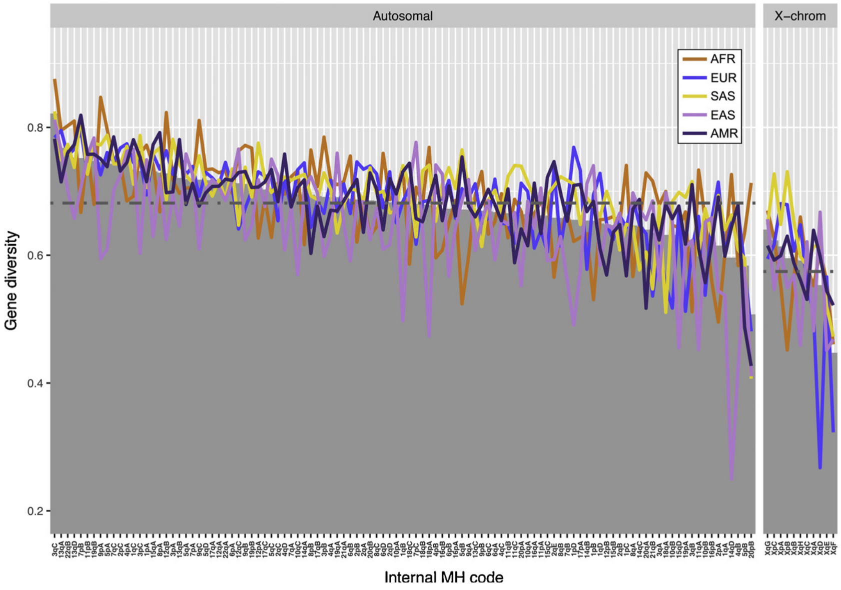

Fig. 3. Gene diversity (GD) values of the autosomal and X chromosome MHs included in the panel. MHs are ranked according to their GD values. Average values for the 26 populations of the 1000 Genomes dataset are represented as grey bars with mean values afor autosomal and X chromosome in dashed lines. Color lines represent mean values for the 1000 Genomes populations grouped into African (AFR), admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS) as in the legend.

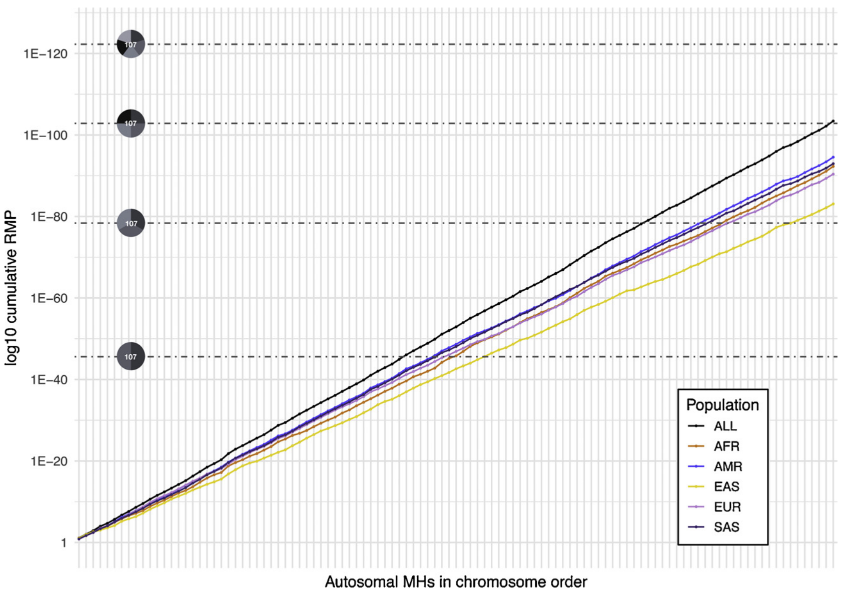

Fig. 4. Cumulative random match probability (RMP) values of the autosomal MH loci for the whole 1000 Genomes Project dataset (ALL), and for population groups: African (AFR), admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). Dashed lines represent cumulative values of a total of 107 perfectly balanced two-haplotype, three-, four- and five-haplotype loci (bottom to top, respectively).

3.2. Implementation of the panel on MiSeq and Ion S5 platforms

The 8-sample concordance set (NA18498, NA06994, NA07000, NA07029, NA11200, 2800 M, 9947A and 007) and a non-template control were run on both platforms to assess sequencing quality and genotype concordance, as well as to test or adjust the manufacturer's recommended protocols.

3.2.1. Assessment of sequencing quality

The two MPS platforms currently applied to forensic DNA analysis and used in this study are based on different sequencing strategies. MiSeq sequencing-by-synthesis [27] uses a reversible terminator chemistry: nucleotides are modified with a label that servers as a terminator and blocks the polymerase reaction. After each extension cycle, the fluorescent dye is recorded to identify the base and the terminator is cleaved off to allow the incorporation of the next base. Ion S5 uses semiconductor sequencing [28]: the different nucleotides are sequentially added to the medium and semiconductor detectors record pH changes produced by the incorporation of nucleotides to the synthesis strand. Nevertheless, a series of universal sequence quality parameters can describe the sequencing output obtained from both platforms.

3.2.1.1. Sequence coverage. Supplementary Table S2 includes performance details of the three MPS runs included in this study: (i) Ion S5 run 1 combining the concordance dataset and the forensic sensitivity study; (ii) Ion S5 run 2 including mixed samples; and (iii) a single MiSeq run, equivalent to the first Ion S5 run. Ion S5 run 1 reached an acceptable 75 % of loading efficiency while Ion S5 run 2 dropped to 30 %. However, this second run included only 18 equimolar pooled libraries (compared to the first run, composed of 32 libraries) - so all the mixture samples reached an adequate coverage of at least 100,000 raw reads. The MiSeq run presented a rather high number of clusters (1317 ± 23 K/mm^2); but a good percentage (88.73 %) passed the quality filter. Calibration of the number of samples per MPS run for forensic purposes is a challenging task that requires previous knowledge of the quality of the sample and the characteristics of

