Forensic Science International: Genetics 35 (2018) 169–175

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Short hypervariable microhaplotypes: A novel set of very short high discriminating power loci without stutter artefacts

Kristiaan J. van der Gaag[a,b], Rick H. de Leeuw[a], Jeroen F.J. Laros[a], Johan T. den Dunnen[a,c], Peter de Knijff[a,*]

[a] Department of Human Genetics, Leiden University Medical Center, Einthovenweg 20, 2333, ZC, Leiden, The Netherlands
[b] Division of Biological Traces, Netherlands Forensic Institute, Laan van Ypenburg 6, 2497GB, The Hague, The Netherlands
[c] Department of Clinical Genetics, Leiden University Medical Center, Einthovenweg 20, 2333, ZC, Leiden, The Netherlands

ARTICLE INFO

Keywords:
Microhaplotypes
Massively parallel sequencing
MPS
Next generation sequencing
NGS
Miseq
Ion pgm
Forensic
Dna
Gonl
Degraded
Fdstools

ABSTRACT

Since two decades, short tandem repeats (STRs) are the preferred markers for human identification, routinely analysed by fragment length analysis. Here we present a novel set of short hypervariable autosomal microhaplotypes (MH) that have four or more SNPs in a span of less than 70 nucleotides (nt). These MHs display a discriminating power approaching that of STRs and provide a powerful alternative for the analysis;1;is of forensic samples that are problematic when the STR fragment size range exceeds the integrity range of severely degraded DNA or when multiple donors contribute to an evidentiary stain and STR stutter artefacts complicate profile interpretation. MH typing was developed using the power of massively parallel sequencing (MPS) enabling new powerful, fast and efficient SNP-based approaches. MH candidates were obtained from queries in data of the 1000 Genomes, and Genome of the Netherlands (GoNL) projects. Wet-lab analysis of 276 globally dispersed samples and 97 samples of nine large CEPH families assisted locus selection and corroboration of informative value. We infer that MHs represent an alternative marker type with good discriminating power per locus (allowing the use of a limited number of loci), small amplicon sizes and absence of stutter artefacts that can be especially helpful when unbalanced mixed samples are submitted for human identification.

1. Introduction

Short Tandem Repeats (STRs) have been the preferred marker for human identification for over two decades. Although the high degree of variation at STR-loci [1] provides useful discriminatory power for forensic and paternity cases, STRs are not the ideal marker type when degraded or mixed samples are involved. The interpretation of samples that have multiple contributors (and especially those with unequal contributions) can be complicated by the effects of slippage of DNA polymerases at the repeat stretches, resulting in stutter peaks that reside foremost at the n-1 position (representing products of one repeat unit less than the original allele length) [2]. Also, STR fragments with higher repeat numbers can be too long to allow amplification in severely degraded DNA samples [3]. The ideal forensic marker has a high degree of variation per fragment, allows for the design of small amplicons and is devoid of the production of stutter artefacts. In 1999, Jin et al. [4] published such a marker: a hypervariable fragment close to the MX1 gene on chromosome 21 containing several single nucleotide polymorphisms (SNPs) within a stretch of 100 nucleotides that proved to be informative in population genetics. However, at that time, the full power of such loci could not be exploited since routine analysis performed by Sanger sequencing only provides consensus information for each position without revealing how the variants of different SNPs within a fragment are connected (as a microhaplotype).

The development of massive parallel sequencing (MPS) platforms has provided promising new possibilities, especially for marker types that reveal their discriminatory value upon sequencing analysis. For STRs, MPS reveals substantial sequence variation in addition to repeat length, thereby increasing the discriminatory power of STRs compared to conventional fragment analysis [5,6]. However, even with MPS, the complication of stutter formation in the interpretation of complex mixtures remains. MPS also allows for the analysis of large panels of SNPs when severely degraded DNA is involved [7,8]. Recently, microhaplotypes (MH) or fragments with two to four SNPs, within a 200 nucleotide (nt) stretch, have been described [9] as an alternative for STR typing of mixtures. Note that both SNPs and MHs do not allow for

* Corresponding author.
E-mail addresses: k.van.der.gaag@nfi.minvenj.nl (K.J. van der Gaag), r.h.de_leeuw@lumc.nl (R.H. de Leeuw), j.f.j.laros@lumc.nl (J.F.J. Laros), ddunnen@humgen.nl (J.T. den Dunnen), knijff@lumc.nl (P. de Knijff).

https://doi.org/10.1016/j.fsigen.2018.05.008
Received 28 February 2018; Received in revised form 3 May 2018; Accepted 16 May 2018
Available online 22 May 2018

searches in DNA databases that are generally built from STR data and that relevant reference samples need to be available. Here we examine a new set of short hypervariable haplotypes, consisting of four or more SNPs contained in genomic fragments of less than 70 nt. We indicate that these MHs represent a discriminating power close to that of STR loci and facilitate mixture analysis without the hindrance of stutter. The data of the 1000 genome [10] and the GoNL projects [11] were used to identify potentially useful MHs. To confirm the genetic variation of these loci, data from 276 individuals of three globally distinct populations and 97 DNA samples from nine large families were analysed using MPS. Variant data of the most promising MHs was made publically available via the Leiden Open (source) Variation Database (LOVD) [12,13].

