Title Page

Title

Massively parallel sequencing of customised forensically informative SNP panels on the MiSeq

Authors

Bhavik Mehta[1,*], Runa Daniel[2], Chris Phillips[3], Stephen Doyle[4], Gareth Elvidge[5], Dennis McNevin[1]

Affiliations

[1] National Centre for Forensic Studies, Faculty of ESTeM, University of Canberra, Canberra, Australia

[2] Office of the Chief Forensic Scientist, Victoria Police Forensic Services Department, Melbourne, Australia

[3] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain

[4]Department of Animal, Plant and Soil Sciences, La Trobe University, Melbourne, Australia

[5] Illumina Inc. USA

*** Corresponding author**

Mr. Bhavik Mehta

Address: National Centre for Forensic Studies

Level D, Building 7, University of Canberra

Bruce, ACT 2617 Australia.

Email: bhavik.mehta@hotmail.com

Abbreviations

BAM	binary alignment / map file format
BGA	biogeographical ancestry
CE	capillary electrophoresis
EVCs	externally visible characteristics
gVCF	genome variant calling file
HA	humic acid
MCS	MiSeq Control Software
MPS	massively parallel sequencing
MSR	MiSeq Reporter
NGS	next generation sequencing
PCR	polymerase chain reaction
SBE	single base extension
SBS	sequencing-by-synthesis
UV	ultra violet
VCF	variant calling file

Received: MONTH DD, YYY; Revised: MONTH DD, YYY; Accepted: MONTH DD, YYY

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/elps.201600190.

Key words

Forensic DNA genotyping

Illumina MiSeq

Next generation sequencing (NGS)

Single nucleotide polymorphisms (SNPs)

SNaPshot

Abstract

Forensic DNA based intelligence, or Forensic DNA Phenotyping (FDP), utilises single nucleotide polymorphisms (SNPs) to infer the biogeographical ancestry (BGA) and externally visible characteristics (EVCs) of the donor of evidential material. SNaPshot® is a commonly employed forensic SNP genotyping technique which is limited to multiplexes of 30-40 SNPs in a single reaction and prone to PCR contamination. Massively parallel sequencing (MPS) has the ability to genotype hundreds of SNPs in multiple samples simultaneously by employing an oligonucleotide sample barcoding strategy. This study of the Illumina MiSeq MPS platform analysed 136 unique SNPs in 48 samples from SNaPshot PCR amplicons generated by five established FDP assays comprising the SNP*for*ID 52-plex, SNP*for*ID 34-plex, Eurasiaplex, Pacifiplex and IrisPlex. Approximately 3 Gb of sequence data was generated from two MiSeq flow cells and profiles were obtained from just 0.25 ng of DNA. Compared with SNaPshot, an average 98% genotyping concordance was achieved. Our customised approach was successful in attaining SNP profiles from extremely degraded, inhibited and compromised casework samples. Heterozygote imbalance and sequence coverage in negative controls highlight the need to establish baseline sequence coverage thresholds and refine allele frequency thresholds. This study demonstrates the potential of the MiSeq for forensic SNP analysis.

1 Introduction

Forensic DNA-based intelligence, commonly known as forensic DNA phenotyping (FDP) or molecular photofitting [1, 2], utilises genetic markers associated with phenotypes including biogeographical ancestry (BGA) and externally visible characteristics (EVCs) to predict the appearance of the donor of evidential material. FDP is rapidly emerging as a potentially powerful tool in criminal investigations particularly when STR genotyping produces partial or non-informative profiles [1].

The most common approach for forensic SNP genotyping has been single base extension (SBE) using the SNaPshot® assay (Applied Biosystems) which utilises capillary electrophoresis (CE) detection [3, 4] and other equipment commonly used in forensic laboratories. Numerous SNP-based forensic intelligence SNaPshot® assays have been developed, including the SNP*for*ID 34-plex [5, 6], Eurasiaplex [7], IrisPlex [8] and HIrisPlex [9]. Some limitations associated with SNaPshot® include an upper multiplexing limit of ~30-40 SNPs in a single PCR assay [10] and the need for multiple tube transfers which increase the risk of contamination [10, 11]. Next generation sequencing (NGS), alternatively termed massively parallel sequencing (MPS), can simultaneously genotype hundreds of markers in multiple samples using small amounts of DNA. High throughput MPS platforms, such as the HiSeq (Illumina) and SOLiD (Applied Biosystems) systems, are cost effective for sequencing whole genomes [12]. Low to medium throughput benchtop sequencers such as the Ion PGM™ (Applied Biosystems) and MiSeq (Illumina) operate at a more appropriate scale for forensic laboratories. Recently, the applicability of the Ion PGM™ for forensic autosomal SNP genotyping has been demonstrated [10, 13]. This study reports on the application of the MiSeq system to genotype autosomal SNPs in a combination of existing customised panels.

The MiSeq employs sequencing by synthesis (SBS) chemistry. Individual DNA molecules are attached to a glass slide (flow cell) and clonally amplified in clusters via bridge PCR [14, 15]. The MiSeq can generate up to 15 Gb (~25 million reads) of sequence data on a single flow cell (version 3) and can be applied to targeted

sequencing of forensically informative markers [12]. This has been demonstrated on the forensic specific MiSeq FGx™ (Forensic Genomics System) with a beta version of the ForenSeq™ DNA Signature Prep Kit (Illumina) [16].