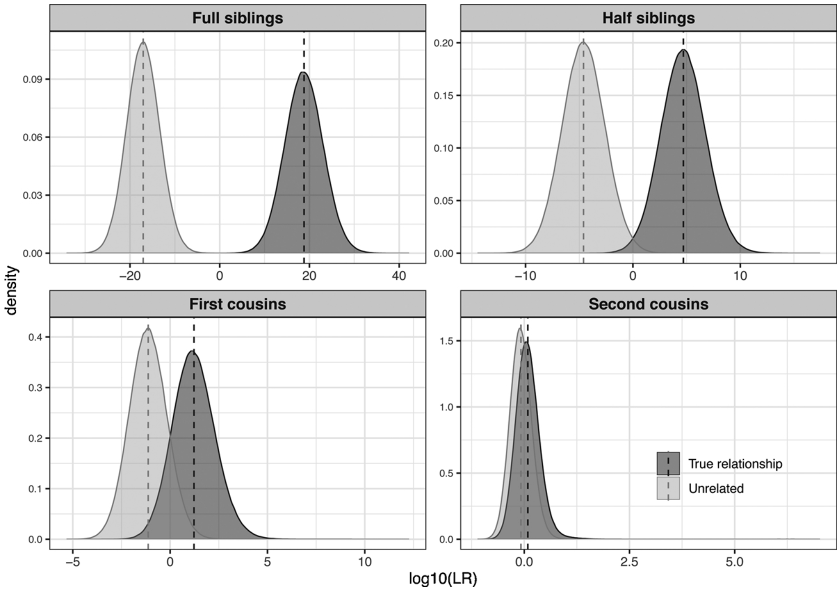

Fig. 5. Distribution of likelihood ratios from one million pedigree simulations of unrelated vs related-as-claimed individuals using *Familias* in the kinship scenarios of: full-siblings; half-siblings; first-cousins and second-cousins. Autosomal MHs were used and European frequencies were applied.

each platform; but libraries often reach quantification results and quantities that allow their analysis across several runs.

Supplementary Fig. S1 shows total coverage values per MH marker in both platforms for the concordance samples. Read coverage means across MHs reached values around two times higher in the MiSeq (6729.6 reads) compared to the Ion S5 (3159.3 reads). Several factors can affect the final coverage, including the capacity of the chip/flow cell, the library normalization process, the loading/cluster formation efficiency or the posterior quality filtering process. However, the main difference between platforms is that the MiSeq sequences each library fragment in both directions, producing paired reads, compared with single, unidirectional reads in the Ion S5. Taking into account the paired reads and the fact that MiSeq libraries undergo 10 cycles of amplification, coverage reached comparable values.

High numbers of sequence reads were obtained for most MHs, with only six markers showing mean coverage values lower than 400 reads in both platforms: 6pB, XpB, 17qC, 16pB, 7pC and 14qA (in increasing order of mean coverage). When comparing normalized coverage (MH coverage/total sample coverage), Supplementary Fig. S2 shows these markers have coverage about 10 times lower than average. The fact that the same markers underperform for both platforms points to differences among the MHs in the amplification success rates of the initial capture PCR. Further optimization is required to achieve better coverage balance across the individual MHs in the multiplex. However, it is not possible to configure the relative primer concentrations of component MHs when designing the primer pool and the low volume of individual primer pair stocks obtained is insufficient for such optimization, as well as a lack of data about initial concentrations of the AmpliSeq primers in the pool.

Low coverage directly affects allele calls as the probability of allele imbalance, allele drop-out and locus drop-out increases. When sufficient coverage is achieved MH markers can be reliably called, especially for single-source samples, although a manual revision of genotypes provides important additional information from the sequence data. Particular care is necessary when genotyping mixtures as the ability to detect a minor contributor is reduced in a proportion of loci.

3.2.1.2. Strand Bias. Supplementary Fig. S3 shows strand bias (as forward or positive coverage/total coverage) obtained from both platforms. Most MHs are balanced, especially on the MiSeq platform, where values are closely clustered around 0.5. Ion S5 values show a wider distribution in the 0.6-0.4 range. This difference likely relates to the fact that paired reads of MiSeq were not filtered. However, extreme values close to 1 (all reads forward) and 0 (all reads reverse) were only found with MiSeq for MH loci 2qC and 12pA (forward) plus XqA and 3qC (reverse). Even when a minimum coverage in both strands was seen, allelic proportions were maintained in both strands and therefore concordant calls can be expected even in those markers with strong strand bias.

3.2.1.3. Base misincorporation. Base misincorporation values (calculated as percentage of reads from non-allelic bases in the SNP site/total SNP coverage) are represented for each sample and as mean values per platform in Supplementary Fig. S4. Prior knowledge of misincorporation helps to reduce the minimum allele read frequency threshold with a certain degree of confidence, to a value that minimizes both the drop-out and drop-in probabilities. This process is crucial for the detection of minor contributors in mixture analysis. Mean percentages of misincorporation across composite SNPs were 0.25 % ± 0.73 % for Ion S5 and 0.16 % ± 0.51 % for MiSeq. Supplementary Table S3 lists SNPs with a mean percentage misincorporation higher than 1.5 % in either platform. The cause of misincorporations was investigated by scrutiny of the alignments in IGV, which indicated 3pC, 19qB and 12qA had context sequence features that could lead to uncertain genotype calls, as described below, so they were excluded from the panel.

First, SNP rs1557912 in MH 3pC had the highest mean misincorporation rate (~5 %) with Ion S5. This SNP is sited in a repetitive region AAAAAAA(A/G)AAGAAA. Misaligned strands that do not reach the target SNP site plus flanking Indels were observed (see Supplementary File S2-D) in a higher proportion of Ion S5 reads compared to MiSeq reads; leading to imbalanced A and G reads in heterozygotes. Second, SNPs rs10404533 and rs10404915 (both in MH 19qB) were unreliably called in both platforms. IGV scrutiny of Ion S5 alignments revealed a high proportion of low mapping quality reads causing genotype discordancies between different mapping quality thresholds (see Supplementary File S2-K.1). These observations are consistent with non-specific amplification of homologous regions that produce reads spuriously aligned to the target MH genomic position. Indeed, some long sequences appeared to be aligned ~20 and ~30 kb downstream of the target region. Consensus sequences of these reads were aligned to the target reference sequence using BLASTn and resulted in percentages of identity over 90 % (see Supplementary File S2-K.2). Third, SNP rs7954300 in MH 12qA was unreliably called with Ion S5, while the other SNPs in 12qA had mean misincorporation values ~1 % in both MiSeq and Ion S5 (see Supplementary Fig. S4). IGV visualization of the sequences (see Supplementary File S2-G) reveals several samples with allelic imbalance for these SNPs leading to genotype discordancies between platforms.

The remaining SNPs listed in Supplementary Table S3; 15qB, 1pC, XqA, 13qD and 2pC are sited in repetitive regions (see Supplementary File S2-I, A, L, H and C, respectively) and consequently had high mean misincorporation values with Ion S5 (previously reported for this platform [29–31]). However, correct calls were obtained by applying a minimum allele read frequency threshold of 10 %. Nevertheless, particular care is needed with mixtures as misincorporations can be mistaken for the alleles from a minor contributor.