2. Material and methods

2.1. Marker selection

We screened the Variant Call Format (VCF) files of the African samples of the 1000 Genome and all of the GoNL project samples (Dutch selected for European ancestry) for genome fragments spanning 100 nt containing six SNPs with Minor Allele Frequencies (MAFs) in the relevant population ≥0.1. To select a subset of fragments for wetlab confirmation from the total set (which was > 100,000 fragments), filtering was performed using the following criteria:

- At least four out of six SNPs need to occur in both the 1000 Genomes and the GoNL projects, we do this for both validation purposes, but also to confirm that the variants are present in samples from different ancestry;
- All six SNPs should be within 70 nt to maximise possibilities for small amplicon design (the number of fragments from the 100 nt interval search allowed us to further reduce the fragment size);
- At least five of the six SNPs should not share the same MAF to maximise the number of possible haplotypes (many identical frequencies suggests perfect linkage and lack of variation);
- One of the SNPs should have a MAF of at least 0.4 to avoid overrepresentation of one haplotype.
- The highest and the lowest MAF of the SNPs within a fragment should have a difference of at least 0.2 to maximise variation in frequencies between haplotypes.
- The genomic distance to the nearest fragment should be at least 100,000 nt.

For the remaining fragments, the MH sequence spanning all SNPs plus 60 nucleotides up- and downstream was checked for homology in the genome using BLAST. Fragments with multiple hits (both within one chromosome and on different chromosomes) were discarded. For the remaining fragments, primer design was performed using primer3 v4.0.0 allowing a Tm of 57–63 °C, a primer length of 18–27 nt and amplicon sizes of 80–120 bp. Fragments containing repeating elements (repeated four or more times) or single nucleotide stretches over 8 nt were discarded. After primer design, the complete amplicon was checked again for homology using BLAST to achieve the final set of fragments for wet-lab testing. The set was completed by designing an amplicon representing the most variable part of the fragment described in Jin et al. [4] which includes seven of the nine SNPs excluding the last SNP of the 248 bp fragment and the SNP in the additional 227 bp fragment.

2.2. Microhaplotype selection by monoplex PCR and ion PGM analysis

To confirm the sequence variation for the selected candidates, 92 MHs were sequenced in 15 samples using the Ion PGM™ System according to the manufacturer's procedures. Five Dutch, three Bhutanese, two Ghanese, two Pygmy and three Amerindian samples from the HGDP CEPH-panel [14] were amplified in monoplex reactions. PCRs were performed using a 10 µl reaction containing PCR buffer (Life Technologies), 3 mM $MgCl_2$, 0.2 µM dNTPs, primer concentrations of 0.1–0.8 µM, 0.6 units Amplitaq Gold (Life Technologies) and 1.5 ng DNA. PCR specificity was checked using the Qiaxcel system (Qiagen) according to the manufacturer's procedures and MHs for which additional bands were visible in eight or more samples were discarded. All monoplex PCR products of the same sample were pooled and adapters were ligated to the amplicon pool using the IonXpress library preparation kit according to the manufacturer's procedures (Ion Torrent/Thermo Fisher). Sequencing was performed using the PGM™ System according to the manufacturer's procedures (Thermo Fisher) and data analysis was performed using FDSTools [5].

2.3. MH confirmation by multiplex PCR and MiSeq analysis

A multiplex PCR was designed (amplicon sizes 87–126 bp including primers) to examine the most informative 16 MHs in more detail. To test for global variation, 99 samples from the Netherlands [15], 87 Asian samples of the Han Chinese and Japan HapMap panel [16], and 90 African samples of the Luhya (Kenya), Yoruba (Kenya/Nigeria) and Maasai (Kenya) HapMap panel were analysed. To confirm stable transmission of the variants, nine CEPH families (family 12, 66, 1328, 1347, 13281, 13291, 13292, 13293 and 13294; 97 samples in total) were analysed. Multiplex PCR was performed using a total volume of 12.5 µl containing PCR buffer (Life Technologies), 4 mM $MgCl_2$, 0.4 µM dNTP, primer concentrations of 0.03–0.35 µM, 2.5 units Amplitaq Gold (Life Technologies) and 1.5 ng DNA. Adapters were ligated using the KAPA HTP Library Preparation Kit for Illumina® platforms according to the manufacturer's procedures (KAPA Biosystems/Roche) and sequencing was performed using the MiSeq® Sequencer according to the manufacturer's procedures (Illumina, v3 chemistry). Data analysis was performed using FDSTools [5]. Data for all observed sequence variants of the final set of MHs was submitted to LOVD (http://databases.lovd.nl/DNA_profiles/) [12,13].

