Forensic Science International: Genetics 28 (2017) 35–43

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Evaluation of the Qiagen 140-SNP forensic identification multiplex for massively parallel sequencing

M. de la Puente[a], C. Phillips[a,*], C. Santos[a], M. Fondevila[a], Á. Carracedo[a,b,c], M.V. Lareu[a]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
[b] Grupo de Medicina Xenómica (GMX), Faculty of Medicine, University of Santiago de Compostela, Spain
[c] Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 6 August 2016
Received in revised form 5 January 2017
Accepted 23 January 2017
Available online 27 January 2017

Keywords:
SNPs
Forensic identification
Massively parallel sequencing (MPS)
Qiagen SNP-ID panel

ABSTRACT

A new forensic 140-SNP genotyping system from Qiagen, designed for massively parallel sequencing (MPS) analysis, was evaluated using the Ion PGM™ MPS system. Assessments consisted of the sequencing of: established control DNAs that had been previously genotyped with alternative PCR and library preparation kits supplied by Thermo Fisher Scientific for the Ion PGM™ system; a simple set of artificial DNA mixtures; DNA extracted from a degraded femur; and a dilution series to gauge forensic sensitivity. In addition to the reagents for the DNA target capture PCR and library preparation, Qiagen offer an alternative sequence analysis software system (Workbench), which was assessed alongside the Ion PGM™ Genotyper software for forensic MPS analysis. The Qiagen SNP genotyping system produced full genotyping concordance with previous data obtained with a similar SNP panel on the Ion PGM™ and in comparison to genotypes listed for 139 of the 140 SNPs in 1000 Genomes. The workbench software was as reliable as Genotyper in calling genotypes, although scrutiny of sequence data with IGV revealed the problem of sequence misalignment plagues a small proportion of the 140 SNPs in the Qiagen panel, a problem already recognized in multiple MPS studies of the same markers in alternative kits. The potential for genotype miscalls from sequence misalignment in certain SNPs will require manual inspection in cases where low-level or degraded DNA reduces the sequence coverage to a point where misalignment influences individual SNP genotype quality.

1. Introduction

In forensic identification cases, single nucleotide polymorphisms (SNPs) that can be amplified efficiently from very short DNA fragments are frequently shown to be more robust than STRs when analyzing highly degraded DNA [1,2]. This characteristic extends to Indels amplified from similarly-sized DNA [3] and indicates that the size of target DNA remaining intact after degradative processes, as well as the control of inhibition [4], are key elements in the successful analysis of challenging DNA typically found in the identification of missing persons or mass disaster victims. One drawback with the use of SNPs and Indels as binary polymorphisms, is the lower discrimination power per marker compared to STRs; requiring much larger multiplexes in order to approach the informativeness of smaller sets of STRs. Luckily, enlarged SNP multiplexes for forensic identification have become available in the last two years, forming the target capture step that prepares the extracted DNA for massively parallel sequencing (MPS) analysis [5–7]. The two MPS forensic identification panels developed so far have adopted the same strategy by combining two SNP sets previously selected to compliment STRs, by offering smaller PCR fragments than all but the shortest Mini-STRs. These two sets comprise 52 SNPs developed by SNP*for*ID [8]; and 45 SNPs extended to 88 SNPs that were compiled by *Kiddlab* [9,10], with four loci in common with SNP*for*ID. The established forensic SNP panels for MPS from Illumina (the ForenSeq DNA Signature Prep Kit [11]) and Thermo Fisher Scientific (the HID-Ion kit [6,7]) were each designed to be applied to their own MPS systems. They have now been supplemented with a third forensic SNP multiplex from Qiagen that can be used flexibly as the target DNA capture step for either MPS system.

This study reports the evaluation of the Qiagen forensic identification SNP PCR multiplex and library preparation chemistries for MPS (herein the Qiagen SNP-ID kit), applied in our assessments to the Thermo Fisher Scientific (TFS) Ion PGM™

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

http://dx.doi.org/10.1016/j.fsigen.2017.01.012
1872-4973/

sequencing system. We used a simple evaluation framework that largely followed the format for a previous study of an almost identically composed prototype version of the TFS Applied Biosystems Precision ID Identity Panel ([7], herein referred to as the HID-Ion panel). One benefit of matched formats is that the same control DNA samples were analyzed in both studies, allowing a direct comparison of the performance of each kit in terms of sequence coverage, genotyping concordance and sensitivity; the key quality parameters for gauging the reliability of MPS for forensic DNA analysis. The Qiagen SNP-ID kit has combined all 140 autosomal SNPs from the SNP*for*ID and *Kiddlab* sets, while the version of the HID-Ion panel that we previously assessed just excludes four SNPs to genotype a total of 136 autosomal loci (plus 34 Y-SNPs). It should be noted that since the original evaluations were made, the HID-Ion multiplex has been reconfigured to analyze 90 autosomal SNPs with revised PCR primer designs generating shorter amplicons in a large proportion of the SNPs.

Lastly, a problem created by adopting all 88 *Kiddlab* candidate markers is the very close linkage of SNP pairs: rs10768550-rs10500617 (separated by 679 nucleotides) and rs9606186-rs5746846 (287 nucleotides). We compiled haplotype frequencies for the 2504 samples of the 1000 Genomes project and this data complements a recent evaluation of the Qiagen SNP-ID panel, which examined its performance in the Illumina MiSeq MPS system and estimated relevant haplotype frequencies from a Swedish population sample [12].

2. Material and methods

2.1. The Qiagen SNP-ID PCR multiplex

The Qiagen PCR primer designs were adopted directly from those constructed by Andreas Tillmar's group and reported in [12]. These primer designs use a novel tiling approach to reduce the effect of flanking region SNPs causing excessive interference of primer binding in any one SNP in the 140-plex and therefore making it prone to reduced amplification efficiency. Double primer pairs were designed per SNP site to provide four different amplicons per marker. As well as balancing the melting temperatures as far as possible, designs incorporated more unstable, AT-rich 3′ primer sequence that was also less frequent in the genome to reduce off-target annealing. Supplementary Table 1 lists the sizes of the shortest amplicons generated from the four primers per SNP, and provides a simple listing of the markers in comparison to the SNP sets of the prototype and current TFS HID-Ion panels and Illumina DNA Signature Prep Kit.

2.2. DNA samples

A compact MPS validation framework was run consisting of sixteen samples of: i. five control DNA samples from a commercial supplier for genotyping concordance; ii. a dilution series of a single control DNA; iii. two library replicates of 1:1, 1:3 and 1:9 mixture ratios; and iv) two library replicates of DNA extracted from a degraded bone sample.