In this study, SNPs from five forensic SBE assays were combined and genotyped on the MiSeq. These were the SNP*for*ID 52-plex for identity [17] ; the SNP*for*ID 34-plex [6], Eurasiaplex [7] and Pacifiplex [18] for BGA; and IrisPlex [8] as an EVC predictive test. Six forensic validation parameters were examined: sensitivity; reproducibility; genotype concordance; effect of different DNA extraction methods; ability to genotype compromised samples including bone and humic acid (HA) inhibited extracts; and ability to genotype UV degraded extracts.

2 Materials and Methods

2.1 Sample Preparation

Ethics approval to collect DNA for this study was granted by the University of Canberra Committee for Ethics in Human Research (project number 11-119 and its extension, 15-64). Seven human DNA templates (S1-S7) were extracted from buccal swabs using the DNA-IQ™ System (Promega) following the manufacturer's recommended protocol. Extracts were quantified using Quantifiler™ Human DNA Quantitation Kit (Applied Biosystems) following the manufacturer's recommended protocol together with the two standard reference materials (SRMs): human male cell line control DNA 007 (Applied Biosystems) and human female cell line control DNA 9947A (Applied Biosystems).

2.2 Preparation of PCR amplicons

PCR amplicons were generated using published primer sequences and reaction protocols for five forensic multiplex PCR assays: SNP*for*ID 52-plex, SNP*for*ID 34-plex, Eurasiaplex, Pacifiplex and IrisPlex. The five multiplex assays together comprise 145 SNP amplicons with nine SNPs (rs1024116, rs1335873, rs12913832, rs16891982, rs1886510, rs204041, rs3827760, rs722098 and rs917118) shared in multiple assays resulting in 136 unique SNP amplicons ranging from 51-156 bp.

2.3 SNaPshot® genotyping of PCR amplicons

SNaPshot® genotyping was performed following the published protocols for each assay [6-8, 17, 18] to assess the efficiency of the PCR reactions used to generate the amplicons for sequencing and to obtain genotypes for concordance studies.

2.4. Forensic validation parameters

The study assessed the following six forensic validation parameters.

2.4.1 Sensitivity

A sensitivity study was conducted on three DNA templates: 9947A, 007 and S1 using template input amounts for each multiplex PCR assay of 0.05, 0.1, 0.2, 0.3 and 0.5 ng (total of 0.25, 0.5, 1.0, 1.5, 2.5ng).

2.4.2 Reproducibility

Four replicates of sample S1 at 0.3 ng for each multiplex PCR assay (total of 1.5 ng) were used to assess reproducibility.

2.4.3 Genotype Concordance

Seven human DNA templates (S1-S7), 9947A and 007 were genotyped on the MiSeq at 0.5 ng for each multiplex PCR assay (total of 2.5 ng) and the resulting genotypes compared with those from SNaPshot®.

2.4.4 Effect of different DNA extraction methods

Sample S2 was extracted using three DNA extraction methods: DNA-IQ™ System (Promega), Isolate II (Bioline), both following the manufacturer's recommended protocols, and a standard phenol/chloroform extraction with ethanol precipitation [19]. The three DNA extracts were genotyped at 0.5 ng for each multiplex PCR assay (total of 2.5 ng) in the same MiSeq run and genotype concordance between the samples was assessed.

2.4.5 Effect of UV irradiation

A one-step UV degradation method was adapted for generating artificially degraded samples [20]. Aliquots (of 5 µL) DNA extracts of S2 and S3 (0.5 ng/µL in 0.2 ml) in PCR tubes were exposed to UV light for 30 and 60 minute intervals. The UV light was generated from a 10W source (Sankyo Denki, 254nm, UV-C) at a distance of 13 cm. PCR amplicons were generated from 1 µL of each irradiated sample for each time interval.

2.4.6 Effect of humic acid inhibition

Humic acid (HA) was used to mimic the inhibition encountered in casework samples from items such as soil. Humic acid (Sigma Aldrich) at 75 ng and 100 ng quantities were added to the PCR reactions of samples S2 and S3 (both at 0.5 ng/µL).

2.4.7 Compromised casework samples

DNA extracts from five aged bone samples (S8, S9, S10, S11 and S12) and one aged blood sample (S13) were obtained from three forensic laboratories. Due to limited sample availability they were only submitted for four multiplex PCR assays: 34-plex, Eurasiaplex, Pacifiplex and IrisPlex (total 93 SNPs). The aged bone samples (S10 and S11) were recovered from Papua New Guinea (suspected to be from World War II). The aged blood sample S13 had been stored at room temperature for 45 years.

2.5 MiSeq MPS library preparation

PCR products (2 µL) from the five multiplex PCR assays for each template were pooled together (10 µL total). A 5 µL aliquot of the pool was used for library preparation. The PCR negative controls from each assay were also pooled. The pooled templates were subjected to the TruSeq® ChIP ligation (Illumina) library preparation step following the manufacturer's protocols. Normalisation of the barcoded libraries was based on quantitation using Qubit (Applied Biosystems) following the manufacturers' recommended protocol. The normalised libraries were pooled into a final concentration of 10 nM.

2.6 MiSeq sequencing template preparation

The 10 nM barcoded library pool was diluted to 4 nM and denatured using 0.2 N NaOH following the manufacturer's recommended protocol [21]. The denatured library was further diluted to 1 pM for loading onto the MiSeq sequencing cartridge. The sequencing control comprised the phiX control library (Illumina). A volume of 600 µL of the 1 pM barcoded denatured library with 5% (v/v) 12.5 pM phiX control was sequenced using the MiSeq v3 600 cycles sequencing kit [21-23]. Paired end sequencing was performed using a 2 x 101 bp cycle setting. Two flow cells were used to sequence 24 samples per chip (SI Table 1).

