Forensic Science International: Genetics 38 (2019) 204–210

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Forensic tri-allelic SNP genotyping using nanopore sequencing

Senne Cornelis[a,b], Yannick Gansemans[a], Ann-Sophie Vander Plaetsen[a], Jana Weymaere[a], Sander Willems[a], Dieter Deforce[a,1], Filip Van Nieuwerburgh[a,1,*]

[a] Laboratory of Pharmaceutical Biotechnology, Ghent University, 9000 Gent, Belgium
[b] Department of Life Sciences and Imaging, imec, 3001 Leuven, Belgium

ARTICLE INFO

Keywords:
Forensic
SNP
Next generation sequencing
Oxford nanopore technologies
MinION

ABSTRACT

The potential and current state-of-the-art of forensic SNP genotyping using nanopore sequencing was investigated with a panel of 16 tri-allelic single nucleotide polymorphisms (SNPs), multiplexing five samples per sequencing run. The sample set consisted of three single-source human genomic reference control DNA samples and two GEDNAP samples, simulating casework samples. The primers for the multiplex SNP-loci PCR were taken from a study which researched their value in a forensic setting using conventional single-base extension technology. Workflows for multiplexed Oxford Nanopore Technologies' 1D and 1D^2 sequencing were developed that provide correct genotyping of most SNP loci. Loci that are problematic for nanopore sequencing were characterized. When such loci are avoided, nanopore sequencing of forensic tri-allelic SNPs is technically feasible.

1. Introduction

Short tandem repeat (STR) profiling using multiplex PCR and subsequent capillary electrophoresis (CE) is currently still the most widely used method to perform DNA based human identification [1]. However, single nucleotide polymorphism (SNP) profiling has gained a lot of interest over the last decades [2]. Both SNP and STR genotyping have distinct advantages and disadvantages over each other [3,4]. In comparison to STRs, SNPs have a lower mutation rate making them better suited for kinship analysis and paternity testing [5]. In addition, SNPs are free from stutters simplifying the interpretation. As the vast majority of SNPs are bi-allelic, a major disadvantage of SNP based identification is the low discriminative power [6]. Moreover, the bi-allelic nature of these SNPs also precludes reliable mixture analysis [7,8]. A possible solution involves using non-binary SNPs. (e.g. tri- or tetra-allelic SNPs) [9,10]. Fewer non-binary SNPs are required to obtain a higher discriminative power with the added advantage that non-binary SNPs enable the identification of multiple donors in a single sample. Westen et al. described a tri-allelic SNP multiplex consisting of 16 loci relevant for forensic identification [11] which relied on the single-base-extension (SBE) technique.

A sequencing based approach would be an elegant alternative to SBE. In recent years, several sequencing based workflows for forensic human identification have been developed and are being commercialized as streamlined, validated methods [12–14]. The use of sequencing techniques for analysis of forensic amplicons has several well described advantages over other techniques [15–17]. A large number of SNP loci can be multiplexed, only being limited by the PCR multiplex capability [18]. Moreover, additional information can be extracted from the regions flanking the SNP, leading to more discriminative power. Despite these advantages, forensic sequencing based approaches are not yet widely used. One of the main hurdles remains the cost per sample which is much higher compared to a SNaPshot SBE analysis. Pooling several samples during a single sequencing run could help to reduce the costs. This is a commonly used technique in sequencing, which requires tagging each amplicon with a sample-specific barcode. In addition to the higher cost per sample, acquiring a state-of-the-art sequencer is a substantial investment [19]. However, Oxford Nanopore Technologies (ONT) recently commercialized the low-cost, pocket-sized MinION nanopore sequencer for general DNA sequencing [20]. This USB-powered device uses disposable flow cells containing an array of nanoscopic pores. While individual DNA molecules pass through the nanopores, they generate an electrogram from which the nucleotide sequence is deduced in real-time [21]. Due to the reduced cost compared to

* Corresponding author at: Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, Gent, Belgium.
E-mail addresses: Senne.Cornelis@UGent.be (S. Cornelis), Yannick.Gansemans@UGent.be (Y. Gansemans), AnnSophie.VanderPlaetsen@UGent.be (A.-S. Vander Plaetsen), Jana.Weymaere@UGent.be (J. Weymaere), Sander.Willems@UGent.be (S. Willems), Dieter.Deforce@UGent.be (D. Deforce), Filip.VanNieuwerburgh@UGent.be (F. Van Nieuwerburgh).
[1] Contributed equally

https://doi.org/10.1016/j.fsigen.2018.11.012
Received 3 April 2018; Received in revised form 9 November 2018; Accepted 9 November 2018
Available online 12 November 2018

