Int J Legal Med (2015) 129:31–36
DOI 10.1007/s00414-014-1108-8

ORIGINAL ARTICLE

Massively parallel sequencing of forensically relevant single nucleotide polymorphisms using TruSeq™ forensic amplicon

David H. Warshauer · Carey P. Davis · Cydne Holt · Yonmee Han · Paulina Walichiewicz · Tom Richardson · Kathryn Stephens · Anne Jager · Jonathan King · Bruce Budowle

Received: 3 September 2014 / Accepted: 30 October 2014 / Published online: 19 November 2014

Abstract The TruSeq™ Forensic Amplicon library preparation protocol, originally designed to attach sequencing adapters to chromatin-bound DNA for chromatin immunoprecipitation sequencing (TruSeq™ ChIP-Seq), was used here to attach adapters directly to amplicons containing markers of forensic interest. In this study, the TruSeq™ Forensic Amplicon library preparation protocol was used to detect 160 single nucleotide polymorphisms (SNPs), including human identification SNPs (iSNPs), ancestry, and phenotypic SNPs (apSNPs) in 12 reference samples. Results were compared with those generated by a second laboratory using the same technique, as well as to those generated by whole genome sequencing (WGS). The genotype calls made using the TruSeq™ Forensic Amplicon library preparation protocol were highly concordant. The protocol described herein represents an effective and relatively sensitive means of preparing amplified nuclear DNA for massively parallel sequencing (MPS).

Keywords TruSeq custom amplicon · Massively parallel sequencing · Ancestry informative markers · Phenotypic SNPs

Electronic supplementary material The online version of this article (doi:10.1007/s00414-014-1108-8) contains supplementary material, which is available to authorized users.

D. H. Warshauer · J. King · B. Budowle
Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Boulevard, Fort Worth, TX 76107, USA

C. P. Davis · C. Holt · Y. Han · P. Walichiewicz · T. Richardson · K. Stephens · A. Jager
Illumina, Inc., 5200 Illumina Way, San Diego, CA 92122, USA

B. Budowle (✉)
Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: bbudowle@hsc.unt.edu

Introduction

Forensic DNA analysis is an extremely valuable tool for human identity testing in a number of situations, including criminal cases, mass disaster scenarios, and instances involving missing persons. Given their high power of discrimination and relatively small amplicon size, short tandem repeats (STRs) usually are the markers of choice for analysis of forensic biological evidence. However, there are a number of situations in which single nucleotide polymorphism (SNP) typing may provide an adjunct, alternative, or better option. The overall amplicon size of SNPs can be designed to be shorter than that of even a mini-STR marker [1] while retaining a level of discriminatory power that is comparable to STRs [2], assuming that a sufficient number of SNPs are typed. This quality makes SNP analysis a powerful tool in situations where, for example, evidentiary DNA is highly degraded.

To date, a variety of typing methodologies have been utilized for the analysis of SNP markers. These approaches include, but are not limited to, single base extension, allele-specific hybridization assays, chip-based microarrays, and mass spectrometry [3–10]. While each of these methods has its merits, they have inherent limitations. The most significant drawbacks are high-input DNA requirements, lack of quantitation, low throughput, high cost, inability to type large numbers of SNPs in a single analysis, and/or limitations requiring typing STRs and SNPs in separate reactions and separate runs.

Recently, massively parallel sequencing (MPS) has been shown to be a promising method for the detection of forensic SNP markers [2]. The number of SNPs that can be detected in a single analysis with MPS is far greater than with the aforementioned methods, and the throughput of MPS is unparalleled. Moreover, up to 384 different samples currently can be typed simultaneously [11]. In addition, advances in sequencing technology have lowered both the cost and time required

for analysis to a point that makes MPS cost-effective and competitive with other typing technologies.