2.7 MiSeq data analysis

Image processing, base calling and base quality scoring were performed with MiSeq Control Software v. 2.5 (MCS: Illumina) using default parameters. The MiSeq Reporter (MSR) software had a default upper limit coverage reporting maximum of 5000 reads per amplicon. The human reference genome hg19 (GRCh37) was used for alignment [24] and sequence output was generated in BAM (binary alignment / map) format. The BAM files were used to generate VCF (variant calling) and gVCF (genome variant calling) files for each sample. The VCF files were analysed by VariantStudio (v2.1) variant analysis software (Illumina) to generate Excel output files. The Excel and gVCF files provided the input for our custom macros to obtain coverage data for each nucleotide (Supporting Information (SI) File 1).

2.8 Allele calls

MSR variant caller default allele call thresholds (80% allele frequency or greater for homozygotes and between 20% and 80% for heterozygotes) and parameters including genotyping quality (GQx) scores were employed. GQx is a phred-scale confidence score for genotype designation [25, 26]. No baseline coverage thresholds were applied. SNPs with no genotype calls and genotypes with GQx < 99 were categorised collectively as 'missing' and discounted from further analysis.

2.9 Statistical analysis

Non-parametric statistical tests using the IBM SPSS package (v. 21) were applied to the data due to skewed (non-normal) amplicon coverage distribution. A Mann-Whitney U test was used to assess the null hypothesis of no significant difference in amplicon length (bp) and GC content (%) between amplicons with the highest and lowest 10% combined coverage for three templates (9947A, 007 and S1). Kendall's tau and Spearman's rho rank correlation coefficients were determined to identify any correlation between amplicon length and GC content over the entire coverage distribution for each template (9947A, 007, and S1). A Kruskal-Wallis test was employed to assess the null hypothesis of no significant difference in combined coverage distribution across all three templates in each assay. A Kruskal-Wallis test was also used to test the null hypothesis of no difference between the combined coverage across all four replicates (of S1 at 1.5 ng) in each assay.

3 Results

The 48 samples, sequenced on two MiSeq flow cells, generated 29.5 million reads in total. Allele frequency variation is compared with depth of coverage in SI Figure 1 for samples 9947A, 007 and S1.

3.1 Sensitivity

MiSeq genotype concordance between template amounts (0.25, 0.5, 1.0, 1.5, and 2.5 ng) for 9947A, 007 and S1 was 97.6%, 99.3% and 97.0% respectively. The genotype concordance between SNaPshot® and MiSeq was 96.0% to 99.3% across all template amounts and samples. Genotypes from the same three samples at 0.5 ng, 1.0 ng and 1.5 ng template amounts were compared with previously obtained Ion PGM™ genotypes reported by Daniel et al. [12] with concordance between 97% to 100%. The percentage of missing data for 9947A, 007 and S1 ranged from 0.7% to 9.5% across all template amounts (SI Table 2). On average, MiSeq genotypes were 97.7% and 98.5% concordant with SNaPshot® and Ion PGM™ respectively (SI Table 3).

3.2 Reproducibility

A Kruskal-Wallis test confirmed that there was a significant difference in coverage distribution across all four replicates (p=0.000). Reproducibility was 97.2 % - 99.3 % (with 4/145 and 1/145 SNPs not present in one

sample and three samples respectively: SI Table 4). However, excluding missing SNPs, genotypes between all four replicates of S1 (at 1.5 ng) were 100% concordant.

3.3 Genotype concordance

Missing SNaPshot® genotypes are shown in SI Table 5. Excluding these missing genotypes, there were between 0 and 5 discordant SNPs for S1, S2, S3, S4, S5, S6, S7, 9947A and 007 respectively (SI Table 5). The overall concordance between MiSeq and SNaPshot® genotypes ranged from 96.5% to 100.0% for all samples (SI Table 6).

3.4 Effect of different DNA extraction methods

The three different extractions of sample S2 (DNA IQ, Isolate II and phenol/chloroform) had 2, 3 and 2 missing SNPs respectively (SI Table 7). Excluding the missing SNPs, 100% genotype concordance was observed between all three extracts.

3.5 Effect of UV irradiation

Exposure of sample S2 to UV for 30 and 60 minutes resulted in 54.5% and 29.0% reportable SNPs respectively (Figure 1). Sample S3 yielded 60.0% and 30.0% reportable SNPs at 30 and 60 minutes UV exposure respectively (Figure 1 and SI Table 7). Genotype concordance between UV exposed samples and corresponding original samples are indicated in SI Table 7. Non-concordant genotypes ranged from 5.0% to 20.0% and were solely due to loss of alleles in the UV exposed samples (an example is shown in SI Figure 2). SNaPshot® genotyping of sample S2 exposed to 60 minutes of UV was unsuccessful using the 52-plex and 34-plex assays, whereas the MiSeq produced partial profiles under these extreme UV degradation conditions.