Genotyping concordance compared genotypes from the Qiagen SNP-ID sequence analyses with SNaPshot genotypes (Applied Biosystems, Foster City, USA) for 52 SNP*for*ID markers obtained using conventional capillary electrophoresis, plus those listed in the 1000 Genomes variant database generated by high coverage sequencing with Illumina HiSeq technology [13]. Five Coriell cell-line derived control DNA samples at 1 ng/μl were used: NA10540, NA18498, NA06994, NA11200 and HG00403. 1000 Genomes has variant genotype data for NA18498, NA06994 and HG00403.

To test the Qiagen SNP-ID kit's sensitivity, a dilution series of NA11200 was prepared at: 0.5 ng/μl, 0.25 ng/μl and 0.125 ng/μl. Mixed DNA samples were prepared with volume mixtures of NA18498 and HG00403 (1 ng/μl) at: 1 to 1; 1 to 3; and 1 to 9. A single degraded sample obtained from skeletal remains (femur) was analyzed. Previous quantification with Quantifiler® Duo (Applied Biosystems) indicated the femur extract had 0.017 ng/μl of DNA and inhibition was not detected. SNaPshot analysis of this DNA with the SNP*for*ID 52-plex test gave ~90% profile completeness.

2.3. Library preparation and sequencing steps

DNA libraries were prepared using the GeneRead™ DNAseq Targeted Panels v2 workflow (Qiagen, Hilden, Germany) and the Qiagen SNP-ID multiplex PCR primer kit. PCR followed manufacturer's guidelines, except for the amount of DNA input that was reduced at least 20 times. Therefore, PCR reactions used 1 μl volumes of the DNA samples described in Section 2.2; except the femur extract plus a negative control sample that each used the maximum 8 μl input volume. PCR cycling comprised 20 amplification cycles of 4 min at 60 °C for annealing/extension.

After purification of PCR products, a simple assessment of amplification efficiency was made using the Agilent High Sensitivity D1000 ScreenTape System (Agilent Technologies, Santa Clara, USA). This step ensured the negative control was free of DNA and it was subsequently removed from further analyses. To maximize the genotyping capacity of the TFS Ion Chip™ used, samples were barcoded during library preparation with either the Qiagen GeneRead Adapter L Set 12-plex (barcodes 1–12) or TFS Ion Xpress™ Barcode Adapters (barcodes 13–16). After purification of all library preparations, quantification and size control of the libraries was performed using the Agilent High Sensitivity D1000 ScreenTape System. Libraries were then diluted when necessary and combined to prepare a 25 pM equimolar pool. The 16 samples were analyzed with a TFS Ion 316™ Chip v2, prepared using the TFS Ion PGM™ Hi-Q™ Chef Kit.

2.4. Data analysis

Sequence data obtained from the Ion PGM™ as .bam files was analyzed in three ways: i. using the Torrent Suite™ 5.0.2 and HID_SNP_Genotyper plugin version 4.2 (herein Genotyper); ii. using Biomedical Genomics Workbench version 2.5.1 (CLC Bio, Qiagen) that applied a custom workflow as previously described [12] (herein Workbench); and iii. by detailed manual scrutiny of aligned sequences with IGV v2.3.40 [14]. Workbench default parameters were applied to genotyping concordance samples, comprising a minimum allele frequency for heterozygote calling of 0.2 and no minimum coverage threshold value. Genotyper default parameters were applied to all samples, comprising a minimum allele frequency for heterozygote calling of 0.1 and a minimum coverage on the SNP site of 6×, with no minimum coverage for each strand. For the revised analysis of mixtures, all default parameters were kept unchanged but minimum allele frequency was reset to 0.02.

Output files or posterior data were analyzed in Excel spreadsheet format. Raw sequences were aligned against the GRCh37/hg19 human reference genome and SNPs in the regions of interest were identified according to dbSNP build 144, using UCSC Genome Browser [15]. Strand bias was calculated as the percentage of forward coverage/total coverage; strand bias per allele was calculated as the percentage of allele forward reads/(allele forward reads + allele reverse reads). Allele frequencies (alternatively the Allele Read Frequency or ARF [7]) were calculated as percentage of allele reads/total coverage per SNP. The nucleotide misincorporation rate was calculated as percentage of non-allelic reads/total coverage per SNP. Note that strand bias, strand bias per allele, ARF

and misincorporation thresholds that we applied for the identification of underperforming SNPs in this study do not correspond with the parameters for SNP calling used in both Genotyper and Workbench software analysis pipelines.

1000 Genomes Phase III SNP genotype data was accessed using the online Data Slicer tool [http://browser.1000genomes.org/Homo_sapiens/UserData/SelectSlice] and processed in Excel to detect haplotypes for SNP pairs: rs10768550-rs10500617 (GRch37 coordinates: 11:5098714- 5099393) and rs9606186-rs5746846 (22:19920359-19920646). Additional SNP combinations with weaker physical linkage (rs2175957–rs8070085; rs2291395–rs4789798; and rs689512–rs3744163–rs2292972) were not analyzed. Users should apply adjustments to statistical tests handling genotypes from these SNP groups with relationship testing packages that handle linkage efficiently, such as ILIR [16].

3. Results and discussion

3.1. Overview of library preparation and sequencing performance

The overall MPS run quality assessments are summarized in Supplementary Fig. 1. PCR amplification and library preparation quantifications are shown in Supplementary Fig. 1A. PCR quantification values are lower for the femur extract and dilution series samples than the others, corresponding to reduced DNA input. Library amplification quantification values show more variability compared to those from the PCR, likely due to yield differences in target PCR, library amplification and multiple purification steps. Nevertheless, the library quantification values successfully equalized the libraries to 25 pM for subsequent sequence template preparation. Only the 1:9 mixed DNA replicate A showed an extra 77 bp peak on the Agilent electropherogram (Supplementary Fig. 1B); considered to indicate presence of adaptor dimers.

The run report indicated a high average load density of 69% (Supplementary Fig. 1C), with all the loaded Ion Sphere™ particles (ISPs) carrying sequence template. A relatively high level of polyclonality (heterogeneous sequences per ISP) of 27% was observed (Supplementary Fig. 1D), which is inside the commonly observed range of 10–30%, although close to the upper limit. The overall run read length histogram showed higher read counts for lengths corresponding to the size range of the PCR amplification; starting at ~160 bp and extending above 200 bp (Supplementary Fig. 1E). However, individual read length histograms, shown in Supplementary Fig. 1A, were less uniform among the different samples. In particular, the femur extract replicated runs had a lower number of sequence counts, which were distributed across a wider size range and detectably skewed towards smaller read length values.