3.2.1.4. Allele read frequency balance. Supplementary Fig. S5 represents major allele read frequencies (as major allele reads/total coverage per SNP). In both platforms, frequency values closely cluster to 1 in homozygotes and close to 0.5-0.6 in heterozygotes. Heterozygous SNPs in the same amplicon, i.e. within any one MH, show similar frequency values for the same sample and platform as would be expected from their phased status. However, some MHs consistently differed from this pattern. First, MHs 7pC, 14qA, 17qC and XpB had several heterozygous samples with major allele frequencies between 0.6 and 0.95 (see Supplementary Fig. S5). These markers are already recognized as underperforming in coverage (Section 3.2.1.1) - underlining that lack of coverage can lead to stochastic imbalance in heterozygotes or even allele drop-out. Second, MHs 3pC, 12qA and 19qB also showed several outliers as a consequence of the misincorporations described above (Section 3.2.1.3). Third, MHs 1qC and 19qA showed imbalanced heterozygous genotypes with Ion S5, confirmed by visual inspection of IGV alignments (Supplementary File S2-B and, J, respectively). However, results from the MiSeq platform showed concordant and balanced genotypes in the same samples, excluding a preferential amplification of one allele during PCR. Therefore, causes of imbalance could be due to the filtering/trimming of raw sequences made by the Torrent Suite software before alignment. These MHs should be manually checked when typed with Ion S5.

3.2.1.5. Sequence baseline. Supplementary Fig. S6 shows total coverage per MH of the non-template control for both platforms. Mean total coverage across MHs reaches values about two times higher in the MiSeq (47.99 reads) compared to the Ion S5 (23.72 reads). These values are comparable when accounting for paired reads and are more than 100 times lower than those from samples at 1 ng of input DNA. Non-template reads were randomly distributed across markers.

3.2.1.6. IGV inspection of sequence alignments. The in-house MH calling pipeline we developed starts with raw reads from both platforms that undergo the same alignment, based on BWA, to synthetic partial reference genome. Alignments from both TFS and Illumina SNP calling analysis pipelines and the in-house MH calling BWA pipeline were scrutinized and compared in IGV.

Despite the three alignment methods providing almost identical outputs, it was notable that two MHs showed differences that pointed to genotyping uncertainty. First, BWA alignments of 5qD (see Supplementary File S2-E) had several contiguous SNPs without locked phases that resulted in more than 3 haplotypes being called for single-source samples. Second, 10qC (see Supplementary File S2-F) showed strongly discordant alignments between platform pipelines. BWA alignments were analogous for both platforms and provided concordant MH-allele calls, but showed reads with several contiguous SNPs that could account for non-specific amplification.

3.2.2. Concordance

Concordance of the obtained MH alleles was evaluated in two ways: between platforms and by comparison with 1KG and SGDP databases. Supplementary Table S4 lists all discordancies and no-calls found from these analyses.

No-call rates of 1.91 % with MiSeq and 2.54 % with Ion S5 were observed in the whole concordance set of 118 MHs in 8 samples, corresponding to 18 and 24 drop-outs out of 944 loci, respectively. All locus drop-outs were restricted to MHs 6pB, 17qC and XpB; already characterized as underperforming in terms of sequence coverage (see Section 3.2.1.1).

Concordance comparisons with 1KG databases were confined to NA06694, NA0700 and NA18498 DNAs and was 96.5 % with MiSeq (12 discordancies in 346 called genotypes) and 97.4 % with Ion S5 (9 discordancies in 345 called genotypes). Concordance comparisons with SGDP NA11200 was based on 117 MHs (no data for MH 10qC) and was 100 % for both. Inter-platform concordance was 98.6 % (13 discordancies in 920 called genotypes).

The bulk of discordant genotypes occurred in MHs 3pC, 5qD, 7pC, 10qC, 12qA, 16pB and 19qB - underperforming as described above. However, MHs 1pD, 2pC and XqB showed discordancies consisting of allele drop-ins only in NA06994, possibly as a result of contamination of this reference DNA.

Finally, no Mendelian incompatibilities were found in the family trio NA06994-NA0700-NA07029; but MHs 12qA and 19qB gave high levels of misincorporation as previously described (Section 3.2.1.3).

3.2.3. Underperforming MHs in the panel

Taking into account the information provided by the evaluation of the sequencing quality parameters in Section 3.2.1 and concordance studies in Section 3.2.2, the panel's component loci were categorized as: (i) MHs that cannot be reliably genotyped; (ii) MHs requiring a manual correction or IGV checks of the obtained genotypes; and (iii) reliable MHs with genotypes that can be used directly. Proportions and details of these three categories are represented in Fig. 6.

First, MHs 3pC, 5qD, 10qC, 12qA and 19qB showed misalignments or high misincorporation rates that significantly impeded their genotyping and were removed from the panel - representing a 4.2 % (5/118) assay failure rate.

Second, MHs with the lowest average coverage per marker were 6pB, XpB, 17qC, 16pB, 7pC and 14qA; representing components of the panel that have a risk of locus or allele drop-out. MHs 1qC and 19qA consistently showed allele imbalance with Ion S5 and add further risks of allele drop-out or drop-in with this platform. Note that the misincorporations seen in 15qB, 1pC, XqA, 13qD and 2pC can be discounted by applying a minimum allele read frequency threshold in these MHs. Applying such thresholds impacts their use for mixture detection, where care would be needed in interpreting any imbalances in haplotype ratios they may show. This second category of 13 MHs represent 11 % of the panel comprising loci whose genotypes need review, with particular care necessary when setting minimum allele read frequency thresholds in mixture analysis. Therefore, one hundred component MHs of the panel require no check and their haplotype data can be used directly in all forensic identification applications.

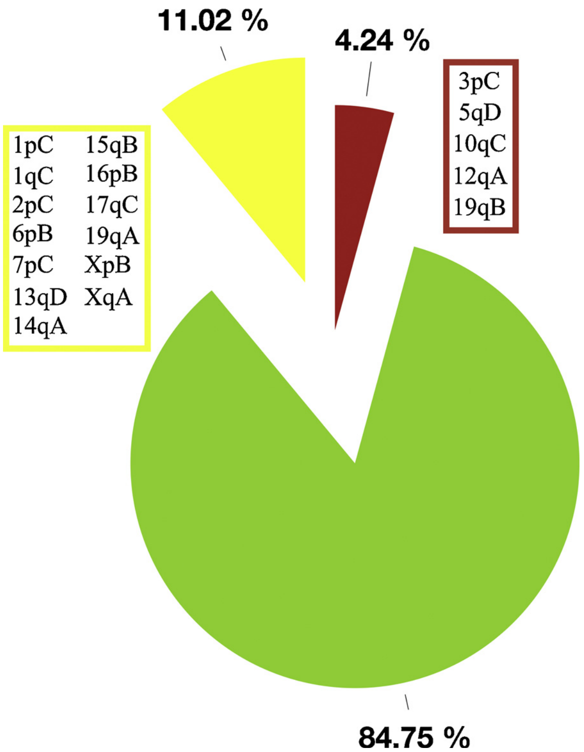

Fig. 6. Pie chart representing percentages of MHs in the panel for the following categories: excluded (red), need a manual review of the genotypes (yellow) and can be reliably genotyped (green) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

3.3. Forensic sensitivity assessments

There are two main differences between MiSeq and Ion S5 protocols that can affect forensic sensitivity to low-level and degraded DNA. First, MiSeq libraries are amplified in 10 cycles (increased from the 7 cycles of the manufacturer's protocol). Second, Ion S5 libraries are quantified with a qPCR assay that does not differentiate between primer dimers and actual library fragments; potentially overestimating quantification values (MiSeq libraries were quantified at the expected size range, in a more conservative approach). These differences affect both the library pooling process and the relative coverage of the samples in the chip/flow cell. We note that protocols can be further optimized for both platforms to enhance forensic sensitivity by increasing capture PCR cycles or performing a library amplification step with Ion S5, so the following evaluations do not represent the final capabilities of each platform.