3.6 Effect of humic acid inhibition

Sample S2 spiked with 75 ng and 100 ng of humic acid returned 65.5% and 60.0% reportable SNPs respectively (Figure 1). Sample S3 similarly returned 69.5% and 59.0% SNPs, respectively (Figure 1). Excluding missing SNPs, the genotype concordance between the humic acid inhibited samples and the original samples ranged between 99.0% and 100.0% (SI Table 7). Figure 2 shows multiplex PCR assays containing bovine serum albumin (BSA) in their PCR reaction mixes (52-plex and 34-plex) generated better coverage compared to those assays without BSA. However, no SNaPhsot® SNP profiles were generated with sample S2 spiked with 100 ng humic acid using the SBE 52-plex and 34-plex assays (data not shown).

3.7 Compromised casework samples

Samples S8, S10 and S11 were below the detection limit for Quantifiler but produced 4%, 10% and 12% reportable SNPs (out of a total of 93) respectively (Table 1). Samples S9, S12 and S13, with DNA concentrations either undetected or less than 0.01 ng/μL, gave 89%, 92% and 86% reportable SNPs respectively (Table 1).

3.8 Amplicon coverage bias

Inconsistent and skewed coverage between different amplicons was observed for all templates. In samples 9947A, 007 and S1, 66% of amplicons with the highest 10% of combined coverage across all template amounts were common to all three. Similarly, all the amplicons with the lowest 10% of coverage were common to all three. Table 2 shows the SNPs with the lowest and highest 10% of combined coverage across the three samples at all template amounts. The data indicates that coverage is amplicon-dependent with rarely sequenced amplicons in common across templates and template amounts, and highly sequenced amplicons also in

common. The effect of amplicon length, GC content and PCR assay on coverage bias was subsequently examined.

3.8.1 Effect of amplicon length on coverage

Sequence coverage as a function of amplicon length for samples 9947A, 007 and S1 at five template input amounts is shown in SI Figure 3. Amplicon lengths for the highest and lowest 10% of amplicons by coverage ranged from 86 to 118 bp and 51 to 156 bp respectively, with medians of 93 bp and 83 bp, respectively (Table 2). A Mann-Whitney U test rejected the hypothesis that there was no difference between the amplicon lengths of the SNPs with the highest and lowest 10% of combined coverage (p = 0.040). In addition, Kendall's tau and Spearman's rho tests performed on the entire coverage distribution showed a weak correlation between amplicon length and coverage (r^2 = 0.251 with p = 0.000 and r^2 = 0.354 with p = 0.000, respectively).

3.8.2 Effect of GC content on coverage

Sequence coverage as a function of amplicon GC content for samples 9947A, 007 and S1 at five template input amounts is shown in SI Figure 4. The average GC content across the amplicons was 44%. The GC content of the highest and lowest 10% of amplicons by coverage ranged from 37% to 51% and 31% to 47%, respectively, with medians of 45% and 43% (Table 2). A Mann-Whitney U test supported the hypothesis of no difference in GC content between the amplicons with the highest and lowest 10% of combined coverage (p = 0.436). In addition, Kendall's tau and Spearman's rho tests performed on the entire coverage distribution indicated that there was no significant correlation between GC content and coverage (r^2 = -0.013 with p = 0.410 and r^2 = -0.017 with p = 0.437, respectively). Thus, there is no evidence to support the hypothesis that coverage bias is associated with GC content.

3.8.3 Effect of PCR assay on coverage

A skewed distribution of coverage was observed within each multiplex PCR assay (SI Figure 5) as well as inconsistency in representation of assays between samples. For example, the 52-plex was under-represented in 007 at 1.0 ng and Eurasiaplex was under-represented in 9947A at 1.5 ng. Seven of the nine amplicons with the highest 10% coverage originated from Eurasiaplex with a combined coverage range from 16442× to 25000×; indicating Eurasiaplex amplicons were generally over-represented (Table 2). A Kruskal-Wallis test rejected the hypothesis of no difference in combined coverage (across all five template amounts for samples 9947A, 007 and S1) between amplicons from different multiplex assays (p = 0.000) indicating significant differences in SNP coverage from different assays in all samples.

3.9 Negative control

Eleven SNPs were observed in the negative control with nine of these from the 52-plex assay (SI Table 8). The coverage ranged from 69× to 5000×. Except for five SNPs (rs1355366, rs1463729, rs1028528, rs734482 and rs2227203), all other genotypes corresponded to one or more possible templates used in the run (9947A, 007, S1 and S2). Daniel et al. [10] also observed coverage of SNPs in a negative control with the same three samples (9947A, 007 and S1) sequenced on the Ion PGM™.

3.10 Cost estimate

Genotyping costs were estimated to be 1.4 US$ per SNP per sample based on library preparation and sequencing reagent costs only (SI Table 9). One Miseq v3 flow cell has the capacity to genotype ~10,000 SNPs per sample at 100× coverage when genotyping 24 samples in a run. Thus, per SNP costs could be further reduced by adding more markers and/or more samples per Miseq run.

4 Discussion

This study demonstrates the potential of the MiSeq as a medium throughput MPS platform for forensic analysis using modular, customised SNP panels that are already established as sensitive forensic assays. The 3 Gb of sequence data obtained from two runs allowed us to obtain SNP genotypes for identity, BGA and EVC inferences from 48 samples.