2.4. Statistical analysis

All statistic calculations were performed on haplotype data (not separately for each SNP). Population statistics were calculated for all populations (Chinese/Japanese and Kenyan/Nigerian were respectively grouped together) using Powerstats [17] and Genalex [18]. The power to detect mixtures (chance to observe a third allele for at least one locus) was calculated as described by Phillips et al. [19] by adding an extra sheet to the Powerstats Excel sheet (file available upon request). An Excel sheet was used to check for correct transmission of variants in the CEPH families. Neighbour joining networks were drawn for the 16 MHs of the final multiplex using Network 5 and Network Publisher [20] using the homologous sequence of a Chimpanzee as outgroup. Recombination rates were retrieved for all MHs from the HapMap recombination maps [21] and the average number of meioses for recombination to occur within the fragment was calculated considering the fragment lengths. To test the potential of these fragments to inform about geographic ancestry, STRUCTURE [22] was run 100 times with a K-value of 2, 3 and 4. CLUMPAK [23] was used to combine and visualize the data of the repeated runs.

3. Results

3.1. MH candidate selection

A search in the VCF files for genomic intervals of 100 nt containing at least six SNPs with a MAF of at least 0.1 resulted in 14,890 potential MHs in the African samples of the 1000 Genomes project and 105,129 MH candidates in the GoNL dataset. An overview of the number of remaining fragments for each chromosome after applying several

Table 1
Numbers of remaining short hypervariable microhaplotypes after applying several filtering criteria for selection of potentially informative fragments.

			Chromosome									
Selection Criteria	Total	% of total*	1	2	3	4	5	6	7	8	9	10
At least four SNPs in 1000 Genome and GoNL selection	10464	10.0%	181	332	277	291	133	5858	338	173	183	185
Clusters <70bp	5612	5.3%	89	155	120	128	64	3435	175	73	91	105
At least five SNPs with different MAF within the fragment	4726	4.5%	80	139	89	102	57	2925	146	58	75	87
Max MAF in fragment at least 0.4	3910	3.7%	61	108	67	81	41	2402	118	57	69	77
Max within fragment freq-distance > 0.2	2882	2.7%	50	71	46	61	31	1863	84	42	47	51
Distance between two fragments > 100,000 nt	410	0.4%	26	16	21	28	13	32	23	20	16	20
Succesful PCR-design	92		4	2	6	4	1	6	5	5	4	5

	Chromosome											
Selection Criteria	11	12	13	14	15	16	17	18	19	20	21	22
At least four SNPs in 1000 Genome and GoNL selection	185	223	210	410	103	227	244	115	255	282	115	144
Clusters <70bp	96	124	112	171	33	103	125	62	120	113	53	65
At least five SNPs with different MAF within the fragment	87	109	91	125	26	82	98	57	96	102	43	52
Max MAF in fragment at least 0.4	70	97	83	113	25	63	96	47	75	82	39	39
Max within fragment freq-distance > 0.2	48	57	53	66	17	50	63	33	41	53	29	26
Distance between two fragments > 100,000 nt	24	17	17	11	11	22	17	13	26	14	11	12
Succesful PCR-design	6	2	4	4	5	6	5	2	6	4	2	4

**percentages are calculated as proportion of the GoNL candidate fragments.*

filtering criteria is shown in Table 1. After checking the remaining 410 fragments for homologous regions in the genome and the possibility for PCR design, 92 fragments dispersed over the genome remained and amplicons were prepared for wet-lab testing.

3.2. MH candidate testing

From the 92 MH candidates amplified for the first set of 15 samples, 83 MHs passed the selection criterion regarding PCR specificity (no differently sized amplification products in at least eight of the 15 samples) and were subjected to sequence analysis.

The sequence variation observed for these 83 candidates was generally very low: in most cases only a single haplotype was observed. In addition, several fragments showed more than two alleles in the same sample reflecting multiple genomic copies of different sequence. Using these results as a selection criterion, 29 fragments remained with more than two haplotypes in 15 samples and no indication of fragment amplification from homologous loci based on the available data.

3.3. Performance of MH set

A multiplex PCR was designed and 23 of the 29 fragments were successfully amplified and sequenced in 276 population samples and 97 CEPH family samples. Three of the 23 fragments revealed multiple amplification products suggesting more than one genomic location. Four fragments showed insufficient sequence variation. Thus, 16 fragments remained for which the genome positions and primer sequences are displayed in Sup. Table 1a. Microhaplotpes were named according to the suggested names by Kidd et al. [24].