3.2. Potential SNP genotyping issues revealed by IGV analyses

Sequence data from the concordance samples was examined in IGV, in particular for the characteristics of: strong strand bias; unusual read structures potentially affecting the allele calls at the target SNP; false Indels created by sequence misalignment; and nucleotide misincorporations at the target SNP position, but not reaching a minimum frequency necessary to be identified as a genotype. Supplementary Table 2 lists the main features found from IGV scrutiny plus flanking region SNPs identified in the reported sequence. Although certain sequence characteristics do not affect the genotyping reliability of the target SNP, they could influence identification of other informative polymorphisms in the reported sequences if these are considered for further analysis. Focusing on characteristics involving the target SNP, nearly 25% of markers had at least one sequence feature that could influence their posterior analysis. Such characteristics can be divided into two categories: i. unusual sequence read structures (e.g. amplicons not uniformly covering the region containing the target SNP), mainly comprising a lack of completed reads or a disproportionately low number of reads on one strand; and ii. misalignment on or around the SNP site. It is important to stress that the issues identified in the SNPs detailed in this section are characteristics of the markers and their surrounding sequences, not the Qiagen SNP-ID multiplex itself.

There were 25 SNPs that showed strand bias or below-average sequence coverage (rs891700; rs9866013; rs13182883; rs7704770; rs2272998; rs727811; rs321198; rs737681; rs4288409; rs4606077; rs1015250; rs2270529; rs1360288; rs430046; rs8070085; rs8078417; rs1024116; rs576261; rs12480506; rs2567608; rs1005533; rs1523537; rs722098; rs2831700; rs5746846). With the exception of rs4606077, rs2270529 and rs1523537, discussed in more detail below, these SNPs did not actually reveal any other characteristic that could influence the reliability of the genotype calls made. Therefore, the bulk of them represent only a higher risk of no-calls (e.g. from potentially reduced sequence data when analyzing low-level DNA), not inaccurate genotyping. An extreme example of strand bias was seen in rs430046 (Supplementary file 1A), where despite balanced detection of both C and T alleles in heterozygotes, a very small proportion of reads were made for the reverse strand. The IGV analysis of the flanking sequence of rs430046 shows the influence of a closely sited A/G SNP in a complex tract of A and G nucleotides (rs381840) that appears to affect sequence extension in the reverse strand direction. In general, we have observed that incomplete sequencing occurs by strand extension stopping in homopolymeric tracts or repetitive regions. As examination of a SNP's surrounding sequence does not provide quantitative information about how sequence coverage and strand bias will be affected by these features, posterior analysis of sequence patterns in IGV can provide important clues about the risk of obtaining lower levels of genotype calls from the sequences generated.

The second sequence feature category included SNPs that had misalignments, misincorporations, false Indels on the SNP site or allele-based strand bias (in contrast to SNP-based strand bias described above). In all, twelve SNPs in the Qiagen SNP-ID multiplex have a higher risk of producing discordant genotypes, so it is important to identify the cause of the miscalled genotypes in each case. SNPs: rs4847034; rs1554472; rs4796362; rs1004357; rs733164; rs1821380 had some degree of non-allelic nucleotide misincorporation in either the forward or reverse strands (Supplementary file 1B, 1C–G, respectively). In the first four SNPs listed above, non-allelic T reads were generated by displacement of the adjacent poly-T tract due to an extra T read. When this occurs, the SNP nucleotide is aligned as an insertion. In rs1821380 and rs733164 the cause of non-allelic T and C reads, respectively, remains unclear. SNP rs2270529 is a T/C SNP with the structure TT[T/C]GTT surrounding the SNP site, as shown in Supplementary file 1H. Several G reads appear in the SNP position on reverse strands, followed by a T insertion displacing the poly-T tract. Also in reverse strands, a proportion show a C insertion positioned before the SNP site when this is read as a T allele. These C insertions occur from the misreading of an extra T in the TT tract next to the SNP site, but both misincorporations occur at a low rate. Although other non-allelic misincorporations can reach threshold ARF values, they are not recognized nucleotides for the SNP variation and can be easily discounted.

In the SNPs prone to sequence misalignments, such corrections are not possible. In the first example, SNP rs1029047 is a T/A SNP with adjacent poly-T and poly-A tracts (TATTT[T/A]AAAAAAAAA); as shown in Supplementary file 1i. The homopolymeric tracts on each side of the SNP produce false insertions and deletions at the

SNP site and at +1 and −2 nucleotide positions, leading to considerable heterozygote imbalance and precluding rs1029047 as a reliable mixed DNA indicator, in line with the same observation in previous studies [5–7,12]. SNP rs2399332 is a G/T SNP within a long adjacent poly-T tract (Supplementary file 1J), so G reads are misaligned as insertions (mostly in forward strands), creating false T reads from miscounting the poly-T nucleotides. SNP rs4606077 is a C/T SNP with a short poly-C tract (Supplementary file 1 K). In the heterozygous sample shown, most forward reads stop in the poly-C tract at position −10, and the remaining reads show a clear imbalance with C reads present in very low numbers. This effect also occurs in flanking SNPs at +10 (rs58774517) and +11 positions (rs1869434). It can be assumed that most forward strands stopping at the poly-C tract carry the C allele; generating both allele imbalance (from several incomplete C-linked forward strands), and strand bias. However, reverse strand reads appear reliable and scrutiny of genotypes is feasible for non-challenging DNA samples.

The G/C SNP rs445251 is sited in the repetitive region GGTT[G/C]GTG (Supplementary file 1L). In this SNP reverse strands show deletions, disproportionately of C alleles, creating heterozygous imbalance and low coverage in C homozygotes. Furthermore, A-nucleotide misincorporations can occur on reverse strands followed by misincorporations in other nucleotide positions. These multiple problems indicate that misalignments are commonly occurring around the SNP site. As a result, rs445251 is characterized by misincorporations, strand bias (both SNP and allele strand bias) and lower overall coverage for the SNP position. Despite this, reads on the forward strand are observed to be reliable and the A misincorporation easily detected. Lastly, the T/C SNP rs1523537 has an unusual sequence read structure (Supplementary file 1M). Most forward strand extensions do not reach the SNP site and in a number of the remaining strands a G misincorporation occurs due to misalignment. In the forward strands the T allele predominates creating allele strand bias, while in homozygous C genotypes ~10% of reads are T alleles. However, reverse strands appear reliable and, as with all the SNPs described in this section, scrutiny and correction of genotypes is feasible for non-challenging DNA samples.