MiSeq sequencing using a pool of five non-commercial multiplex PCR assays produced uneven sequence coverage which was also observed for the same amplicons by the Ion PGM™ system [10]. The overall sequence data profile has the following characteristics: (a) uneven coverage of amplicons between multiplex PCR assays; (b) uneven coverage within each assay; and (c) non-normal (skewed) distribution of coverage (SI Figure 5). High and low coverage is consistent between amplicons, suggesting that coverage bias is not random but amplicon-dependent. Additionally, coverage was weakly associated with amplicon length (SI Figure 3), possibly due to sequence length bias in the magnetic bead clean up steps of library preparation, favouring longer amplicons.

GC content is often implicated as a source of coverage bias in MPS and associated library preparation [31]. Several studies have identified GC bias in MiSeq sequencing [31-33]; however, we did not encounter such bias (SI Figure 4). This may be because the amplicons sequenced in this study were from optimised SNaPshot® assays where optimal GC content of both primer and amplicon sequence had been important considerations during primer design. Also, there were no extremes of GC content in our amplicons (28-65%). This result matches a lack of detectable GC bias in previous Ion PGM™ sequencing of the same templates [10].

The most likely reason for the observed amplicon coverage bias is the amplification imbalance between and within each of the five multiplex assays. The bias may have arisen from differing amplicon representation between multiplexes prior to library preparation, with Eurasiaplex clearly showing over-representation (SI Figure 5). SNaPshot® PCR assay protocols were applied without modification for amplicon generation and were not optimised for MPS. Furthermore, PCR products from the five assays were pooled in equal volumes, whereas equimolar pooling may have reduced the imbalance between PCR assays. Any bias within assays may be addressed by further fine-tuning of primer concentrations. Nevertheless, the equal volume pooling strategy used resulted in high genotype concordance with SNaPshot® for both the MiSeq and Ion Torrent [10] suggesting that this approach can be utilised effectively without investing resources in amplicon quantitation or multiplex optimisation to achieve balanced amplicon production.

Baseline coverage thresholds were not applied as discounting genotypes with GQx < 99 resulted in filtering out most of the genotypes with less than 20× coverage. Genotype non-concordance did not appear to be related to low coverage (S1 Table 5) and contaminating alleles in the negative control were similarly not related to low coverage (S1 Table 8: only one of 11 contaminating alleles with less than 250× coverage). The genotypes of six of eleven SNPs observed in the negative controls corresponded to those of one or more templates that could be attributed to contamination between samples. However, the unmatched genotypes of the other five SNPs must either have been due to external DNA or PCR/sequencing errors. In this study, the single negative control was a pool of the negative controls from individual multiplex PCR assays and sequencing those individual negative controls would have been more informative. Thus, while this approach is not sufficient to evaluate baseline coverage thresholds, it is an informative preliminary study of negative controls on this platform for any customised forensic assay designs.

This customised MiSeq approach was sensitive enough to provide reliable genotypes with a total template amount as low as 0.25 ng (0.05 ng for each multiplex assay) using default allele frequency thresholds, yielding an average overall concordance of 98% with SNaPshot and Ion Torrent data (for 9947A, 007 and S1). The sensitivity study was performed on the same three templates with the same five PCR assays employed by a similar Ion PGM study [10] as a means of comparison and is indicative only. Daniel *et al*. [10] similarly found > 98% genotype concordance down to 0.1 ng template amount per assay (0.5 ng total). Greater resolution would require the use of replicates for each dilution.

The reproducibility study indicated a high genotype concordance between the four replicates (100 % concordant, excluding missing SNPs). This occurred in spite of a significant difference in coverage distribution between the replicates, the probable result of highly uneven amplicon coverage. Genotyping of rs1592672 consistently failed in all replicates and the primers for this SNP may require redesigning if this customised assay was to be routinely employed. One of the replicates failed to produce genotypes for a further three SNPs (rs1357617, rs188650 and rs938283) in a total of 136 unique SNPs.

Genotype concordance between MiSeq and SNaPshot for 9947A, 007 and S1 to S7 ranged from 97.8 to 100 %. Applying stringent allele frequency thresholds (such as 95% for homozygotes and 40-60% for heterozygotes) may increase the already high concordance by eliminating some of the ambiguous allele

frequencies shown in Fig. 1 (in the range 10-30% and 70-90%). However, this may reduce the number of usable reads and genotypes [10].

Some of the MiSeq SNP genotypes which were non-concordant with SNaPshot® were concordant with Sanger genotypes (from data in Daniel *et al.* [10]). Two SNPs (rs1029407 and rs717302) were non-concordant across all three platforms (MiSeq, SNaPshot and Sanger) likely due to homopolymeric stretches in flanking regions prompting misalignments. MPS sequencing is known to perform poorly in homopolymeric regions [31]. SNP rs1029407 has been mistyped by the Ion PGM™ [12] as well as by the MiSeq in this study (SI Table 5), whereas the GAIIx (Illumina) MPS platform has produced a correct AA genotype for 9947A in another study [24]. The MiSeq omitted a single base in the flanking homopolymer region which the alignment algorithm then mis-aligned (SI Figure 6), whereas the GAIIx alignment software could align the sequences properly and call the correct genotype [24]. This provides further evidence that SNP mistyping in markers sited in homopolymeric regions can occur from misalignment as well as from mis-incorporation of nucleotides.

MiSeq genotyping was not affected by the methods used to extract DNA and was able to produce partial SNP profiles from samples exposed to 60 minutes of UV radiation and 100 ng of humic acid, whereas SNaPshot® SNP genotyping and standard STR profiling failed to detect any alleles in these samples. In addition, this approach successfully typed degraded casework (compromised) samples, producing genotypes for up to 92 % of SNPs for aged blood and bones, when real time PCR quantitation using Quantifiler failed to detect DNA in most cases (< 0.01 ng/μL in all cases). This demonstrates the robustness and applicability of MiSeq using customised SNP panels for highly degraded and inhibited sample analysis typical of disaster victim identification (DVI) and exhumed remains.