The observed number of variable SNP-positions within a MH varied from four to 22 and the number of unique haplotypes varied from 4 to 26 as displayed in Table 2.

A sequence alignment of the observed haplotypes of each MH is displayed in Sup. Fig. 1 and Sup. Table 2 displays the allele frequencies in each of the three tested populations.

Networks were drawn from the population samples for each MH to visualise the SNP-distance between the separate haplotypes and the observed number of haplotypes for each population. An example of the network of mh07PK-38311 is displayed in Fig. 1. For this figure, an illustration of the fragment was included connecting the position of the SNPs with the branches in the network. Sup. Fig. 2 displays the networks for each of the 16 MHs, statistics of the Chi-Square tests for Hardy-Weinberg Equilibrium are displayed in Sup. Table 3.

The observed degree of variation is different for each MH. 13 of the 16 loci result in a simple network with either no or one reticulation. The MH with the most haplotypes (mh17PK-86511) has a slightly more complex structure with a few low-frequency haplotypes that result in reticulations. MHs mh11PK-62906 and mh14PK-72639 result in complex web-like structure. mh11PK-62906 is the only fragment located in a region with a substantially elevated recombination rate. For the tested allele transfers of the selected 16 MHs in the CEPH families (144 allele transfer events for each locus in total), no inconsistent haplotype inheritance was observed. As an example, Sup. Fig. 3 displays the joined family tree for CEPH family 1328, 13281, 13291, 13292, 13293 and 13294 (50 individuals in total) with the corresponding genotypes and read counts of each haplotype for mh16PK-83544.

Forensic and paternity statistics are summarised for each tested population in Sup. Table 4. The random match probability (RMP) of the total set of 16 MHs is 9.2×10^{-13} for the African population, 4.4×10^{-11} for the Dutch population and 1.0×10^{-9} for the Asian population. In comparison, Table 3 displays the RMP for several panels of different kind of loci and Table 4 displays the power to detect a mixture (PMD) for the MHs and the tri-allelic [25] and tetra-allelic SNPs [19].

To test the power of the 16 MHs to differentiate populations of different ancestry, 100 Structure runs were performed using two to four groups (K = 2, K = 3 and K = 4, Fig. 2). A major cluster (76 of the 100

Table 2
Overview of the observed variation for each MH.

Locus	Unique observed haplotypes[a]	Number of observed SNP-positions in MH
mh06PK-24844	9	10
mh06PK-25713	6	6[b]
mh07PK-38311	4	5
mh08PK-46625	5	4
mh10PK-62104	5	7[b]
mh11PK-62906	19	7
mh11PK-63643	7	7
mh14PK-72639	15	9[b]
mh15PK-75170	12	13[b]
mh16PK-83362	7	8
mh16PK-83483	8	9
mh16PK-83544	5	6[b]
mh17PK-86511	26	22[b]
mh18PK-87558	4	6
mh22PK-104638	11	12
mh21PK-MX1s	5	4

[a] Each haplotype is defined as a unique observed combination of the SNP-variants within a fragment (in the tested human samples).

[b] For mh06PK-25713, mh10PK-62104, mh14PK-72639, mh15PK-75170, mh16PK-83544 and mh17PK-86511, one of the SNPs is tri-allelic. For each fragment, the number of observed unique haplotypes is displayed and the number of SNP-positions in the fragment from which these haplotypes are comprised.