3.3. Genotyping concordance

Concordance of genotypes of the five Coriell control DNAs was assessed by comparing 3 of the 5 samples listed in the 1000 Genomes SNP variant database, which has 139 of the 140 SNPs genotyped (rs938283 not currently listed); and by comparing 52 of the 140 SNPs genotyped with SNaPshot. Genotype calls were collected independently using Genotyper and Workbench software to analyze the sequence output from the control DNAs. Genotyping concordance between MPS genotypes generated from the Qiagen SNP-ID kit and 1000 Genomes data reached a very high rate of 99.52%. However, the two genotype discordances found were different in each analysis regime and were centered on rs445251 with Workbench and SNPs rs1004357, rs5746846 in Genotyper. The explanations for each discordant genotype are outlined in Table 1, with the SNPs already identified as underperforming markers and the reasons for genotype miscalls described in detail in section 3.2. It is noteworthy that the Workbench genotype calls for rs445251 are not discordances as such, but partial genotypes that highlight a potential sequence deletion on one strand. Furthermore, the correct genotype call for rs1004357 and the successful call for the low threshold sequence coverage of rs5746846 suggest 100% concordance of Qiagen SNP-ID genotypes with 1000 Genomes data and slightly better performance than the Genotyper software supplied for forensic analysis with the Ion PGM™ system.

Concordance between Qiagen SNP-ID genotypes and SNaPshot was 98.64% (Table 1). The seven discordances were all due to interpretive difficulties that can be encountered when reading 52-plex SNaPshot profiles [17]. The discordant genotypes were correctly called by analyzing singleplex SNP genotypes with SNaPshot.

As far as our SNP genotyping accuracy assessments allowed, Workbench MPS analysis software performed as well as Genotyper. A key feature we would wish to see included in both sequence analysis regimes, is the ability for the user to fine-tune the analysis parameters of individual SNPs in a multiplex. The problems of low level nucleotide misincorporation (particularly if allelic), sequence misalignments of poly-base tracts, strong strand bias and spurious identification of Indels; detailed in section 3.2, all lend weight to the need for the user to adjust parameter thresholds to suit the characteristics of each SNP and its surrounding sequence.

3.4. Sequencing quality parameters for Qiagen SNP-ID data

Sequence data from the five concordance control DNAs were used to define the sequencing quality parameters: average coverage per marker; strand bias; reference and alternative allele

Table 1
Discordances found from comparisons between Qiagen SNP-ID genotype calls (using Genotyper and Workbench analysis software) vs. 1000 Genomes database and vs. SNaPshot.

	SNP	Sample	Genotyper	Workbench	1000 Genomes	Cause of discordance
Coriell control DNA: concordance with 1000 Genomes	rs445251	NA18498	CC	C	CC	Deletion on the SNP site
		NA06994	CC	CC	CC	Deletion on the SNP site
	rs1004357	NA18498	AT	AA	AA	T misincorporation caused by poly-base tract
	rs5746846	HG00403	NN	GG	GG	Forward coverage below the threshold

	SNP	Sample	Genotyper	Workbench	52-plex	singleplex	
Coriell control DNA: concordance with SNaPshot	rs251934 (A43)	NA10540	CC	CC	NN	CC	SNaPshot problems
		NA06994	CC	CC	NN	CC	SNaPshot problems
		NA11200	CT	CT	NN	CT	SNaPshot problems
	rs729172 (A16)	NA18498	AA	AA	AG	AA	SNaPshot problems
		NA06994	AA	AA	AG	AA	SNaPshot problems
		HG00403	AA	AA	AG	AA	SNaPshot problems
	rs917118 (A07)	NA11200	GT	GT	GN	GT	SNaPshot problems

	SNP	Replicate 1	Replicate 2	52-plex	
Femur DNA extract: concordance with SNaPshot	rs938283 (A33)	TT	TT	CT	Likely SNaPshot mistyping
	rs917118 (A07)	AC	AC	CN	Likely SNaPshot mistyping

strand bias; misincorporation rate; and ARF values, in analyses outlined in previous publications [6,7,17,18]. Supplementary Table 3 lists average values for these parameters, except ARF.

Average sequence coverage/reads per SNP for the Qiagen SNP-ID kit was 1220.3, with the range 22.2 to 2239.2 (lowest average coverage 211.2 in the concordance DNA controls). These levels are well above the 15–20 x minimum coverage thresholds adopted in many forensic MPS studies [7,17,19,20]. The average coverage heat map of Fig. 1, that ranks SNPs in increasing average coverage levels (based on concordance samples), indicates a reasonably homogenous distribution of sequence output from Qiagen's PCR multiplex. Although the 1:9 mixed DNA replicate A had some unexpected sequence coverage problems, all SNPs from the dilution series and femur extract sequence analyses were above a 20× minimum threshold and only some 5–6% of SNPs had coverage below 200×. When comparing these results to those of Grandell et al. [12], a study evaluating the same panel with a different sequencing platform (Illumina MiSeq), the main differences amongst component marker coverage can be explained by the initial multiplex PCR. As this initial amplification step is common to both analyses, similarities in the relative distribution of coverage values across markers can be expected. Two SNPs identified as under-represented by Grandell: rs1360288 and rs105883; are among the markers we identified with lower coverage values (SNPs on the left side of Fig. 1.); with average coverage values below 400× in Supplementary Table 3. Of the five markers with lower normalized coverage proportions in Grandell's study (Fig. 1 in [12]), the above SNPs rs1360288 and rs105883 plus rs5746846, rs873196 and rs2567608 are listed among the seven SNPs with lower average coverage values in this study (see Fig. 1 and Supplementary Table 3). The differences of sequence coverage in rs9951171 and rs1005533, identified in this study as low coverage SNPs, could be due to differences in input DNA, which was 20 times lower in this study, or a stochastic effect because of the limited number of samples we analyzed (5 vs. 49 by Grandell).

Strand bias data indicates 11 SNPs have highly skewed reads outside the 25% to 75% thresholds established previously [7], as highlighted in Fig. 2 (rs891700; rs9866013; rs13182883; rs727811; rs321198; rs430046; rs576261; rs2567608; rs1005533; rs722098; rs5746846). All 11 SNPs were identified as underperforming from IGV scrutiny (section 3.2), so application of a stringent minimum strand coverage parameter increases the rate of no-calls in these markers. SNP rs4606077 had very high allele strand bias, where alternative allele reads are strongly skewed to detection on reverse strand sequence extensions (as detailed in section 3.2).