This customised approach offers modularity and flexibility to add and subtract SNP panels providing better ancestry resolution (to sub-population level) and EVC inclusion in contrast to the commercially-available ForenSeq™ DNA Signature Kit (Illumina) consisting of only 56 ancestry informative SNPs which limits ancestry resolution to three or four continental populations only. MPS is a potential DNA-based intelligence tool that can type a large battery of forensically informative markers with consequent reduction in inter-run variability, cost, time and effort.

5. Concluding remarks

MiSeq MPS sequencing employing customised, modular SNP panels has been demonstrated here to be able to genotype over one hundred identity, BGA and EVC markers simultaneously in multiple samples. This offers the potential to maximise the use of scarce evidentiary material in comparison to the standard SNaPshot® genotyping. In addition, our customised method provides the option of adding optimised marker sets to increase the resolution and accuracy of ancestry and phenotype prediction in a single run. Future work should be conducted to evaluate the baseline coverage thresholds which may increase genotyping accuracy.

6 References

[1] Butler, K., Peck, M., Hart, J., Schanfield, M., Podini, D., *Forensic Science International: Genetics Supplement Series* 2011, *3*, e498-e499.

[2] Kayser, M., de Knijff, P, *Nature Reviews* 2011, *12*, 179-192.

[3] Sobrino, B., Brion, M., Carracedo, A., *Forensic Science International* 2005, *154*, 181-194.

[4] Pati, N., Schowinsky, V., Kokanovic, O., Magnuson, V., Ghosh, S., *Journal of Biochemical and Biophysical Methods* 2004, *60*, 1-12.

[5] Phillips, C., Salas, A., Sánchez, J. J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., Calaza, M., de Cal, M. C., Ballard, D., Lareu, M. V., Carracedo, Á., *Forensic Science International: Genetics* 2007, *1*, 273-280.

[6] Fondevila, M., Phillips, C., Santos, C., Freire Aradas, A., Vallone, P. M., Butler, J. M., Lareu, M. V., Carracedo, Á., *Forensic Science International: Genetics* 2013, *7*, 63-74.

[7] Phillips, C., Aradas, A. F., Kriegel, A. K., Fondevila, M., Bulbul, O., Santos, C., Rech, F. S., Carceles, M. D. P., Carracedo, Á., Schneider, P. M., Lareu, M. V., *Forensic Science International: Genetics* 2013, *7*, 359-366.

[8] Walsh, S., Lindenbergh, A., Zuniga, S. B., Sijenb, T., Knijff, P. d., Kayser, M., Ballantyne, K. N., *Forensic Science International: Genetics* 2011, *5*, 464-471.

[9] Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W., Kayser, M., *Forensic Science International: Genetics* 2013, *7*, 98-115.

[10] Daniel, R., Santos, C., Phillips, C., Fondevila, M., van Oorschot, R. A. H., Carracedo, Á., Lareu, M. V., McNevin, D., *Forensic Science International: Genetics* 2015, *14*, 50-60.

[11] Mehta, B., Daniel, R., McNevin, D., *Forensic Science International: Genetics Supplement Series* 2013, *4*, e376-e377.

[12] Berglund, E. C., Kiialainen, A., Syvänen, A.-C., *Investigative Genetics* 2011, *2*, 23-23.

[13] Børsting, C., Fordyce, S. L., Olofsson, J., Mogensen, H. S., Morling, N., *Forensic Science International: Genetics* 2014, *12*, 144-154.

[14] Grada, A., Weinbrecht, K., *J Invest Dermatol* 2013, *133*, e11.

[15] Illumina, Illumina Corporation, San Diego, CA, 2010.

[16] Churchill, J. D., Schmedes, S. E., King, J. L., Budowle, B., *Forensic Science International: Genetics* 2016, *20*, 20-29.

[17] Sanchez, J. J., Phillips, C., Børsting, C., Balogh, K., Bogus, M., Fondevila, M., Harrison, C. D., Musgrave-Brown, E., Salas, A., Syndercombe-Court, D., Schneider, P. M., Carracedo, A., Morling, N., *Electrophoresis* 2006, *27*, 1713-1724.

[18] Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R. A., Burchard, E. G., Schanfield, M. S., Souto, L., Uacyisrael, J., Via, M., *Forensic Science International: Genetics* 2016, *20*, 71-80.

[19] Köchl, S., Niederstätter, H., Parson, W., *Forensic DNA Typing Protocols*, Springer 2005, pp. 13-29.

[20] Pang, B. C. M., Cheung, B. K. K., *Analytical Biochemistry* 2007, *360*, 163-165.

[21] Illumina, MiSeq System User Guide, Rev O.

[22] Illumina, Preparing sequencing libraries for Loading on a MiSeq, Rev B.

[23] Illumina, TruSeq CHIP sample preperation guide, Rev A.

[24] GenomeReferenceConsortium, Human Genome Overview: GRCh37, National Institutes of Health.

[25] Li, H., Durbin, R., *Bioinformatics* 2010, *26*, 589-595.

[26] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *Genome research* 2010, *20*, 1297-1303.