3.3.1. Low-level DNA

Taking into account the guidelines from other commercially available Ampliseq™-based forensic assays containing similar marker numbers, optimal DNA input was set at 1 ng. Supplementary File S3 shows STR-like profiles obtained for both platforms through the in-house MH-allele calling pipeline for 1 ng input of 2800 M. Typically, from the 113 MHs validated as reliably genotyped in Section 3.2, one to three MHs from loci described in Section 3.2.1.1 drop-out, as they do not reach the 15 reads per allele coverage threshold. This is not unexpected as, in contrast to commercially available forensic panels, the initial capture PCR has yet to be fully optimized for our panel.

Fig. 7A shows the distribution of coverage per marker obtained from

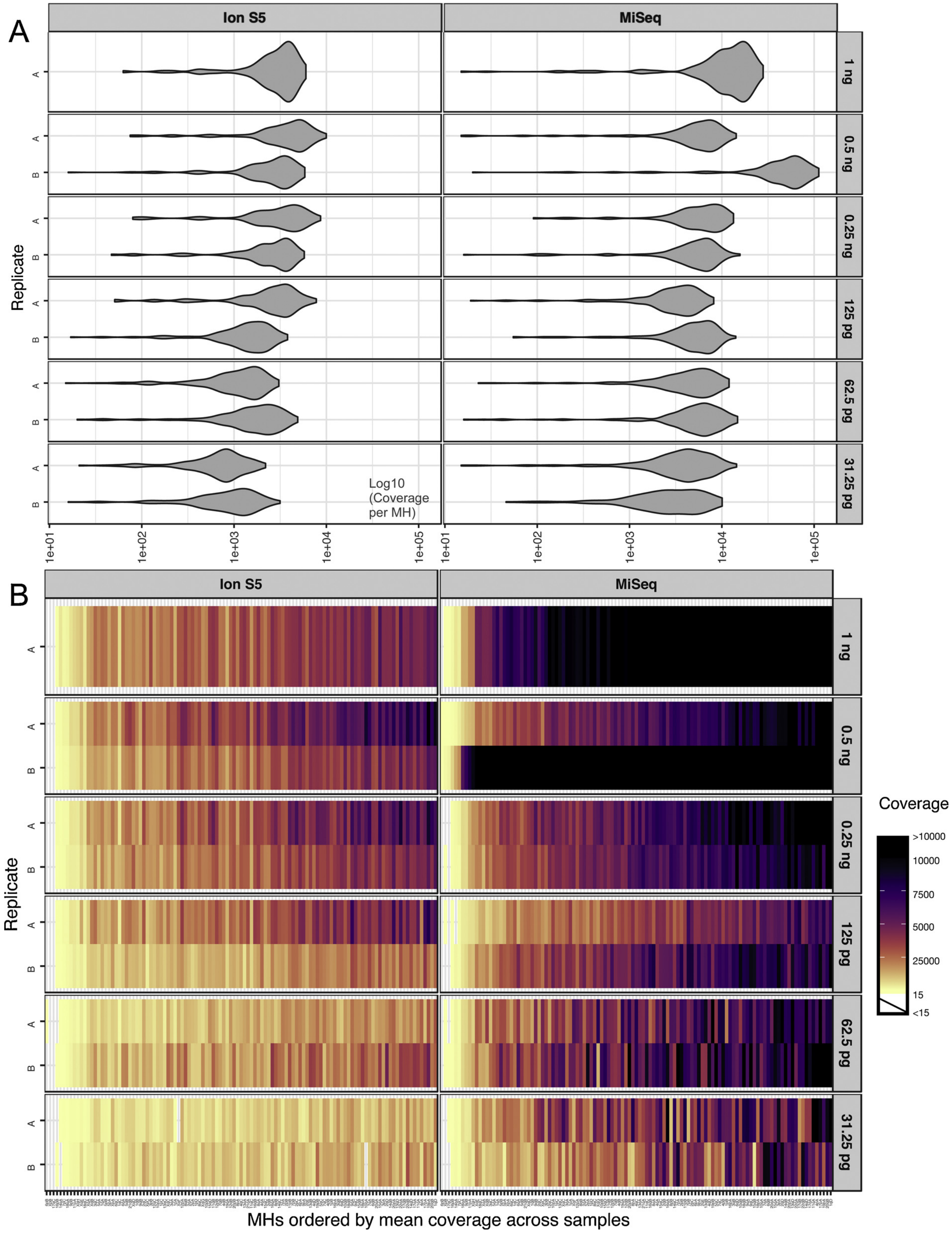

Fig. 7. A. Violin plot of coverage per MH from the sensitivity samples with Ion S5 (right) and MiSeq (left). 7B. Heat map of coverage per MH from the sensitivity samples on Ion S5 (right) and MiSeq (left) platforms. MHs are ordered by mean coverage across samples.

1 ng, 0.5 ng, 0.25 ng, 125 pg, 62.5 pg and 31.25 pg inputs (A–B replicates, both platforms). These distributions share a common pattern with a large number of MHs at highest levels of coverage followed by a narrow range of loci with the lowest values. When reducing input DNA, a trend is seen where the wider range of coverage distribution spreads and moves towards lower values, reaching sequence coverage mean values in both platforms for the 31.25 pg replicates about three times lower than 1 ng.

Section 3.2.1.1 highlighted that coverage is not evenly distributed across component MH markers. Fig. 7B reveals that, even for the lowest

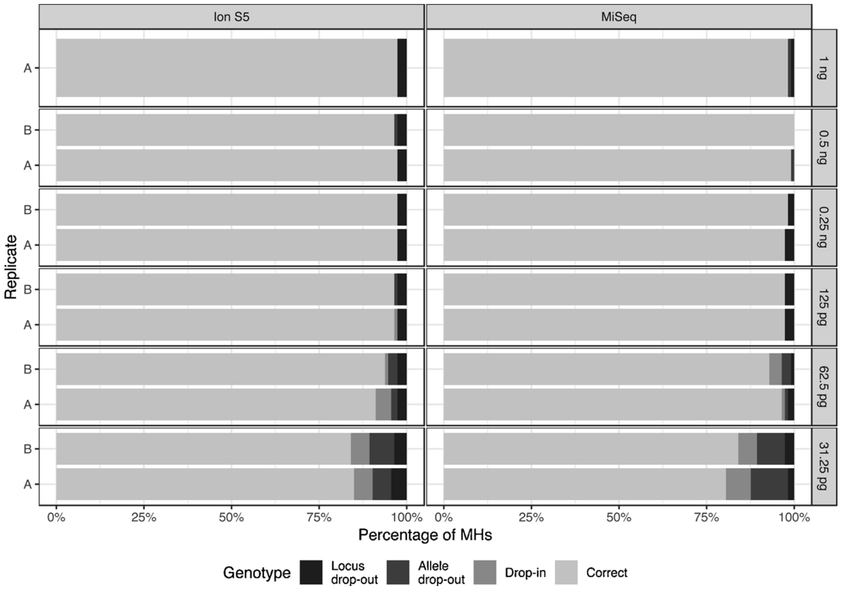

Fig. 8. Barplot of percentage locus drop-outs, allele drop-outs, drop-ins and correct genotypes obtained for the sensitivity samples in Ion S5 (right) and MiSeq (left).

input amounts, most MHs obtained sufficient sequence coverage (considering a threshold of 15 reads) and locus drop-outs were mainly restricted to those MHs previously described.