Nine SNPs showed high misincorporation rates (non-allelic reads >1%): rs4847034, rs1554472, rs2270529, rs1821380, rs4796362, rs1004357, rs445251, rs1523537 and rs733164, as described in section 3.2. Notably, rs1004357 and rs733164 results suggest nucleotide misincorporation can sometimes reach values as high as 10%, although this would not be enough to create strong departures from perfectly balanced ARF values (0, 50, 100) when allelic nucleotides are misincorporated.

The ARF values, shown in Fig. 3A, indicate patterns that match well with the ranges expected for heterozygous (40–60%) and homozygous samples (up to 5% or over 95%). The ten SNPs indicated in Fig. 3A deviated in a small number of samples from these expected ranges: rs7520386; rs2046361; rs2056277; rs2833736; rs1029047; rs1478829; rs4606077; rs430046; rs3744163; rs1523537. The first four of these SNPs had not been identified as problem markers in IGV analyses (section 3.2) and only had one observation outside of the defined ARF thresholds with no apparent reason for these outlier values. Grandell et al. [12] identified three SNPs with ARF values deviating from thresholds: rs2399332, rs4530059 and rs1029047. The authors explained this effect in rs2399332 and rs4530059 by the presence of neighbouring SNPs in the primer regions; but these two markers did not show deviated ARF values in our study. However, rs2399332 has some misincorporation that can be explained by the presence of a poly-base tract (see Supplementary file 1.J). It is worth mentioning that rs2399332 was also identified using a different set of primers by Eduardoff et al. [7], as showing misincorporations, and by Børsting et al. [6], as showing allelic imbalance. SNP rs4530059 was identified as showing allelic imbalance by Børsting [6], but not Eduardoff [7]. Furthermore, only rs430046 and rs1523537 coincided with four SNPs showing outlier ARF ratios in the previous Ion PGM™ study of Eduardoff [7], which also identified rs8037429 and rs803749 (within-range ARF values in our study). Lastly, rs1029047 had ARF deviations due to a poly-base tract issue (see Supplementary file 1.i) and was also identified by Grandell et al. [12] and, with a different set of primers, by Seo et al. [5], Børsting et al. [6] and Eduardoff et al. [7]. Therefore, establishing a set of reference ARF value plots for the identification of the most balanced SNPs (to act as the most reliable mixed DNA detectors) remains a step that must be undertaken for each laboratory set-up and MPS pipeline.

3.5. Forensic sensitivity assessments

Dilution series analyses observed 100% genotype completeness and concordance for 0.5 ng, 0.25 ng and 0.125 ng input DNA. Levels of sequence coverage were also comparable with those of standard DNA input, with average coverage of 0.5 ng = 650.7; 0.25 ng = 1193.7 and 0.125 ng = 966.2. Although variation in coverage might be affected by equalizing all samples to 25 pM before template preparation, a relationship between the initial target DNA concentration and both PCR and library quantifications is evident in the data of Supplementary Fig. 1A. It is also important

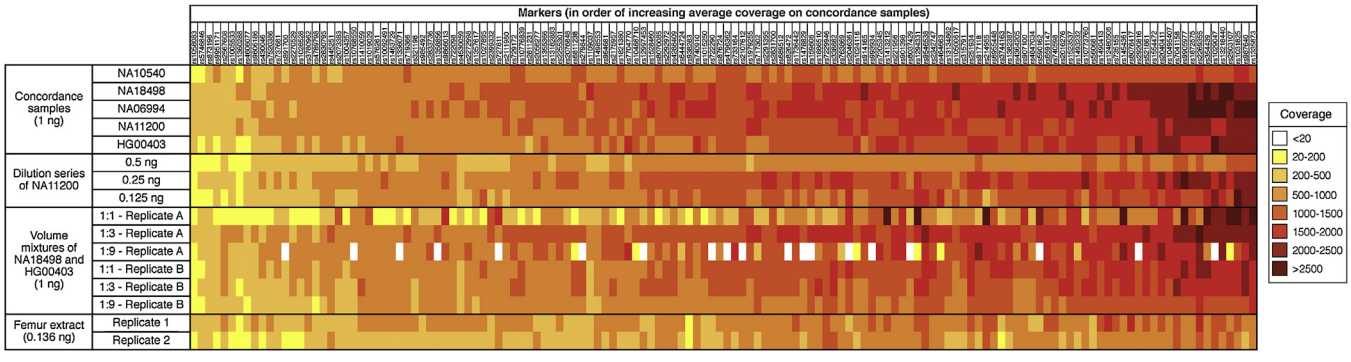

Fig. 1. Heat-map presenting coverage values for each of the 140 markers included in the panel in each sample analyzed. Markers are ordered by increasing average coverage in the concordance samples.