Chromatin immunoprecipitation sequencing (ChIP-Seq) is a technology in which genomic DNA is cross-linked with chromatin and enriched before being subjected to MPS [12]. Traditionally, it has been used to investigate the distribution, abundance, and characteristics of DNA-bound protein targets across a genome of interest. The TruSeq™ ChIP sample preparation kit (Illumina, Inc., San Diego, CA) provides a simple workflow that allows preparation of chromatin-bound DNA for sequencing via the attachment of TruSeq™ adapters.

In this study, the TruSeq™ ChIP protocol was modified to enable library preparation of forensically relevant SNP-containing amplicons. This modified protocol, known as TruSeq™ Forensic Amplicon, was used to detect a battery of 160 human identification SNPs (iSNPs), ancestry, and phenotypic SNPs (apSNPs) in a set of 12 reference samples. The resulting data were analyzed for both sequence coverage and heterozygote allele balance. Results presented here illustrate the efficacy of this method.

Materials and methods

Nuclear DNA amplicons containing iSNPs and apSNPs were subjected to the TruSeq™ Forensic Amplicon protocol and subsequently sequenced on the MiSeq™ platform. Following the University of North Texas Health Science Center Institutional Review Board approval, quantitated human DNA control samples from 12 unrelated individuals (obtained from Coriell Institute for Medical Research, Camden, NJ) were used for this proof-of-concept study.

Normalization, primer design, and amplification

The 12 DNA control samples were normalized to 1 ng/μL. The normalized samples were verified to be 1 ng/μL using the Quantifiler® Human DNA Quantification Kit on the ABI 7900HT Fast Real-Time PCR System (ThermoFisher, Carlsbad, CA) following the manufacturer's recommendations.

PCR primers were designed manually using OligoAnalyzer 3.1 (Integrated DNA Technologies (IDT), Coralville, IA), Primer3, and UCSC Genome Browser [13, 14]. Two sets of desalted primer (IDT) pools were created by adding each locus-specific primer (forward and reverse) into a multiplex set. A pool of 94 iSNPs and a separate pool of aSNPs and pSNPs, totaling 56 and 10, respectively, were created (Supplemental Table 1). For the iSNP master mix, 12.5 μL of 2× Qiagen Multiplex PCR Master Mix (Qiagen Inc., Valencia, CA), 2.4 μL of the iSNP primer mix, 10.1 μL of laboratory grade water, and 1 μL of the respective normalized sample were added to each well of a 96-well plate. For the apSNP master mix, 12.5 μL of 2× Qiagen Multiplex PCR Master Mix, 1.65 μL of the apSNP primer mix, 10.85 μL of laboratory grade water, and 1 μL of the respective normalized sample were added to each well of a 96-well plate.

The samples then were amplified using a Bio-Rad Tetrad 2 thermal cycler (Bio-Rad Laboratories, Inc., Hercules, CA) with the following PCR parameters: 95 °C for 11 min, 96 °C for 1 min, 35 cycles of 94 °C for 30 s, 58 °C for 30 s with a 0.5 °C/s ramp rate, 68 °C for 45 s with a 0.2 °C/s ramp rate, then 60 °C for 30 min and a hold at 10 °C.

Library preparation

The TruSeq™ Forensic Amplicon library preparation protocol recommends an amplified DNA input volume of 50 μL, at a concentration of 20–2000 pg/μL (i.e., 1–100 ng total input DNA). Following these guidelines, the amplified products generated from each PCR were normalized at 0.5 ng/μL at a volume of 50 μL in a 96-well plate, for a total of 24 wells each containing 25 ng of amplified DNA. A second laboratory (at Illumina) used 1 μL of 1 ng/μL amplicons instead.