The obtained MH genotypes were evaluated for concordance and results are shown in Fig. 8. As expected, stochastic events such as allele drop-outs and drop-ins occur at a higher frequency at the lowest input levels. Over 95 % concordant genotypes could be obtained for both platforms with input amounts of 125 pg, going down to around 90 % for 62.5 pg and 80 % for 31.25 pg. Therefore, a manual review of genotypes for single-source sample inputs under 125 pg is recommended in order to exclude MHs with low frequency alleles or high heterozygous imbalance. The revision of low-level DNA genotypes from mixtures could prove to be particularly challenging.

3.3.2. Degraded DNA

Figs. 9A and B shows the sequence coverage obtained from artificially degraded DNA. The trend and distribution of coverage values per marker are similar to those of the sensitivity dilution series, which directly relates to the reduction of intact DNA as sonication times increase.

Fig. 10 compares MH genotyping results from both platforms and profile completeness from standard STR profiling. As a consequence of the deliberate design of amplicon sizes of 125–175 nt, the panel's amplification success exceeded that of STRs. Indeed, MH profiles with ~95 % correct genotypes were obtained in the 240-minute sonicated sample (both platforms) that had no detectable genotypes for STRs.

3.4. Mixture analysis

Analyzing mixtures with Ion S5 at volumes 1:1, 1:3, 1:7, 1:15 of 2800 M (first contributor) and 9947A (second contributor) underlined the strong performance of microhaplotypes to detect mixed DNA and identify contributors, as has been widely reported [1–3,6,7]. Fig. 11 shows concordance for the recognized haplotypes with the expected data from the single-source DNA used, applying two min_allele_frequency thresholds of 0.1 and 0.02. For 0.1 min_allele_frequency thresholds, the capacity to detect minor alleles decreases at extreme mixture ratios, resulting in ~45 % allele drop-out in the 1:7 mixture and ~60 % at 1:15. This is explained by the fact that the expected allele read frequency of a heterozygous allele from the minor contributor drops to 6.25 % for the 1:7 mixture and 3.125 % at 1:15.

As proposed in other studies [29,30,32], applying a lower min_allele_frequency threshold can improve minor mixture contributor detection. A threshold of 0.02 resulted in allele drop-out rates lower than 5 % even for the most skewed ratios. However, a second effect of lowering the min_allele_frequency threshold is that drop-ins occur at an increased rate of ~5 %. For this reason, any prior knowledge of the mixture ratio (for example, as estimated from previous STR profiles) can be used to determine the expected frequency in the mixture of the minor contributor alleles and set the min_allele_frequency threshold accordingly.

Fig. 12 and Supplementary File S4 show profiles for the single-source samples and different mixture ratios. Taking into account coverage, alleles can be assigned to a major or minor contributor in imbalanced mixtures, in a similar way to common STR profiles. However, a more comprehensive study, including a greater range of mixture ratios and contributors, is needed in order to fully assess the potential of this MH panel for mixture analysis.

3.5. Cross-checks with previously published forensic microhaplotype panels

The 118 MHs compiled in the panel were compared with seven published sets of forensic MH loci [3,6–9,33,34] to ensure no overlapping sites had been identified in parallel. The full list of all 118 + 211 other microhaplotypes selected for forensic purposes, is given in Supplementary Table S5, ordered by chromosome and position. The 87 microhaplotypes recently developed as a forensic MPS assay by Turchi et al. [35] are based on the 130 loci identified by Kiddlab [3].

One pair of overlapping sites was identified: 1pB (rs59090359-rs12273809) and mh11PK-63643 (rs1291417249-rs1188483930); an MH developed by van der Gaag et al. and published during our studies [8].

4. Concluding remarks

In this study, the necessary assessments of marker informativeness made before designing a de-novo MPS multiplex, proved a critical step in confirming that the selected MH loci would be much more powerful for forensic identification and mixed DNA analysis than individual SNPs in similar sized multiplexes. In particular, the simulations of complex kinship analyses indicated the power that can be achieved with microhaplotypes efficiently genotyped by MPS from short fragments. Only nine microhaplotypes had three haplotypes and so have comparable