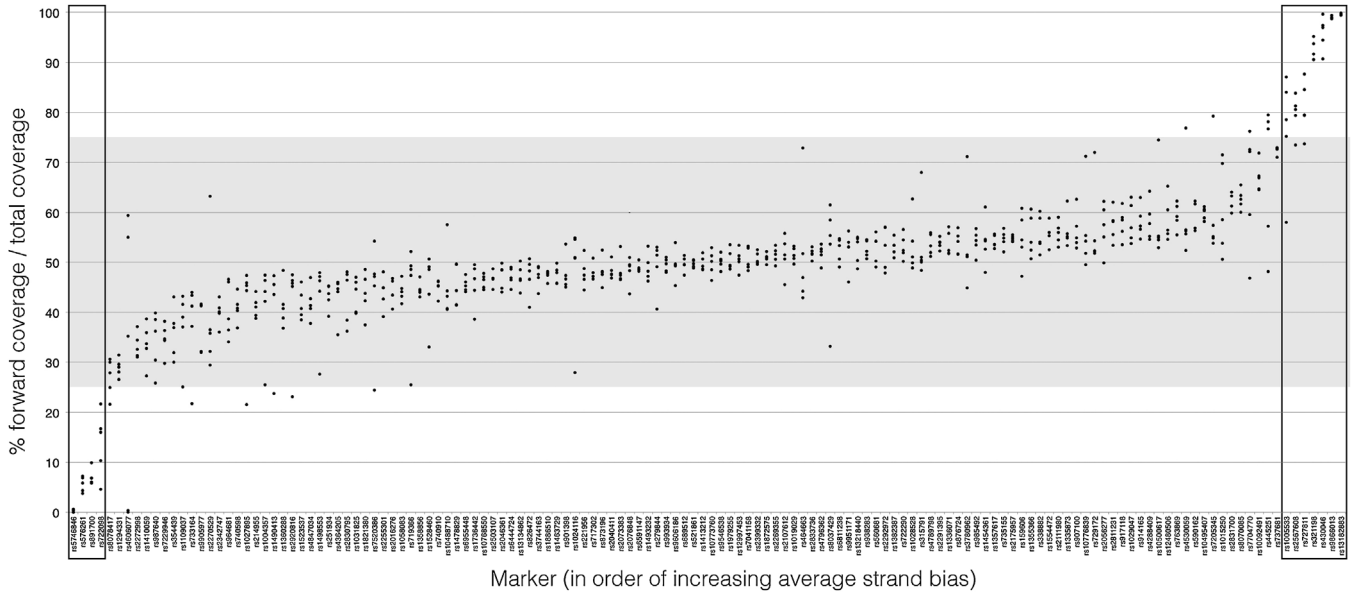

Fig. 2. Strand bias estimates (as% of forward coverage/total coverage), based on observed values in the concordance samples, for the 140 markers included in the panel. Markers with highly skewed reads (below 25% or over 75% of average strand bias) are indicated by a black box. Markers are ordered by increasing average strand bias.

to emphasize that ARF values of the dilution series samples tend to deviate from typical frequency ratios observed in heterozygous genotypes (see Fig. 3B); an effect also noticed by Grandell et al. (Fig. 3 in [12]). The largest deviations from normal balanced ARF frequency ratios were observed in the analysis of the degraded DNA from the femur extract (see Fig. 3C).

PCR of the femur extract (0.136 ng of input DNA) gave more than 0.4 ng of DNA for library preparation, and library quantification values were marginally lower than the 0.125 ng diluted non-degraded DNA. The mean sequence read length of these samples was also slightly lower than average (158 vs. 168.74 nucleotides), as shown in Supplementary Fig. 1A. Comparison of the genotypes obtained from the two femur extract library replicates gave 100% completeness and concordance. Genotype concordance between Qiagen SNP-ID genotypes and those of the 46 SNPs detected with the 52-plex SNaPshot test was 96.73%. The cause for the observed discordances in two SNPs (rs938283 and rs1031825) remains unclear, but each SNP has high reliability, shows 100% control DNA concordance and above-average coverage values in control DNA analyses, suggesting the SNaPshot genotypes were most likely to have been misidentified. Average coverage values of 704.28 and 524.54 were obtained for replicates 1 and 2 (Supplementary Fig. 1A, femur extract), which, although slightly lower than values from other samples, did not fall below a 20 x coverage threshold in any SNP.

3.6. Mixed DNA detection

The balanced number of sequence reads observed for each SNP's allele when analyzing single-source samples allows the establishment of homozygote and heterozygote ARF thresholds, to provide reference plots, as shown in Fig. 3A. However, the sequence data from mixed DNA samples disrupts the balanced patterns normally observed, from the addition of extra copies of each allele (Supplementary Fig. 2). Mixed DNA ARF distributions are quite distinct from those of single-source samples, with high numbers of heterozygous SNPs outside the 40–60% ARF region. These patterns were discernible in the mixed DNA analyzed with the Qiagen SNP-ID multiplex, particularly in the 1:1 and 1:3 ratios. Furthermore, heterozygosity proportions increased from approximately 48% in the single-source component samples (NA18498 and HG00403) to 80% in the 1:1 mixture.

Following a previous study of mixed DNA with SNPs genotyped using the HID-Ion™ kit [7], we compared the performance of two different parameter sets in the Genotyper analysis software using Germline parameters (including min_allele_freq = 0.1) and Somatic parameters considered to be more effective to detect low frequency variants (min_allele_freq = 0.02). As the version of Genotyper we used prevents choice of Germline or Somatic parameters, Germline parameters were made default and min_-allele_freq was manually adjusted to 0.02 (herein, "Lower allele freq", other parameters unchanged). This reduced stringency of the analysis parameters aims to reduce discordancy between the expected mixture genotypes and the reported genotypes by detecting minor alleles with ARF values of 2%-10%. Use of the Lower allele freq parameter reduced the dropout rate from 11.4% to 0.8% (7 allele dropouts in 6 SNPs in the 1:9 mixture). However, in both 1:9 mixture replicates, rs10488710 had an allele dropout despite the ARF being 0.023 and 0.025, which could not be explained. The use of different analysis parameters did not significantly affect the no-call rate (4.3% with default settings; 3.5% with 0.02 min allele freq), with all the missing genotypes due to low SNP coverage.

When applying Genotyper software to mixed DNA analysis, it is important to consider that setting a less stringent minimum allele frequency threshold value raises the risk of incorrectly calling a homozygous SNP as a heterozygote (dropin), as a number of homozygous SNPs show ARF values that deviate from an 'ideal' 0 and 100 on single source samples (see Fig. 3A). In this study, no dropins were found when applying both 0.1 and 0.02 thresholds. Moreover, particular care must be taken when analyzing the underperforming SNPs described above. In the present study two interventions had to be made. First, in the HG00403 component DNA the rs5746846 genotype had to be corrected as described in section 3.4. This SNP has strong strand bias with only reverse strand sequence extensions, creating a no-call genotype. Second, all genotypes of rs1004357 were corrected to homozygous A genotypes after scrutiny of the samples using IGV. In section 3.2 this SNP was described as problematic due to a small poly-T tract adjacent to the SNP position that inserts a non-allelic T nucleotide.