The TruSeq™ Forensic Amplicon library preparation process is similar to that of TruSeq™ ChIP, except that it uses PCR amplicons as starting material rather than chromatin-bound DNA. The process began with end repair, where the 5′ ends of the amplicons were made blunt and phosphorylated during a 30-min incubation at 30 °C in an Applied Biosystems® GeneAmp® PCR System 9700 thermal cycler (Life Technologies). All subsequent incubation and amplification processes were carried out on this thermal cycler platform. Next, the samples were washed using AMPure XP beads and 80 % ethanol. The blunt ends then were adenylated, which prevented them from ligating to each other during adapter ligation. Adenylation was performed by thermal cycling using the following parameters: 37 °C for 30 min, 70 °C for 5 min, and a final hold at 4 °C. Following adenylation, adapter ligation was performed, wherein TruSeq™ indexed adapters were bound to the adenylated 3′ ends of the amplicons. Each sample was bound to adapters with a unique index sequence for multiplexed sequencing. Adapter ligation required a 10-min incubation at 30 °C, followed by washing using AMPure XP beads and 80 % ethanol. For enrichment of adapter-bound amplicons, PCR was carried out using primers designed to amplify only those amplicons with adapters bound to them. The enrichment PCR parameters were: 98 °C for 30 s, 18 cycles of 98 °C for 10 s, 60 °C for 30 s, and 72 °C for 30 s, a final extension at 72 °C for 5 min, and a final hold at 4 °C. Enrichment PCR was followed by washing with AMPure XP beads and 80 % ethanol.

Following library preparation, the adapter-ligated amplicons were quantified using the Qubit® platform (Life Technologies), according to the manufacturer's protocol. Based on the quantification results, the samples were normalized to a concentration of 10 nM with 10 mM Tris–HCl buffer

at pH 8.5 with 0.1 % Tween 20, as per Illumina® guidelines. A total of 5 μL of each sample were used to pool samples together for a total 10 nM sample pool of 120 μL.

MiSeq™ sequencing and data analysis

To prepare for sequencing on the MiSeq™ (Illumina), 10 μL of the 10 nM sample pool were combined with 40 μL of 10 mM Tris–HCl buffer at pH 8.5 with 0.1 % Tween 20, for a resultant concentration of 2 nM. Illumina®'s library preparation guidelines for the MiSeq™ were followed, and the concentration of the pooled sample was brought down to 12 pM using chilled HT1 buffer. Paired-end sequencing was performed, with a read length of 120 bases.

The sequencing sample sheet for these samples was created using the Illumina Experiment Manager. For this modified protocol, the "TruSeq™ Amplicon" workflow was used, and the samples were treated as custom amplicons. Once the sample sheet was created, it was edited by changing the index sequences used in the "TruSeq™ Amplicon" workflow to those used in the TruSeq™ Forensic Amplicon protocol. A custom manifest file was used for the sequencing run to define the position and names of each of the SNPs of interest. Using this manifest, MiSeq Reporter was able to produce *vcf* files for each sample which identified each SNP detected during sequencing.

Since MiSeq Reporter limits sequence coverage values for SNPs to 5000× by default, a separate method of variant-calling was required to ascertain the actual coverage at each locus of interest so that conclusions could be drawn with regard to the depth of sequencing and heterozygote balance afforded by the TruSeq™ Forensic Amplicon library preparation method. To this end, *bam* files were subjected to variant-calling without downsampling using the GATK [15]. Heterozygote balance was calculated by dividing the lower allele coverage value at each heterozygous SNP locus by the higher coverage value, yielding a heterozygote balance percentage.