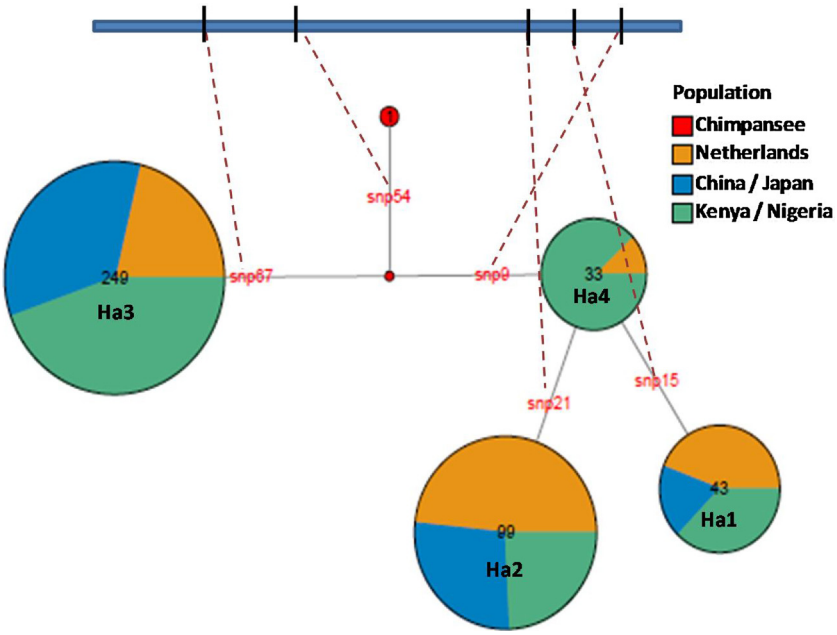

Fig. 1. Illustration of the fragment of mh07PK-38311 and the corresponding network of the haplotypes.
On top, the fragment of mh07PK-38311 is displayed with the observed SNP positions indicated by vertical lines. Below, the network displays the distribution of each haplotype over the different tested populations and the SNP-distance between each haplotype. The circles are sized by the number of haplotypes observed in each population with colours representing the haplotypes of each analysed population. Each branch of the network is connected to the corresponding SNP in the fragment by a dotted line.

runs) and a minor cluster (12 of the 100) was obtained for K = 2 separating the African or the Asian samples respectively from the other two populations. For K = 3, 98 of the 100 runs resulted in an almost complete separation of all three populations involved. For K = 4, a major and a minor fourth cluster were obtained resulting in a poor differentiation of either the African or the Dutch population (data not shown).

4. Discussion

Detection of degraded DNA and of minor contributions in mixed samples is often complicated when conventional forensic STR typing is applied. Due to the large range of amplicon sizes for some loci and the occurrence of stutter products, it can be difficult to generate reliable and reproducible STR profiles. It would therefore be ideal to use a marker type of small amplicon sizes with a discriminating power equivalent to STRs but without the burden of stutter artefacts.

We selected hypervariable micro haplotype loci with at least six

Table 3
Overview of the Random Match Probability for different panels of forensic loci.

Panel	number of loci	Type of loci	RMP[b]	Based on population	Source
Short hypervariable microhaplotypes	16	Micro haplotypes	4.4×10^{-11} 1.0×10^{-9} 9.2×10^{-13}	NL China/Japan Kenya/Nigeria	This study
SGM Plus® Kit	10	STRs	7.9×10^{-14} 3.0×10^{-13}	African American US Caucasian	ThermoFisher
NGM™	15	STRs	1.6×10^{-19} 4.6×10^{-20} 2.2×10^{-19}	US Hispanic African American US Caucasian	ThermoFisher
NGM™ < = 200 bp*	9[a]	STRs	3.1×10^{-12} 8.8×10^{-13} 2.6×10^{-12}	US Hispanic African American US Caucasian	ThermoFisher
Powerplex Fusion		STRs	1.6×10^{-28} 2.4×10^{-27} 2.1×10^{-27} 1.4×10^{-25}	African American US Caucasian US Hispanic US Asians	Promega
SNPforID	52	SNPs	5.0×10^{-21} 1.1×10^{-19} 5.0×10^{-19}	European Somali Asian	Sanchez et al. [28]
tri-allelic SNPs	13	SNPs (tri-allelic)	3.2×10^{-6} 4.4×10^{-7}	Dutch Dutch Antilles	Westen et al. [25]
tetra-allelic SNPs[a]	24	SNPs (tetra-allelic)	1.5×10^{-12} 5.2×10^{-10} 2.0×10^{-15}	European East Asian African	Phillips et al. [19]
Kidd MicroHaps	31	Micro haplotypes	$1 \times 10^{-13} - 4 \times 10^{-21}$	Various global populations	Kidd et al. [9]

[a] The boundary of 200 nt is within the range of some loci, on average 9 loci determine the RMP.
[b] Abbreviations: RMP = Random Match Probability.

Table 4
Overview of the Power of Mixture Detection for different panels of forensic loci.

Panel	number of loci	Type of loci	PMD[a]	Based on population	Source
Short hypervariable micro-haplotypes	16	Micro haplotypes	0.9989 0.9947 0.9999	NL China/ Japan Kenya/ Nigeria	This study
tri-allelic SNPs	10	SNPs (tri-allelic)	0.7490 0.9471	Dutch Dutch Antilles	Westen et al. [29]
tetra-allelic SNPs*	24	SNPs (tetra-allelic)	0.9939 0.9260 0.9999	European East Asian African	Phillips et al. [19]

[a] abbreviations: PMD = Power of Mixture Detection.