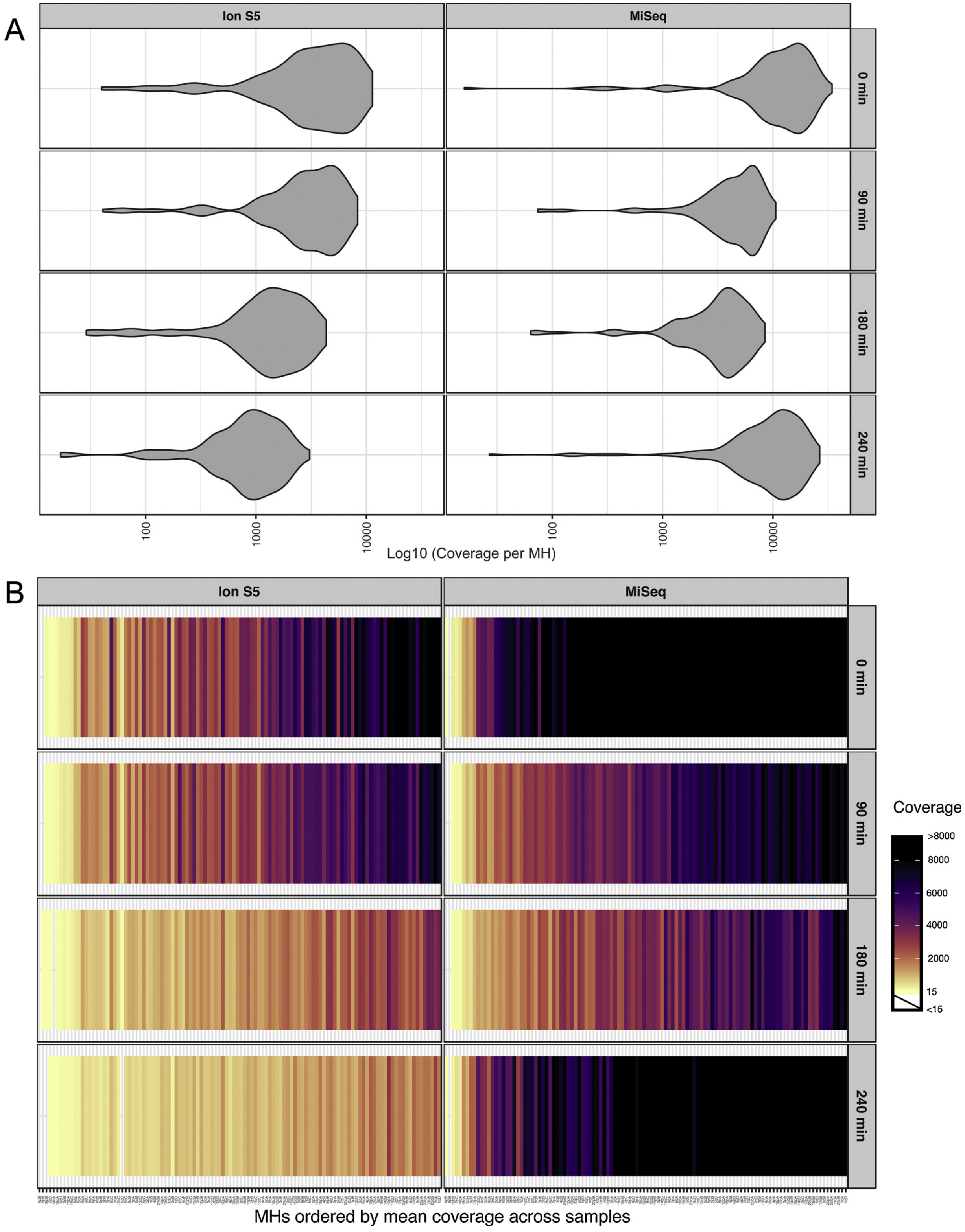

Fig. 9. A. Violin plot of coverage per MH from the degraded samples with Ion S5 (right) and MiSeq (left). 9B. Heat map plot of coverage per MH from the degraded samples with Ion S5 (right) and MiSeq (left) platforms. MHs are ordered by mean coverage across samples.

levels of informativeness to tri-allelic SNPs. The largest proportion of loci had 4 common haplotypes (with 5 or 6 haplotypes observed overall), and three had more than 8 haplotypes - culminating in 3qC with a total of 17 different haplotypes in 1KG Africans; 17 in Europeans; and 18 in East Asians. Despite the enhanced differentiation power obtained from combining over 100 microhaplotypes, simulations of pairwise kinship analyses of many first cousins and most second cousins (in deficient pedigrees) show it is not possible to obtain sufficiently high relationship probabilities with enough regularity to use this microhaplotype panel alone in such tests. Luckily, the single overlap of 1pB with mh11PK-63643 of van der Gaag's panel [8], means almost 200 other microhaplotypes already exist from previous publications and

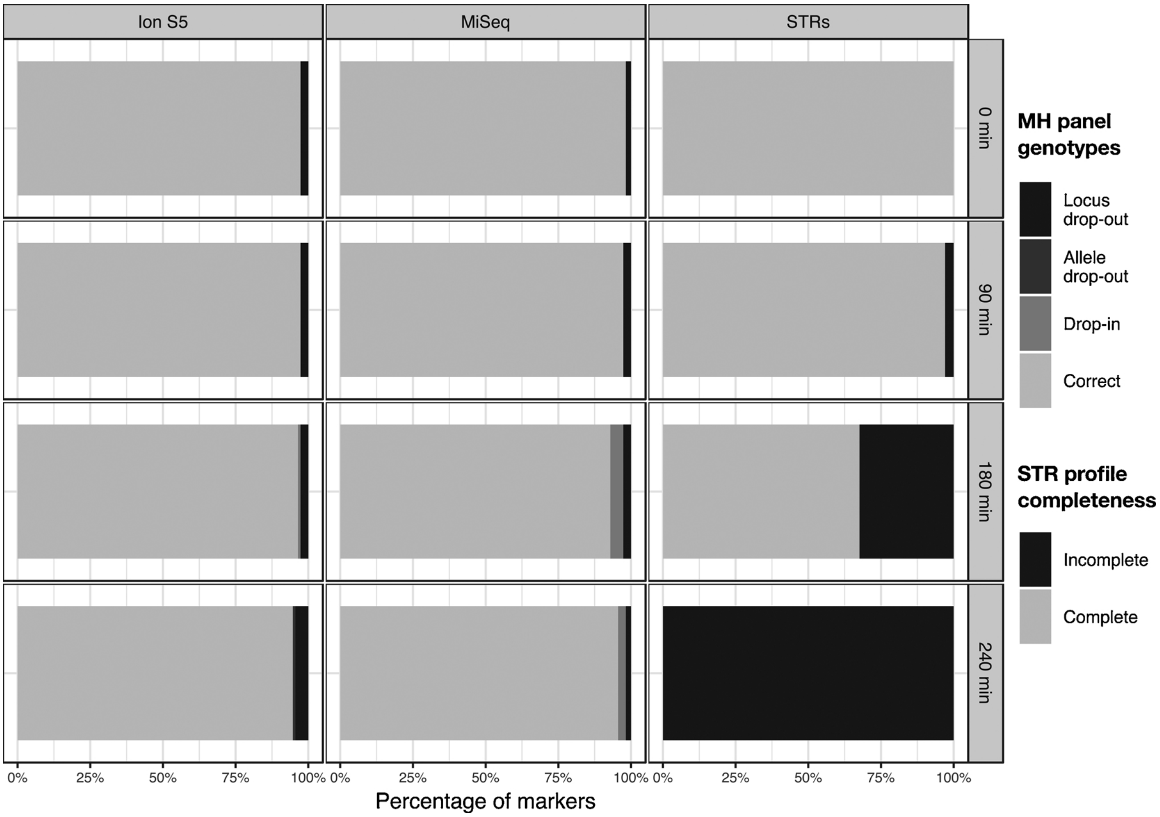

Fig. 10. Bar plot of percentage locus drop-outs, allele drop-outs, drop-ins and correct genotypes obtained for the degraded samples with Ion S5 (right) and MiSeq (middle). The leftmost plot shows percentage profile completeness from the standard STR set for the same samples.