Results

Through the use of the TruSeq™ Forensic Amplicon library preparation protocol, SNP genotypes were generated for all 160 targeted iSNPs and apSNPs in 11 of the 12 samples analyzed. In sample 9, rs10776839 was not called due to low coverage (this particular SNP displayed low sequencing read depth across all samples). The amplified products of 11 of the samples tested in this study also were analyzed by a separate laboratory at Illumina. The SNP genotypes yielded were highly concordant between the two laboratories. Of the 11 samples compared, 7 were 100 % concordant at all 160 SNPs. The concordance between the four remaining samples was between 98.75 and 99.38 %. Discordant genotype calls can be found in Supplemental Table 2. It should be noted that discordant genotype calls between these two datasets were mainly due to differences in the calling of heterozygous versus homozygous genotypes, based on the heterozygosity thresholds used during analysis. For example, at rs2399332 in sample 2, the "T" allele displayed a coverage value of 312 reads, while the "G" allele had a coverage value of 3474 reads, which equates to a heterozygosity balance of approximately 9 % (Supplemental Table 2). The in-house heterozygosity threshold was set at 5 %, while a 10 % threshold was used by the second laboratory. Thus, this SNP was called in-house as a heterozygote, while the second laboratory determined that the SNP was homozygous for the "G" allele. Such results are, in effect, concordant. Indeed, if the in-house results are interpreted using a 10 % heterozygosity threshold, the concordance values rise to 100 % in 10 out of the 11 samples compared. The recalculated concordance value for the remaining sample (sample 10) would be 99.38 %, due to a single discordance at rs2399332, where the in-house heterozygosity value was 12.9 %, and thus only slightly above the 10 % threshold. While the in-house heterozygosity threshold value was chosen arbitrarily for this study to simply demonstrate proof of concept, this occurrence highlights the need for reliable thresholds developed through proper validation in each testing laboratory. Overall, the results are similar across all SNPs and differ only due to thresholds and variation of the lower signal SNP. Primer redesign for these loci may improve allele imbalance.

Whole genome sequencing (WGS)-based SNP calls were obtained from the Complete Genomics FTP site [16] for additional concordance testing. The allele calls derived from the in-house data produced by the TruSeq™ Forensic Amplicon library preparation method displayed a high concordance (96.23 to 98.74 %) with the WGS data across all 12 samples. Discordance between the WGS-derived SNP calls and the in-house calls was found at a total of nine out of the 160 SNPs (rs1029047, rs1058083, rs10776839, rs10954737, rs12997453, rs182549, rs2399332, rs430046, and rs907100). It should be noted that the discordances between the in-house calls and the Illumina calls listed above were corroborated by the WGS data, consistent with the calls made by the second laboratory. This agrees with the explanation that these discordances were simply the result of threshold differences. The remaining discordant SNP loci between the WGS and in-house data appear to be discordance "hotspots" for this particular multiplex design, as all but one of the loci showed discordance in at least four of the samples tested (Table 1). Phase 3 data from the 1000 Genomes Project were available for samples 1 and 12, and a comparison showed that the phase 3 genotype calls for these samples were consistent with the WGS calls. The vast majority of the discordance (all but 3 of the total 53 discordant calls, across all samples) consisted of

Table 1 SNP discordance (in-house versus WGS)

	1	2	3	4	5	6	7	8	9	10	11	12
rs1029047	A/T : A	A/T : A		A/T : A		A/T : A				A/T : A	A/T : A	
rs1058083					A/G : G		A/G : G	A/G : G		A/G : G		
rs10776839	G : G/T			G : G/T	G : G/T		G : G/T			G : G/T	G : G/T	G : T
rs10954737	T : C/T	T : C/T	T : C/T					T : C/T	T : C/T	T : C/T		
rs12997453		G : A/G										
rs182549	C/T : T	C/T : T	C/T : T	C/T : T					C/T : T	C/T : T		
rs2399332	G/T : G	G/T : G					G/T : G	G/T : G		G/T : G	G/T : G	
rs430046	C : C/T	C : C/T		C : C/T	C : C/T	C : T	C : C/T	C : C/T	C : C/T		C : C/T	C : T
rs907100						G : C/G	G : C/G	G : C/G	G : C/G	G : C/G		G : C/G

Discordance between the SNP calls generated in this study and those obtained through whole genome sequencing are listed. Discordance is shown in the following format: "in-house call: WGS call"

differences between heterozygous and homozygous allele SNP calls, which can once again be explained by differences in heterozygosity thresholds. However, a nucleotide variation within the primer binding site may have resulted in a failure to amplify one of the alleles at a given locus. Other explanations include factors such as multiplex inefficiency, low coverage leading to skewed SNP calls, and simple alignment errors.