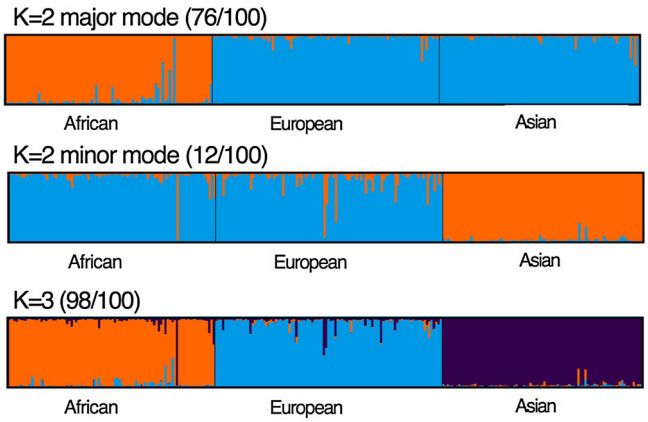

Fig. 2. STRUCTURE/CLUMPAK population differentiation for 16 MHs in the tested populations.
The figure displays the CLUMPAK results of 100 STRUCTURE runs, every bar displays one individual. On top, the major mode is displayed of STRUCTURE runs with K = 2 derived from 76/100 repeated analyses where the African samples are mostly differentiated from Europe and Asia. In the middle, the minor K = 2 mode is displayed derived from 12/100 repeated analyses where most Asian samples are differentiated from Africa and Europe. At the bottom, the results for K = 3 are displayed derived from 98/100 repeated analyses where most of the samples of the three continents are properly differentiated.

SNPs within a range of 100 nt from genomic reference data of a European and African populations and tested the final set on additional populations (including Asian samples) in order to provide a set of markers which is likely to be informative in the majority of global populations. Since the data available to us consisted merely of SNP-frequencies and did not contain any information about haplotype frequencies of the combined SNPs within a fragment, we used variation of SNP allele frequencies within each fragment as a means to maximise haplotype variation.

BLAST results of the first selection of 410 fragments exposed that many (≈25%) of the hypervariable fragments contained homologous regions in the genome, which suggested that part of the variation in the databases might have resulted from something else than actual SNP variation. After discarding those fragments, sequencing of the remaining fragments still revealed much less variation than we observed in the data of the two genome projects. Since most of such reference data is derived from alignment of short reads (for these projects reads of mostly ≤150 nt) to a reference sequence, there are two likely issues that could cause discrepancy in the estimated frequencies for these hypervariable fragments:

1 Homologous fragments may map to the same position, falsely suggesting a heterozygous genotype.
2 Fragments with many SNPs in a short range may exceed the number of allowed mismatches for mapping reads to the reference during analysis, meaning that only the reads that overlap part of the SNPs and haplotypes that are most similar to the reference sequence will be mapped to the correct location.

In combination with relatively low coverage, these two issues can result in erroneous variant calling for separate SNP positions within one (heterozygous) sample. An extensive wet-lab confirmation of new possibly hypervariable loci is therefore essential. Testing of samples from globally dispersed populations will not only give information about discriminating power in different populations, but also increase the chance to find different heterozygous allele combinations that can help to identify possible co-amplified homologous regions. Testing of samples from large families will confirm correct inheritance of the haplotypes and assist the internal validation of genotyping results.

Although many of the initial candidate loci were rejected, a final set of 16 MHs remained with expected inheritance of the haplotypes in the tested families and a high degree of variation in the population samples. With a varying number of haplotypes for each MH (2–19) and corresponding haplotype frequencies, the discriminating power is not as strong as STRs but the set of 16 loci still reaches strong random match probabilities (RMP) of: 1.0×10^{-9} in the Asian population, 4.4×10^{-11} in the Dutch population and 9.2×10^{-13} in the African population. For identification purposes, our set of loci prove to be more informative than other alternative non-STR loci as can be observed from Table 3. Since the populations tested for the different loci are not exactly the same, a direct comparison of the RMP should be interpreted carefully. Notwithstanding, the discriminating power of the 16 short hypervariable MHs roughly resembles that of nine STRs [26], 25 tri-allelic SNPs [25], 21 tetra-allelic SNPs [19] or 23 of the earlier described MHs [9].

An important advantage of the use of MHs for mixture analysis is the number of haplotypes that is observed for several loci. The statistical power (likelihood ratio) of matching a person with a two-person mixture is substantially increased when more than two alleles are present for a specific locus. The power to detect a third allele for a two-person mixture in at least one of the 16 loci ranges from 0.995 in the Asian population to 0.9989 in the Dutch population and even 0.99992 in African population. For detecting additional contributors in mixtures, the assay outperforms the published sets of tetra-allelic and tri-allelic SNPs (Table 4). For the 130 MHs of Kidd et al. [27], the average PMD is estimated based on the top 28 loci for different numbers of loci divided in ranges of effective number of alleles (3–4, 4–5 and > 5). These 28 loci together reach a PDM of 0.9999999875 from which 16 loci contain all SNPs within a 150 nt span. However, only three these 28 loci contain all SNPs within a 100 nt span as is the case for the loci described in this paper. The two sets together could complete an even more optimal set of loci for mixture detection.