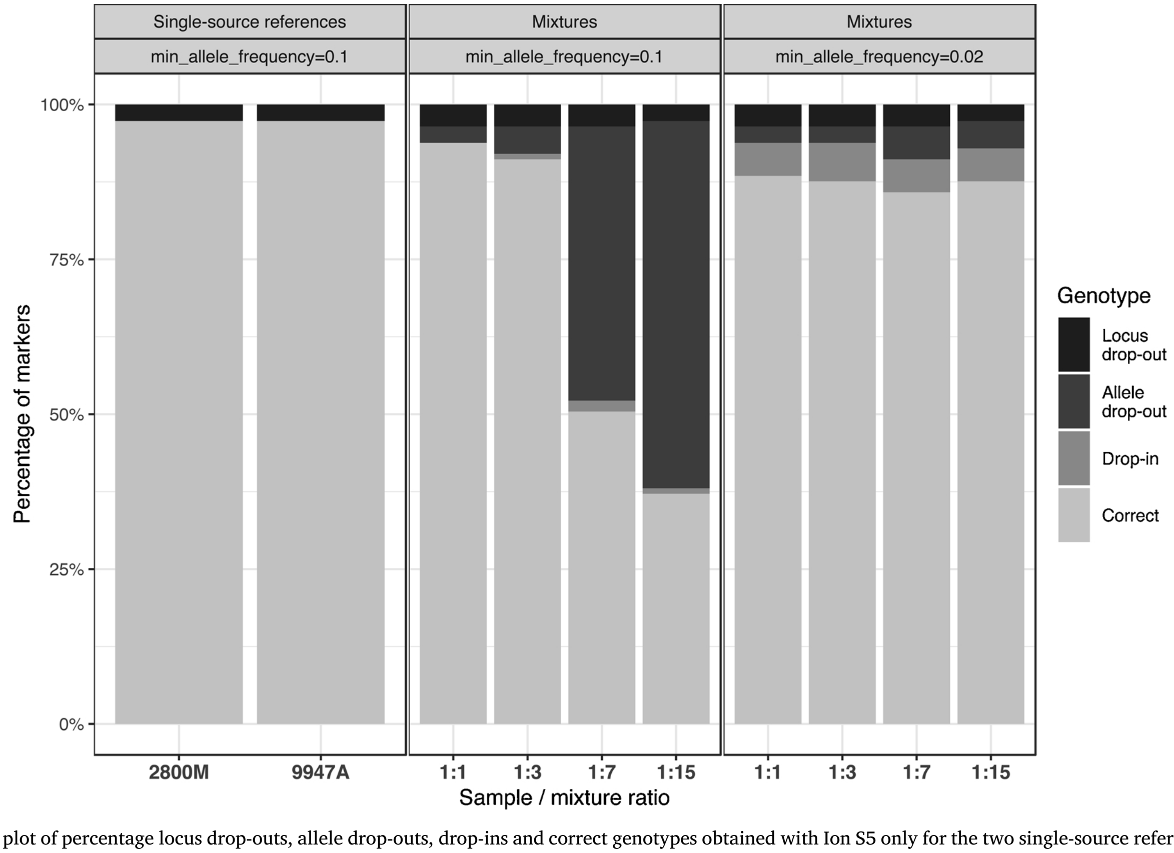

Fig. 11. Bar plot of percentage locus drop-outs, allele drop-outs, drop-ins and correct genotypes obtained with Ion S5 only for the two single-source reference samples (left) and for different mixture ratios using a min_allele_frequency threshold of 0.1 (middle) and a 0.02 (right).

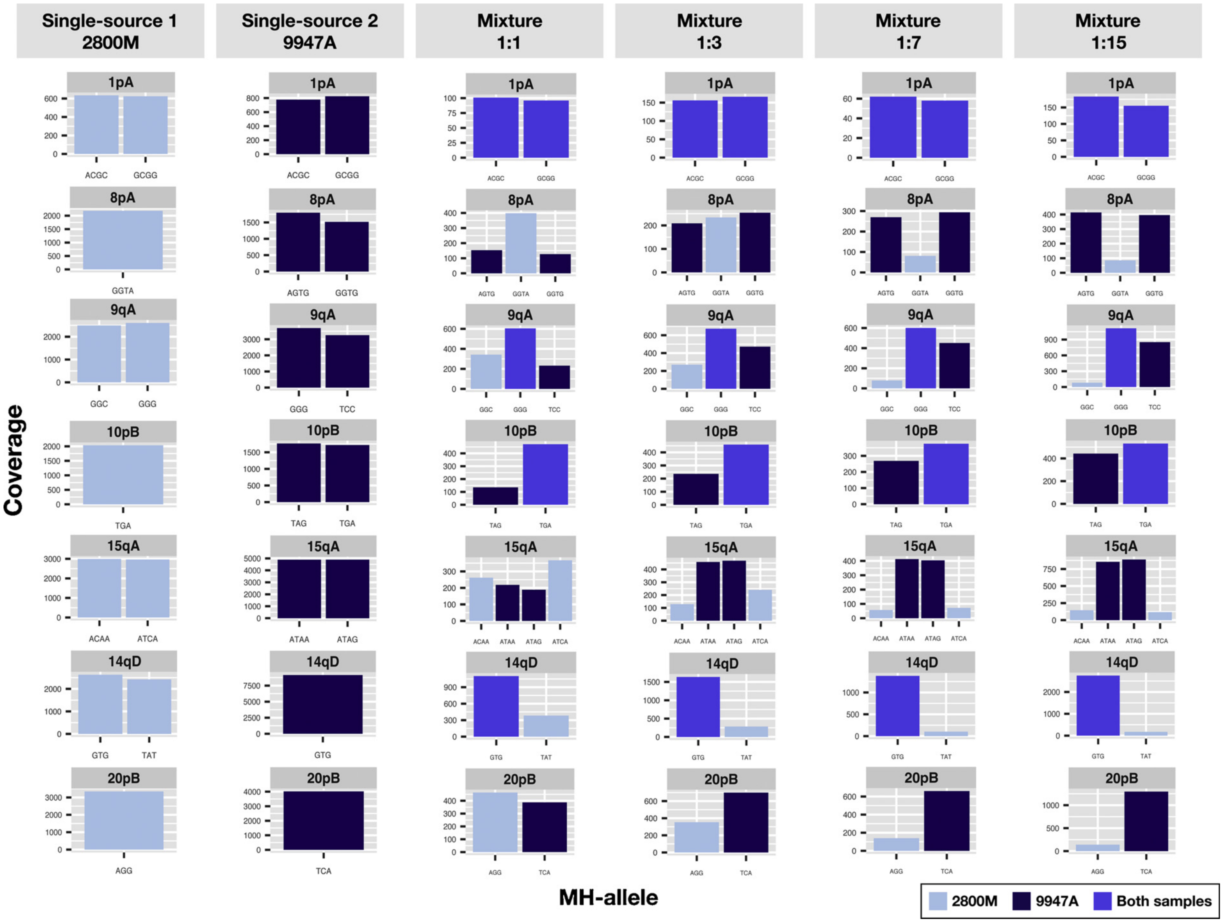

Fig. 12. Summary of profiles obtained from 1 ng input of single-source samples 2800 M and 9947A and volume mixture ratios 1:1, 1:3, 1:7 and 1:15 with Ion S5 platform. Bar colors indicate the source of each allele in the mixture according to the legend. Complete profiles are included in Supplementary File S4.

could be added to further extend microhaplotype MPS tests applied to challenging kinship analyses. If these are applied to the identification of missing persons, suitable adjustment of haplotype spans in many of these extra loci will be necessary, but a large proportion of them will retain sufficient informativeness to far exceed the levels achievable with multiple-allele SNPs sequenced from similarly sized amplicons.