Overall, the heterozygote balance achieved through the use of the TruSeq™ Forensic Amplicon library preparation method was quite even. Across all samples, between 91.9 and 100 % of the heterozygous loci showed allelic balance ratios of 1:2 (50 % balance) or better. An example of heterozygous allele balance is shown in Fig. 1. The heterozygous loci for which allelic balance ratios dropped below 1:2 are shown in Supplemental Table 3. In some cases, allelic imbalance was explained by low coverage (e.g., rs1029047 in sample 2, which had a relatively low coverage of 281 reads and displayed a heterozygosity balance value of 12.6 %), but other factors such as those noted above may explain imbalance in heterozygous loci with higher coverage values.

The average sequencing coverage per locus across all 12 samples with both panels (i.e., effectively 24 samples) ranged from 142× to 46,908×, and coverage was relatively consistent between samples at each locus. Figure 2 illustrates the sequence coverage across the apSNP loci, as an example. The wide range of coverage is most likely due to differences in amplification efficiency of the multiplex PCR. Further optimization is underway to reduce the coverage range.

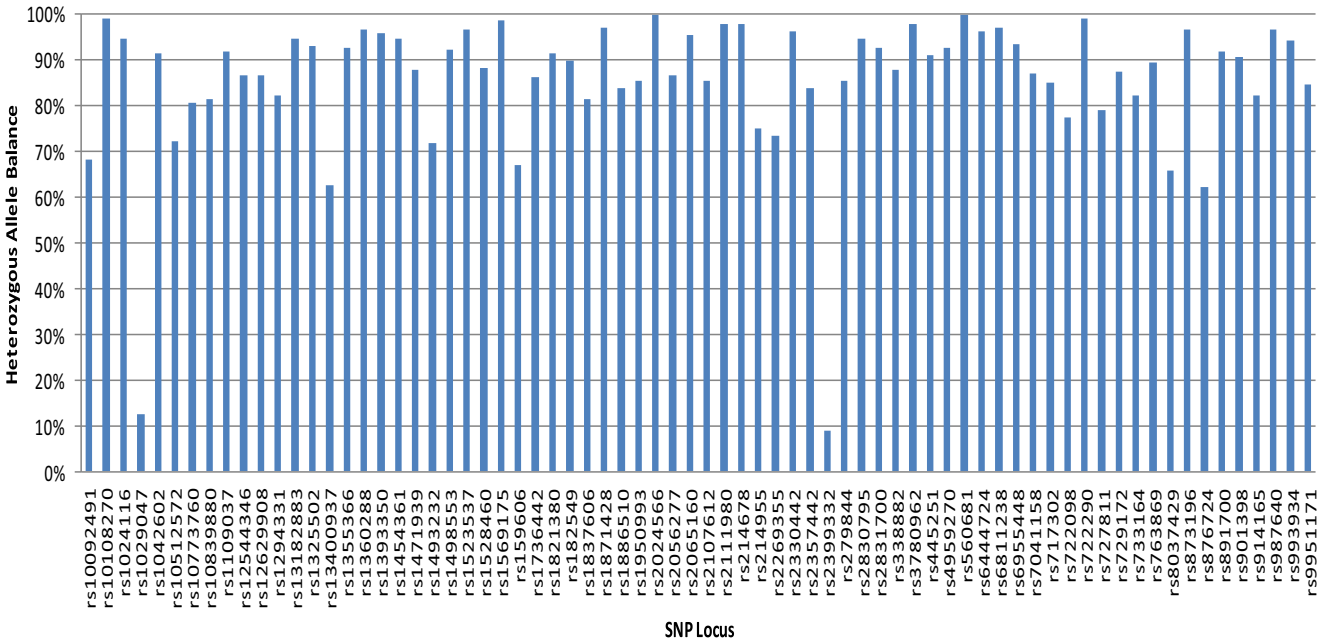

Fig. 1 Heterozygous allele balance for sample 2. Allele balance at heterozygous loci, expressed as a percentage, is shown. A value of 100 % denotes a perfect 1:1 balance of alleles. In this sample, only 2 loci (rs1029047 and rs2399332) display an allele balance value of less than 50 %