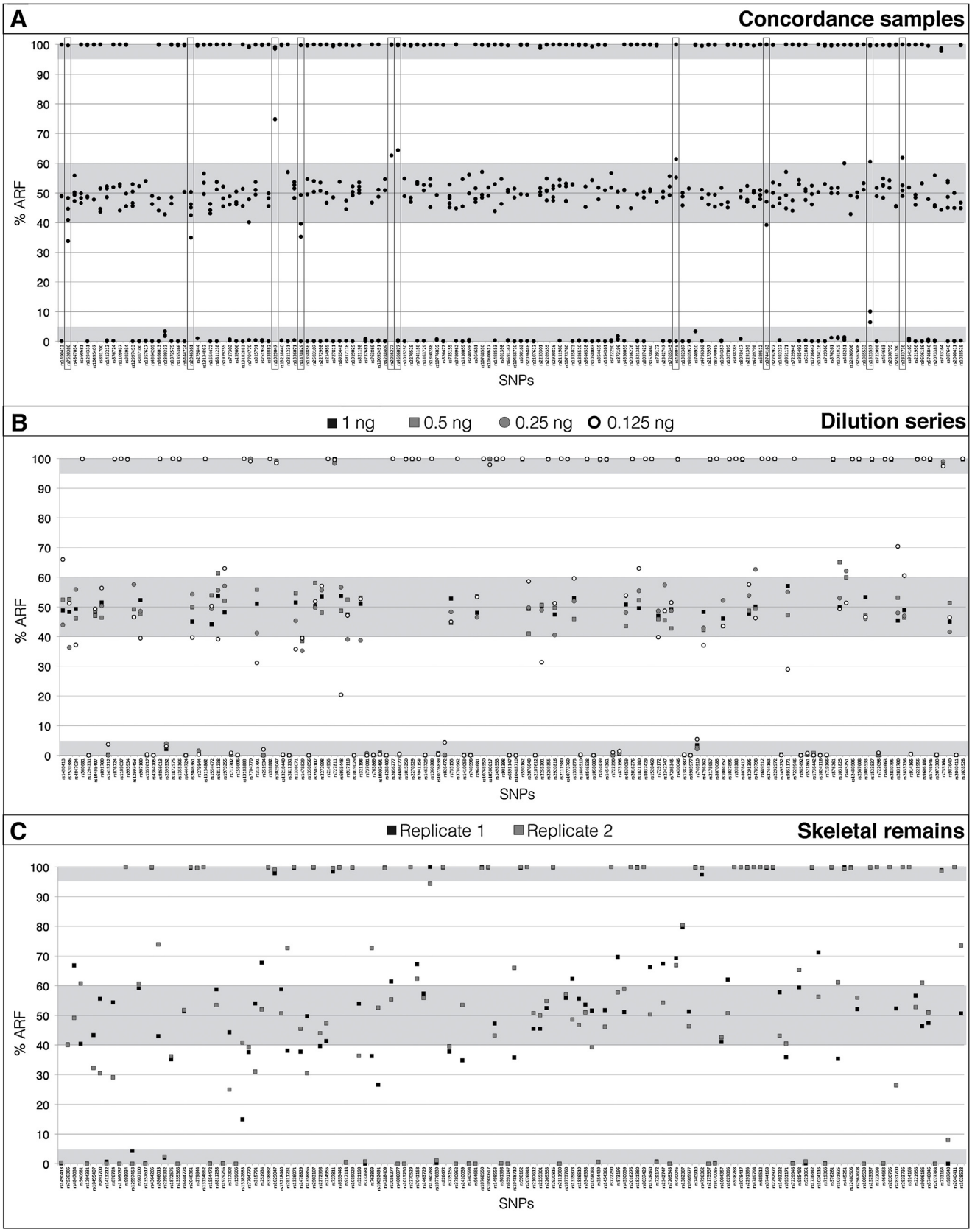

Fig. 3. ARF values as the percentage of reference allele reads/coverage for different set of samples. Markers displayed on the original output order. Grey areas represent the expected heterozygous (between 40 and 60%) and homozygous values (below 5% or over 95%). A: Concordance samples. Markers deviating in one or two samples from expected are indicated by a black box. B: Dilution series of NA11200 samples. Legend indicates each concentration symbol. C: Femur extract sample. Legend indicates each replicate symbol.

To summarize, the analysis of a limited set of mixtures indicates that the reduction of the min_allele_freq parameter in Genotyper is a crucial step to achieve high quality genotype calls for mixed DNA. This adjusted parameter should be applied retrospectively in cases where a sample is suggested to be a probable mixture from the observation of higher heterozygosity levels and disrupted ARF patterns.

3.7. Haplotype frequency estimates from 1000 genomes data

The summary haplotype frequencies from 1000 Genomes data for SNP pairs rs10768550-rs10500617 and rs9606186-rs5746846 are shown in Table 2 (for four population groups and admixed American populations) and Supplementary Table 4 (individual haplotypes and 26 population haplotype frequencies). In SNP pair rs10768550-rs10500617 the CA haplotype is only seen twice and the TT is very rare in non-African populations (the presence of TT haplotypes in the admixed American population samples is likely from African admixture). In SNP pair rs9606186-rs5746846 the GC haplotype occurs more frequently, but interestingly there are no CG haplotypes recorded, suggesting an extremely low recombination rate between these two SNPs. Therefore, the inclusion of these two SNP pairs slightly reduces the overall informativeness of the panel and necessitates the use of haplotype frequencies in place of independent allele frequency estimates for each pair's component SNPs.

4. Concluding remarks

In this study a new 140-plex kit for forensic SNP genotyping with MPS was assessed and data from our evaluations indicated it performed well. The results reported are based on a limited number of samples, but because most of them were previously genotyped with the same MPS detection system but a different PCR and library preparation; we observed comparable levels of sequence coverage from each kit. Although we carefully checked the sequence alignments and genotype calls by scrutiny of IGV data, none of the SNP genotyping issues detected were related to the Qiagen chemistry but to the context sequence characteristics of the SNPs themselves. Not surprisingly, many of these problem SNPs have been removed from the final commercial version of the original SNP set developed by TFS for their Ion PGM™ system (the *Precision ID Identity Panel*, Applied Biosystems, TFS, Carlsbad, USA). Therefore, use of the Qiagen SNP-ID PCR kit will require the analysis of certain SNPs with more care and individual genotype calls may require extended checks with IGV.

Two software analysis systems were compared by applying each regime to the same sequence data. Both Workbench and Genotyper gave near identical performance in accomplishing consistent genotype calls with low-level input DNA and an example degraded DNA extract, as well as control DNA samples. A single SNP and sample had non-allelic nucleotide misincorporation called as a genotype and one no-call genotype with Genotyper, to form the only differences in precision between the software regimes. If the deletion recorded on one strand of SNP rs445251 with Workbench is discounted (as this genotype would be individually checked during the analysis of sequencing results), the Qiagen SNP-ID kit and Workbench software analysis system achieved 100% genotyping concordance, underlining its reliability.

A case can be made for replacing or removing SNPs that may require manual inspection to check their genotyping reliability when analyzing degraded or low-level DNA. Similarly, the two very

Table 2
Haplotype frequencies from 1000 Genomes variant data, for the two closest SNP pairs in the Qiagen SNP-ID kit: rs10768550-rs10500617 (679 nucleotide separation) and rs9606186-rs5746846 (287 nucleotide separation). RA: reference allele, AA: alternative allele.