From evaluations of sequencing performance, MHs 3pC, 5qD, 10qC, 12qA and 19qB were removed from the set of 118, leading to a final core panel of 113 loci. A further 13 MHs require care with interpretation of sequence coverage data - which will need manual review of their genotypes when analyzing low-level DNA. However, this is a relatively straightforward task since analysis of the challenging DNA typical of missing persons identification and for which this panel is well suited, is not a high-throughput forensic application. The same can be said of mixed DNA analysis, with the remaining 100 MHs of the panel providing comprehensive and detailed relative sequence coverage data, as well as haplotype diversity levels which will be informative enough for interpretation of many of the forensic mixtures commonly encountered in routine criminal casework [36]. We believe the low coverage of MHs 6pB, 7pC, 14qA, 16pB, 17qC, and XpB can largely be addressed by upward adjustments of individual primer ratios, when initial primer pools are constructed - a process of rebalancing over which we had no control. Although minimum allele read frequency thresholds of 10 % can be applied to Ion S5 data for MHs 1pC, 2pC, 13qD, 15qB and XqA, this impacts their effectiveness for mixture analysis with this platform, particularly if mixture ratios between minor and major DNA contributors are more extreme.

Finally, the care taken with marker selection with regard to designing short amplicon PCR primers to analyze highly degraded DNA, and the data we obtained from MPS analysis of dilution series and degraded control DNAs, suggest similar or superior levels of sensitivity for this panel to those from established forensic marker sets, including Mini-STRs and SNPs. This indicates that the microhaplotypes assembled will be ideal markers for the identification of missing persons. In this application a full range of markers can be applied, and their data combined to reach the highest possible likelihoods for the relationship hypotheses that need to be compared. The use of over 100 novel microhaplotypes sequenced with a single MPS assay, will also enhance mixed profile analysis, and the mixture interpretation pipeline we developed will take this complex and challenging application of forensic MPS further forward.

Acknowledgements

The studies reported and authors MdlP, CP, MVL 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. 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 studies reported have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 740580 within the framework of the Visible Attributes Through Genomics (VISAGE) Project and Consortium.

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.2019.102213.

References

[1] 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.

[2] K.K. Kidd, W.C. Speed, Criteria for selecting microhaplotypes: mixture detection and deconvolution, Investig. Genet. 6 (2015) 1.

[3] 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.

[4] J. Costas, A. Salas, C. Phillips, Á. Carracedo, Human genome-wide screen of haplotype-like blocks of reduced diversity, Gene 349 (2005) 219–225.

[5] 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.

[6] P. Chen, C. Yin, Z. Li, Y. Pu, Y. Yu, P. Zhao, D. Chen, W. Liang, L. Zhang, F. Chen, Evaluation of the Microhaplotypes panel for DNA mixture analyses, Forensic Sci. Int. Genet. 35 (2018) 149–155.

[7] P. Chen, C. Deng, Z. Li, Y. Pu, J. Yang, Y. Yu, K. Li, D. Li, W. Liang, L. Zhang, F. Chen, A microhaplotypes panel for massively parallel sequencing analysis of DNA mixtures, Forensic Sci. Int. Genet. 40 (2019) 140–149.

[8] K.J. van der Gaag, R.H. de Leeuw, J.F.J. Laros, J.T. den Dunnen, P. de Knijff, Short hypervariable microhaplotypes: a novel set of very short high discriminating power loci without stutter artefacts, Forensic Sci. Int. Genet. 35 (2018) 169–175.

[9] L. Voskoboinik, U. Motro, A. Darvasi, Facilitating complex DNA mixture interpretation by sequencing highly polymorphic haplotypes, Forensic Sci. Int. Genet. 35 (2018) 136–140.

[10] The Genomes Project Consortium, A global reference for human genetic variation, Nature 526 (2015) 68–74.

[11] 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.

[12] M. Nei, F. Tajima, DNA polymorphism detectable by restriction endonucleases, Genetics 97 (1981) 145–163.

[13] F. Oldoni, K.K. Kidd, D. Podini, Microhaplotypes in forensic genetics, Forensic Sci. Int. Genet. 38 (2018) 54–69.

[14] 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.

[15] 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.

[16] 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.

[17] SNPCheck3. https://secure.ngrl.org.uk/SNPCheck/credits.htm.

[18] T. Egeland, P.F. Mostad, B. Mevag, M. Stenersen, Beyond traditional paternity and identification cases. Selecting the most probable pedigree, Forensic Sci. Int. 110 (2000) 47–59.

[19] D. Kling, A.O. Tillmar, T. Egeland, Familias 3 - Extensions and new functionality, Forensic Sci. Int. Genet. 13 (2014) 121–127.

[20] R: a Language and Environment for Statistical Computing, (2019) http://www.r-project.org/.

[21] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, et al., The Simons Genome Diversity Project: 300 genomes from 142 diverse populations, Nature 538 (2016) 201–206.

[22] Bam-readcount software at: https://github.com/genome/bam-readcount.

[23] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760.

[24] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[25] N. Thomas, R Package - Microhaplot, (2019) https://github.com/ngthomas/microhaplot.

[26] J.T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, Nat. Biotech. 29 (2011) 24–26.

[27] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, et al., Accurate whole human genome sequencing using reversible terminator chemistry, Nature 456 (2008) 53–59.

[28] J.M. Rothberg, W. Hinz, T.M. Rearick, J. Schultz, W. Mileski, M. Davey, J.H. Leamon, K. Johnson, M.J. Milgrew, M. Edwards, et al., An integrated semiconductor device enabling non-optical genome sequencing, Nature 475 (2011) 348–352.

[29] M. Eduardoff, C. Santos, M. de la Puente, T.E. Gross, M. Fondevila, C. Strobl, B. Sobrino, D. Ballard, P.M. Schneider, Á. Carracedo, et al., Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™, Forensic Sci. Int. Genet. 17 (2015) 110–121.

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

[31] L.M. Bragg, G. Stone, M.K. Butler, P. Hugenholtz, G.W. Tyson, Shining a light on dark sequencing: characterising errors with Ion Torrent PGM data, PLoS Comput. Biol. 9 (2013) e1003031.

[32] 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.

[33] 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.

[34] P. Chen, W. Zhu, F. Tong, Y. Pu, Y. Yu, S. Huang, Z. Li, L. Zhang, W. Liang, F. Chen, Identifying novel microhaplotypes for ancestry inference, Int. J. Legal Med. 133 (2019) 983–988.

[35] C. Turchi, F. Melchionda, M. Pesaresi, A. Tagliabracci, Evaluation of a microhaplotypes panel for forensic genetics using massive parallel sequencing technology, Forensic Sci. Int. Genet. 41 (2019) 120–127.

[36] F. Oldoni, D. Podini, Forensic molecular biomarkers for mixture analysis, Forensic Sci. Int. Genet. 41 (2019) 107–119.