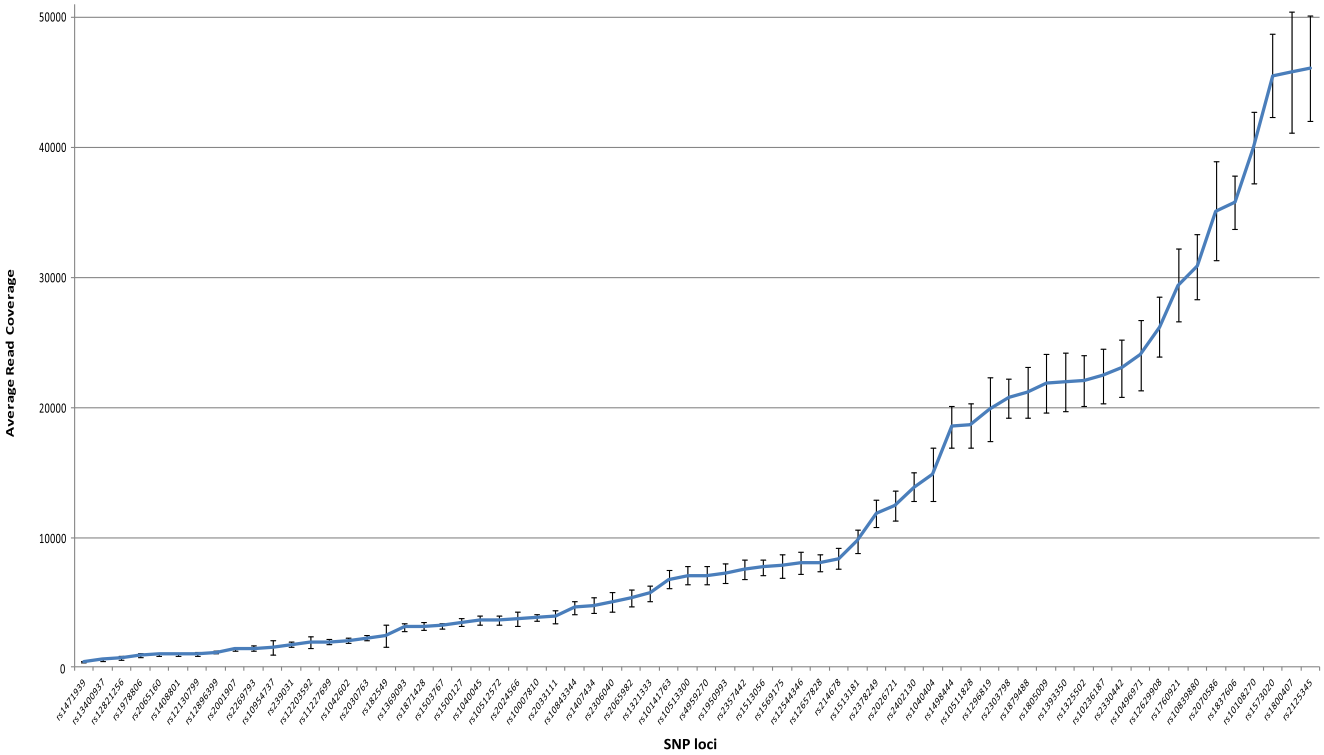

Fig. 2 Average sequence coverage for apSNP loci. The average depth of coverage across all samples for each apSNP locus is shown here. *Bars* represent the standard deviation

Conclusions

The results of this proof-of-concept study indicate that the TruSeq™ Forensic Amplicon library preparation protocol is an effective method of preparing amplified nuclear DNA for massively parallel sequencing. This method is less labor-intensive than alternative techniques. Unlike the TruSeq™ Custom Amplicon workflow, TruSeq™ Forensic Amplicon does not require the use of custom-designed oligonucleotide probes for library preparation. Additionally, the TruSeq™ Forensic Amplicon library preparation method is highly sensitive, with a relatively low input DNA requirement (1 ng of input DNA was amplified and 25 ng of amplified DNA were used for each sample, and at the second laboratory, 1 μL of 1 ng/μL amplicons was used, as opposed to the recommended 500 ng of input DNA recommended for the TruSeq™ Enrichment protocol). This lower DNA input is more suited for the quantities of sample DNA often encountered in forensic casework. In conjunction with a properly designed multiplex PCR, this preparation method is capable of producing reliable sequencing results with relatively even allele balance at heterozygous loci. Though not tested in this study, it is likely that the TruSeq™ Forensic Amplicon kit could be used for the preparation and detection of STR markers. The results of this proof-of-concept study suggest that this novel use of the original TruSeq™ ChIP protocol could support forensic genetic typing by MPS.