SNP pair	rs10768550		rs10500617		rs9606186	rs5746846
RA	C		T		C	C
AA	T		A		G	G
	SNP haplotype	Haplotype counts	Haplotype frequencies	SNP haplotype	Haplotype counts	Haplotype frequencies
African[a]	CT	631	0.6260	GG	468	0.4643
	TA	227	0.2252	CC	388	0.3849
	CA	1	0.0010	CG	0	0.0000
	TT	149	0.1478	GC	152	0.1508
		1008			1008	
European	CT	740	0.7356	GG	545	0.5417
	TA	265	0.2634	CC	416	0.4135
	CA	0	0.0000	CG	0	0.0000
	TT	1	0.0010	GC	45	0.0447
		1006			1006	
South Asian	CT	763	0.7802	GG	528	0.5399
	TA	215	0.2198	CC	340	0.3476
	CA	0	0.0000	CG	0	0.0000
	TT	0	0.0000	GC	110	0.1125
		978			978	
East Asian	CT	611	0.6062	GG	716	0.7103
	TA	396	0.3929	CC	279	0.2768
	CA	1	0.0010	CG	0	0.0000
	TT	0	0.0000	GC	13	0.0129
		1008			1008	
Admixed American	CT	482	0.6945	GG	413	0.5951
	TA	198	0.2853	CC	239	0.3444
	CA	0	0.0000	CG	0	0.0000
	TT	14	0.0202	GC	42	0.0605
		694			694	

[a] Excludes Americans of African Ancestry in SW USA and African Caribbeans in Barbados.

closely linked SNP pairs are not applicable as independent loci; with informativeness as haplotypes almost the same as using one of the SNPs per pair (with the exception of African frequency data and South Asian frequencies for rs9606186-rs5746846). Nevertheless, for the application of a short-amplicon SNP set that is an appropriate choice for forensic identification cases involving analysis of degraded DNA, MPS workflows already require detailed checks on the genotype calls made from such material, and the same care is required with SNaPshot SNP genotyping using smaller multiplexes.

Acknowledgments

The authors would especially like to thank Andreas Tillmar of the Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden, for very helpful support in optimizing the MPS protocols and adapting the Workbench software system for forensic SNP analysis. MdlP is supported by funding awarded by the Consellería de Cultura, Educación e Ordenación Universitaria of the Xunta de Galicia as part of the Plan Galego de Investigación, Innovación e Crecemento 2011–2015 (Plan I2C).

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.2017.01.012.

References

[1] L.A. Dixon, A.E. Dobbins, H.K. Pulker, J.M. Butler, P.M. Vallone, M.D. Coble, W. Parson, B. Berger, P. Grubwieser, H.S. Mogensen, et al., Analysis of artificially degraded DNA using STRs and SNPs –results of a collaborative European (EDNAP) exercise, Forensic Sci. Int. 164 (2006) 33–44.

[2] M. Fondevila, C. Phillips, N. Naverán, L. Fernandez, M. Cerezo, A. Salas, Á. Carracedo, M.V. Lareu, Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur, Forensic Sci. Int. Genet. 2 (2008) 212–218.

[3] C. Romanini, M.L. Catelli, A. Borosky, R. Pereira, M. Romero, M. Salado Puerto, C. Phillips, M. Fondevila, A. Freire, C. Santos, et al., Typing short amplicon binary polymorphisms: supplementary SNP and Indel genetic information in the analysis of highly degraded skeletal remains, Forensic Sci. Int. Genet. 6 (2012) 469–476.

[4] M. Fondevila, C. Phillips, N. Naverán, M. Cerezo, A. Rodríguez, R. Calvo, L.M. Fernández, Á. Carracedo, M.V. Lareu, Challenging DNA: assessment of a range of genotyping approaches for highly degraded forensic samples, Forensic Sci. Int. Genet. Supplement Series 81 (2008) 1–3.

[5] S.B. Seo, J.L. King, D.H. Warshauer, C.P. Davis, J. Ge, B. Budowle, Single base polymorphism typing with massively parallel sequencing for human identification, Int. J. Legal Med. 127 (2013) 1079–1086.

[6] C. Børsting, S.L. Fordyce, J. Olofsson, H. Smidt Mogensen, N. Morling, Evaluation of the Ion Torrent™ HID SNP 169-plex: a SNP typing assay developed for human identification by second generation sequencing, Forensic Sci. Int. Genet. 12 (2014) 144–154.

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

[8] J.J. Sanchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, et al., A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713–1724.

[9] A.J. Pakstis, W.C. Speed, J.R. Kidd, K.K. Kidd, Candidate SNPs for a universal individual identification panel, Hum. Genet. 121 (2007) 305–317.

[10] A.J. Pakstis, W.C. Speed, R. Fang, F.C. Hyland, M.R. Furtado, J.R. Kidd, K.K. Kidd, SNPs for a universal individual identification panel, Hum. Genet. 127 (2010) 315–324.

[11] J.D. Churchill, S.E. Schmedes, J.L. King, B. Budowle, Evaluation of the Illumina® beta version ForenSeq™ DNA signature prep kit for use in genetic profiling, Forensic Sci. Int. Genet. 20 (2015) 20–29.

[12] I. Grandell, R. Samara, A.O. Tillmar, A SNP panel for identity and kinship testing using massive parallel sequencing, Int. J. Legal Med. 130 (2016) 905–914.

[13] 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature 526 (2015) 68–74.

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

[15] 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.

[16] A.O. Tillmar, C. Phillips, Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets, Forensic Sci. Int. Genet. 26 (2017) 58–65.

[17] R. Daniel, C. Santos, C. Phillips, M. Fondevila, R.A. van Oorschot, Á. Carracedo, M.V. Lareu, D. McNevin, A SNaPshot of next generation sequencing, Forensic Sci. Int. Genet. 14 (2014) 50–60.

[18] M. Eduardoff, T.E. Gross, C. Santos, M. de la Puente, D. Ballard, C. Strobl, C. Børsting, N. Morling, L. Fusco, C. Hussing, B. Egyed, L. Souto, J. Uacyisrael, D. Syndercombe Court, P.M. Schneider, Á. Carracedo, M.V. Lareu, The EUROFORGEN-NoE Consortium, W. Parson, C. Phillips, 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.

[19] 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.

[20] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, Y. Gu, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC Genomics 13 (2012) 1–13.