4.1. Observed variation

Reticulations in a neighbour joining network can be caused by either recombination or by recurrent mutations. The only fragment located in a region with exceptionally high recombination rate is mh11PK-62906, but considering the small fragment length, recombination would only be expected to occur within the fragment once every 5.5×10^4 meioses. This might suggest that the web-like networks of mh11PK-62906 and mh14PK-72639 (and in lower extent mh17PK-86511) are more likely to be explained by mutation hotspots concentrated on a few specific positions rather than by recombination. Indeed, in none of the tested allele transfers of the CEPH families (144 allele transfer events for each locus in total), recombination has occurred in such a way that the allele inheritance of any of the loci was impacted. When using these loci for paternity cases, it should be

considered that mh11PK-62906 and mh14PK-72639 are more likely to display mutations than an average fragment.

For the network of mh06PK-25713, a fairly even distribution of the haplotype frequencies was observed for all populations but for most of the loci, several haplotypes vary substantially in frequency between the tested populations. This suggests that the MHs provide ancestry information although the design and selection of the loci was not intended for this purpose. STRUCTURE analysis indeed showed that the three analysed populations are differentiated almost completely based on the data of these 16 MHs. Data from a larger set of samples with a more global representation would be needed to test the full potential of these MHs as ancestry informative markers. From the 16 MHs that remained after all selection criteria, several of the loci failed Hardy-Weinberg equilibrium test since the frequency of some homozygous genotypes (usually with low frequency) is higher than would be expected.

None of the fragments is located in gene regions, so strong natural selection is not expected for these fragments. An explanation for this could be that some samples in the tested populations are somewhat genetically distinct from the rest of the population, which is not unlikely since we grouped samples of two Asian populations and of three African populations in order to achieve comparable sample sizes. It also cannot be excluded that some fragments could have occasional SNPs under the primer binding sites although we did not observe any discrepancy of inheritance in the nine CEPH families.

4.2. Sequence data analysis

It should be noted that not every software for sequence data analysis is capable to analyse single-fragment haplotype data. When using an analysis software that maps the complete sequences to a reference, results are often summarised by SNP instead of haplotypes. In this study we used FDSTools [5] since variant frequencies in the data are always reported for the complete sequence between two flanks instead of a summary for each position.

5. Conclusions

A new set of short hypervariable microhaplotypes were selected as potential loci for application in forensic DNA analysis. For 16 MFs, confirmation of the variation and inheritance was performed by analysing 276 samples of three globally dispersed populations and 97 samples of nine large families. MHs provide an alternative type of loci for cases where STR stutter or degradation of DNA limits or complicates the analysis. Since the discriminating power of the selected hypervariable MHs is larger than other published non-STR loci, they provide a practical and financially advantageous method with a relatively small number of loci. For the purpose of increased discriminating power and ancestry informative information, a combination of these loci with (part of) the loci of Kidd et al. [21] could provide an even more powerful tool.

The selection of short hypervariable MHs from genomic reference data is complicated since the generally short read length of reference data is not ideal to resolve the exact variation in short range hypervariable fragments. Since the read length of most MPS platforms is increasing, future reference data will most likely be better suited for selection and analysis of additional MHs.

National forensic DNA databases currently consist of STR data. Although it is not expected that all database samples will be typed for new loci in the near future, loci such as MHs could provide a powerful tool in cases where reference samples are available for comparison.

Competing interests

The authors declare no conflict of interest.

Acknowledgements

This study was supported by a grant from the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands. The authors wish to thank Titia Sijen and Jerry Hoogenboom (Netherlands Forensic Institute) for carefully reviewing the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2018.05.008.