Acknowledgments This work was supported in part by award no. 2012-DN-BXK033, awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the US Department of Justice. The authors also would like to thank Illumina, Inc. for its support during this study.

Conflict of interest C.P. Davis, C. Holt, Y. Han, P. Walichiewicz, T. Richardson, K. Stephens, and A. Jager are employed by Illumina, Inc.

References

1. Dixon LA, Murry CM, Archer EJ, Dobbins AE, Koumi P, Gill P (2005) Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. Forensic Sci Int 154:62–77
2. Seo SB, King JL, Warshauer DH, Davis CP, Ge J, Budowle B (2013) Single nucleotide polymorphism typing with massively parallel sequencing for human identification. Int J Legal Med 127:1079–1086
3. Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N (2009) A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis 27:1713–1724
4. Pakstis AJ, Speed WC, Fang R, Hyland FC, Furtado MR, Kidd JR, Kidd KK (2010) SNPs for a universal individual identification panel. Hum Genet 127:315–324
5. Tomas C, Axler-DiPerte G, Budimlija ZM, Børsting C, Coble MD, Decker AE, Eisenberg A, Fang R, Fondevila M, Fredslund SF, Gonzalez S, Hansen AJ, Hoff-Olsen P, Haas C, Kohler P, Kriegel

AK, Lindblom B, Manohar F, Maroñas O, Mogensen HS, Neureuther K, Nilsson H, Scheible MK, Schneider PM, Sonntag ML, Stangegaard M, Syndercombe-Court D, Thacker CR, Vallone PM, Westen AA, Morling N (2011) Autosomal SNP typing of forensic samples with the GenPlex™ HID System: results of a collaborative study. Forensic Sci Int Genet 5:369–375

6. Børsting C, Sanchez JJ, Morling N (2005) SNP typing on the NanoChip electronic microarray. Methods Mol Biol 297:155–168
7. Mengel-Jørgensen J, Sanchez JJ, Børsting C, Kirpekar F, Morling N (2005) Typing of multiple single-nucleotide polymorphisms using ribonuclease cleavage of DNA/RNA chimeric single-base extension primers and detection by MALDI-TOF mass spectrometry. Anal Chem 77:5229–5235
8. Freire-Aradas A, Fondevila M, Kriegel A-K, Phillips C, Gill P, Prieto L, Schneider PM, Carracedo Á, Lareu MV (2012) A new SNP assay for identification of highly degraded human DNA. Forensic Sci Int Genet 6:341–349
9. Musgrave-Brown (2007) Forensic validation of the SNP*for*ID 52-plex assay. Forensic Sci Int Genet 1:186–190
10. Phillips C, Fang R, Ballard D, Fondevila M, Harrison C, Hyland F, Musgrave-Brown E, Proff C, Ramos-Luis E, Sobrino B, Carracedo A, Furtado MR, Syndercombe Court D, Schneider PM, the SNPforID Consortium (2007) Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel. Forensic Sci Int Genet 1: 180–185
11. NuGEN (2013) Encore™ 384 Multiplex System. NuGEN. http://www.nugeninc.com/nugen/index.cfm/products/pl/library-preparation/encore-384-multiplex-system/. Accessed 30 January 2013
12. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680
13. Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Nucleic Acids Res 40(15):e115
14. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12(6):996–1006
15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303
16. Complete Genomics FTP site: ftp://ftp2.completegenomics.com