[27] Walsh, S., Liu, F., Ballantyne, K. N., Oven, M. v., Lao, O., Kayser, M., *Forensic Science International: Genetics* 2011, *5*, 170-180.

[28] Walsh, S., Chaitanya, L., Clarisse, L., Wirken, L., Draus-Barini, J., Kovatsi, L., Maeda, H., Ishikawa, T., Sijen, T., de Knijff, P., Branicki, W., Liu, F., Kayser, M., *Forensic Science International: Genetics* 2014, *9*, 150-161.

[29] Westen, A. A., Matai, A. S., Laros, J. F., Meiland, H. C., Jasper, M., de Leeuw, W. J., de Knijff, P., Sijen, T., *Forensic Science International: Genetics* 2009, *3*, 233-241.

[30] Phillips, C., Amigo, J., Carracedo, Á., Lareu, M., *Forensic Science International: Genetics* 2015, *19*, 100-106.

[31] Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., Jaffe, D. B., *Genome Biol* 2013, *14*, R51.

[32] Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., Gu, Y., *BMC genomics* 2012, *13*, 341.

[33] Chen, S., Li, S., Xie, W., Li, X., Zhang, C., Jiang, H., Zheng, J., Pan, X., Zheng, H., Liu, J. S., *PloS one* 2014, *9*.

Acknowledgments

The authors gratefully acknowledge technical support and consumables from Illumina Inc. and financial support from the Australian Research Council (LP110100121 - *From genotype to phenotype: Molecular photofitting for criminal investigations*). SRD was supported by an Illumina MiSeq Grant. The compromised forensic samples were provided by Kirsty Wright (School of Natural Sciences, Griffith University), Jodie Ward (Forensic & Analytical Science Services, NSW) and David Bruce (Forensic & Analytical Science Services, NSW). We also acknowledge the Unrecovered War Causalities-Army (UWC-A) unit and Jeremy Austin (Australian Centre for Ancient DNA, Adelaide) for their support with aged bone sample recovery.

Conflict of interest

The authors have declared no conflict of interest.

Figures

Figure 1 The effects of UV irradiation (30 and 60 minutes) and humic acid (HA, at 75 ng and 100 ng) on genotype concordance (as a percentage of a total of 136 unique SNPs) for samples S2 and S3.