References

[1] K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR allele sequence variation: current knowledge and future issues, Forensic Sci. Int. Gent. 18 (2015) 118–130.
[2] A.A. Westen, L.J. Grol, J. Harteveld, A.S. Matai, P. de Knijff, T. Sijen, Assessment of the stochastic threshold: back- and forward stutter filters and low template techniques for NGM, Forensic Sci. Int. Genet. 6 (2012) 708–715.
[3] Reza Alaeddini, Simon J. Walsh, Ali Abbas Forensic implications of genetic analyses from degraded DNA—a review, Forensic Sci. Int. Genet. 4 (2010) 148–157.
[4] L. Jin, P.A. Underhill, V. Doctor, R.W. Davis, P. Shen, L.L. Cavalli-Sforza, P.J. Oefner, Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations, Proc. Natl. Acad. Sci. U. S. A. 96 (March 30 (7)) (1999) 3796–3800.
[5] J. Hoogenboom, K.J. van der Gaag, R.H. de Leeuw, T. Sijen, P. de Knijff, J.F. Laros, FDSTools: a software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise, Forensic Sci. Int. Genet. 27 (November (27)) (2016) 27–40.
[6] K.J. van der Gaag, R.H. de Leeuw, J. Hoogenboom, J. Patel, D.R. Storts, J.F. Laros, P. de Knijff, Massively parallel sequencing of short tandem repeats-population data and mixture analysis results for the PowerSeq™ system, Forensic Sci. Int. Genet. 24 (September) (2016) 86–96.
[7] S. Elena, A. Alessandro, C. Ignazio, W. Sharon, R. Luigi, B. Andrea, Revealing the challenges of low template DNA analysis with the prototype ion ampliSeq™ Identity panel v2.3 on the PGM™ sequencer, Forensic Sci. Int. Genet. 22 (May) (2016) 25–36.
[8] I. Grandell, R. Samara, A.O. Tillmar, A SNP panel for identity and kinship testing using massive parallel sequencing, Int. J. Legal Med. 130 (July (4)) (2016) 905–914.
[9] 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 (September) (2014) 215–224.
[10] The 1000 Genomes Project Consortium, A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74.
[11] Genome of the Netherlands Consortium, Whole-genome sequence variation, population structure and demographic history of the Dutch population, Nat. Genet. 46 (2018) 818–825.
[12] I.F. Fokkema, J.T. den Dunnen, P.E. Taschner, LOVD: easy creation of a locus-specific sequence variation database using an LSDB-in-a-box approach, Hum. Mutat. 26 (August (2)) (2005) 63–68.
[13] I.F. Fokkema, P.E. Taschner, G.C. Schaafsma, J. Celli, J.F. Laros, J.T. den Dunnen, LOVD v.2.0: the next generation in gene variant databases, Hum. Mutat. 32 (May (5)) (2011) 557–563, http://dx.doi.org/10.1002/humu.21438 (Epub 2011 Feb 22).
[14] H.M. Cann, C. de Toma, L. Cazes, M.F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G.B. Ferrara, J.S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R.J. Herrera, X. Huang, J. Kidd, et al., A human genome diversity cell line panel, Science 296 (5566 April (12)) (2002) 261–262.
[15] A.A. Westen, T. Kraaijenbrink, E.A. Robles de Medina, J. Harteveld, P. Willemse, S.B. Zuniga, K.J. van der Gaag, N.E. Weiler, J. Warnaar, M. Kayser, T. Sijen, P. de Knijff, Comparing six commercial autosomal STR kits in a large Dutch population sample, Forensic Sci. Int. Genet. 10 (2014) 55–63.
[16] The International HapMap Consortium, Integrating ethics and science in the International HapMap Project, Nat. Rev. Genet. 5 (June (6)) (2004) 467–475.
[17] Promega Corporation, Powerstats v1.2. Unpublished work.
[18] R. Peakall, P.E. Smouse, GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update, Bioinformatics 1 (October 28(19)) (2012) 2537–2539.
[19] 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 (November) (2015) 100–106.
[20] H.J. Bandelt, P. Forster, A. Röhl, Median-joining networks for inferring intraspecific phylogenies, Mol. Biol. Evol. 16 (1) (1999) 37–48.
[21] International HapMap Consortium, et al., International HapMap Consortium et al. A second generation human haplotype map of over 3.1 million SNPs, Nature 18 (5602 December 449(7164)) (2007) 851–861.
[22] 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 (5602 December (20)) (2002) 2381–2385.
[23] 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 (September (5)) (2015) 1179–1191.
[24] K.K. Kidd, Proposed nomenclature for microhaplotypes, Hum. Genomics 10 (1 June (17)) (2016) 16.
[25] A.A. Westen, A.S. Matai, J.F. Laros, H.C. Meiland, M. Jasper, W.J. de Leeuw, P. de Knijff, T. Sijen, Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples, Forensic Sci. Int. Genet. 3 (September (4)) (2009) 233–241.
[26] ThermoFisher. cms_073986 - The AmpFℓSTR® NGM™ PCR Amplification Kit: The Perfect Union of Data Quality and Data Sharing Forensic News ThermoFisher.
[27] K.K. Kidd, W.C. Speed, A.J. Pakstis, D.S. Podini, R. Lagacé, J. Chang, S. Wootton, E. Haigh, Soundararajan U.' Evaluating 130 microhaplotypes across a global set of 83 populations, Forensic Sci. Int. Genet. 29 (July) (2017) 29–37.
[28] J.J. Sanchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P.M. Schneider, A. Carracedo, N. Morling, A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (9) (2006) 1713–1724.