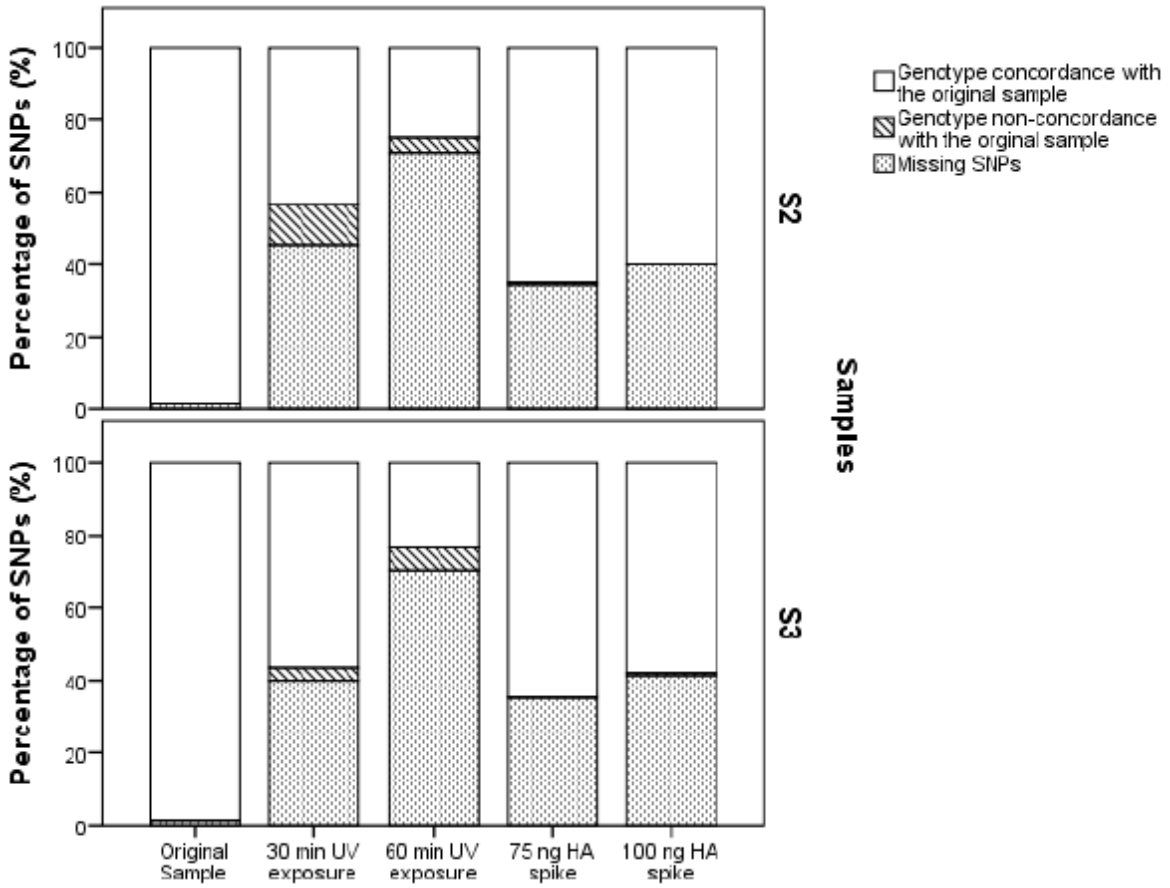
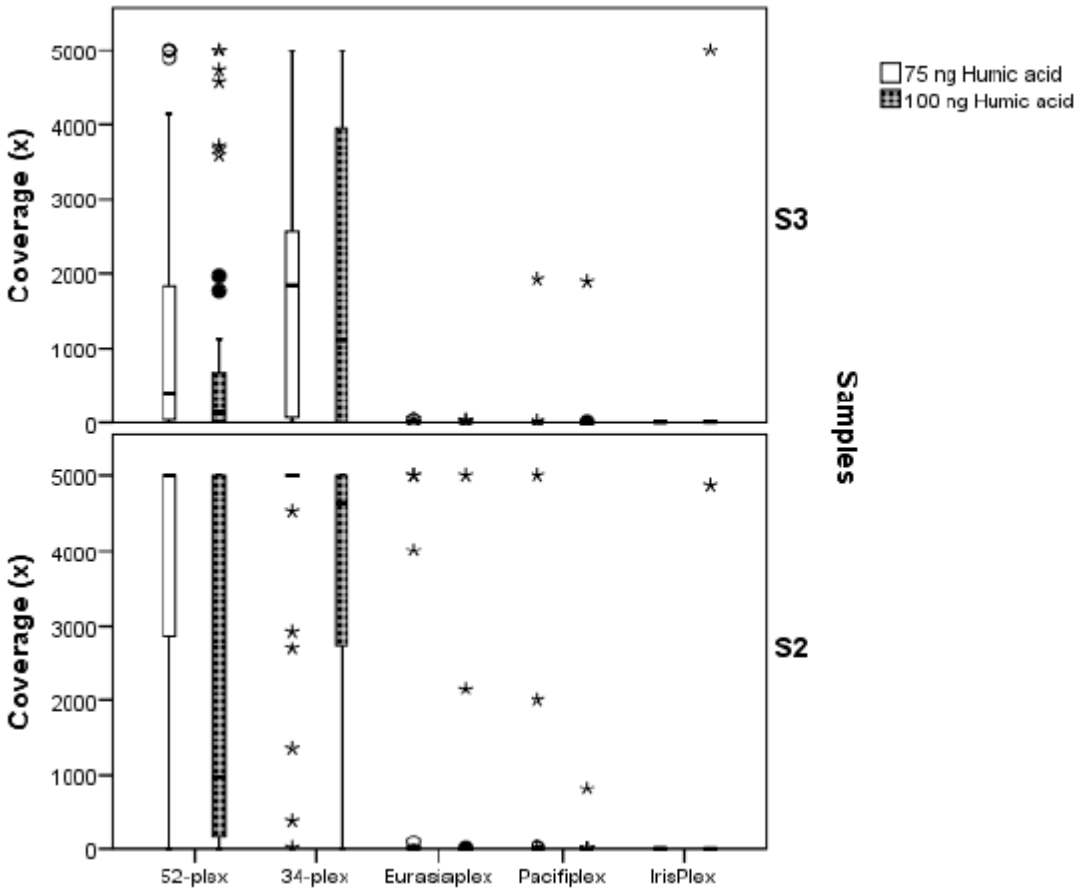


Figure 2 Coverage distributions in each of the five multiplex PCR assays spiked with Humic acid (HA) at 75 ng and 100 ng for samples S2 and S3. The 52-plex and 34-plex assays both contain bovine serum albumin (BSA) in their PCR reaction mix.



Tables

Table 1 Reportable SNPs (from a total of 93) for the compromised samples.

Sample	Substrate	Quantity (ng/µL)	Percentage of reportable SNPs (%)
S8	bone	Undetected	4
S9	bone	0.006	89
S10	bone	Undetected	11
S11	bone	Undetected	13
S12	bone	0.004	92
S13	blood	Undetected	86

Table 2 SNPs with lowest and highest 10% of combined coverage across all template amounts for 9947A, 007 and S1.

SNP	GC content (%)	Amplicon length (bp)	Multiplex	9947A ×	007 ×	S1 ×
Lowest 10% coverage						
rs3785181	47	156	34-plex	82	126	314
rs2069945	43	83	Pacifiplex	47	137	876
rs1357617	44	90	52 Auto 1	69	111	970

rs12434466	33	51	Pacifiplex	168	155	998
rs239031	47	70	34-plex	222	384	1124
rs876724	40	83	52 Auto 1	107	108	1401
rs2046361	28	79	52 Auto 1	212	92	3025
rs2274636	46	81	Pacifiplex	253	446	1652
rs826472	31	85	52 Auto 1	269	420	4483
Median	43	83				
Highest 10% coverage						
rs9809818	45	89	Pacifiplex	16416	18431	25000
rs39897	50	78	Eurasiaplex	16442	21083	25000
rs1544656	46	90	Eurasiaplex	17418	24999	25000
rs1519654	51	86	Eurasiaplex	20004	25000	25000
rs10008492	47	94	Eurasiaplex	20008	24999	25000
rs2196051	34	115	Eurasiaplex	20013	25000	25000
rs734482	38	118	Eurasiaplex	20022	25000	25000
rs17625895	41	99	Eurasiaplex	20023	24999	25000
rs354439	37	93	52 Auto 2	20264	18432	25000
Median	45	93				