traditional sequencers, nanopore sequencing holds promise to become the technology of choice to perform forensic profiling [22–24]. The potential of the MinION sequencer for forensic SNP profiling was has already been demonstrated [25]. Individually PCR-amplified bi-allelic SNP loci of a forensic sample were equimolarly pooled and sequenced using a MinION flow cell (R9.4) generating 2D reads. The current study builds on those findings. Forensic tri-allelic SNP profiles were generated using a 16-plex SNP loci PCR, followed by a ligation mediated sample-specific barcoding step, a nanopore library preparation on the equimolarly pooled ligation products of different samples, and finally a single nanopore sequencing run. Currently, ONT isn't supporting their 2D sequencing method anymore, but offers a 1D and a $1D^2$ method instead. Both methods were assessed. Thus, the current study shows the potential of the state-of-the-art MinION sequencer methods to analyze a forensic tri-allelic SNP multiplex PCR product. The possibility to barcode and pool several samples in a single sequencing run is tested as well. Three positive control reference samples and two GEDNAP (German DNA Profiling, www.gednap.org) samples simulating casework samples were genotyped. All profiles generated by nanopore sequencing were compared with reference profiles obtained via Illumina sequencing.

2. Materials & methods

2.1. Samples

The results presented in this paper were obtained from five samples. Three Promega (Madison, USA) single contributor reference DNA samples (9947, 9948 & 2800) as well as two reference saliva samples obtained from the GEDNAP (German DNA Profiling, www.gednap.org) proficiency tests G50 and G51 were used. These GEDNAP samples, which simulate forensic reference samples, were subjected to a Chelex DNA extraction procedure prior to PCR amplification [26].

2.2. PCR amplification

The 16 SNP loci were amplified using a multiplex PCR (primer sequences available in Supplementary Table 1) based on the report by Westen et al. [11]. Note that the SNP ID *rs9274701* was relabeled to *rs9275142* because the primers provided by Westen et al. [11] amplify an amplicon that is consistent with the latter SNP ID in dbSNP build 150. PCR was completed in a total volume of 12.5 µl containing 1 ng template DNA, 1 X Gold Buffer I (Thermo Fisher Scientific), 9 mM MgCl2, 2 mM dNTPs (Thermo Fisher Scientific, Waltham, USA) and 0.5 U AmpliTaq Gold (Thermo Fisher Scientific). An optimized primer concentration of 100 nM each was used. The primer mix composition can be found in Supplementary Table S1. The temperature profile consisted of an initial denaturation step at 94 °C for 10 min followed by 35 cycles of denaturation at 94 °C for 30 s, primer annealing at 60 °C for 30 s, and extension at 72 °C for 30 s. Subsequent to the amplification cycles a final elongation step of 5 min at 72 °C was performed. The Agilent High-Sensitivity DNA kit (Bioanalyser, Agilent Technologies, California, USA) was used to assess the quality of the generated PCR products. Concentration was checked fluorometrically using a Qubit fluorometer (Life Technologies, Paisley, UK).

Table 1
SNP profiles of all samples obtained from the different sequencing runs. A single genotype indicates all runs resulted in the sample's reference genotype obtained via Illumina sequencing. Discordant results are shown in bold next to the reference genotype and marked with the following suffixes to identifying the sequencing experiments: a) Illumina sequencing, b) 1D reads from the 1D nanopore sequencing run, c) 1D reads from 1D2 nanopore sequencing, and d) $1D^2$ reads obtained from nanopore $1D^2$ sequencing.

SNP locus	2800	9947	9948	GEDNAP 50 C	GEDNAP 51 C
rs1008686	TT	TT	TT	AT	AA
rs1112534	CC	CT	CC	CC	TT
rs17287498	AG	GG	GG	GT	GG
rs2032582	AT	GT	GG	AT	GG
rs2069945	GG[a,d]/**CG**[b,c]	CC	GG	AC	CG
rs2307223	AA	AT	AT	AG	AA
rs2853525	CT	CT	CC	TT	TT
rs3091244	CC	CT	AT	CC	CC
rs34741930	CC	CC	CC	CC	CC
rs35528968	AA	AA	AA	AA	AA
rs356167	AG	CG	AG	CG	CG
rs433342	GG	CG	GG	AG	GG
rs5030240	GT	CG	GG	GT	CG
rs727241	TT	TT	TT	CT	TT
rs9275142	GG	CG	CG[a]/**CC**[b,c,d]	CC	CC
rs9329104	AG	AA	AG	AA	AA

2.3. Nanopore library preparation and sequencing

Oxford Nanopore Technologies' (ONT) base calling software currently requires DNA fragments to have a minimum length of 100 bp to be processed. To circumvent this restriction, PCR amplicons were ligated to create longer DNA fragments. Sample-specific nucleotide barcodes were added to the ligation reaction. This way, the amplicons of the SNP loci are ligated into longer fragments, and at the same time sample-specific barcodes are incorporated into these fragments. The used barcode sequences were identical to those of the Native barcoding protocol developed by ONT. To accomplish the ligation, the PCR products were first purified via gel electrophoresis (E-gel 2% and a 1 kb Plus DNA ladder, Thermo Fisher) in order to remove primers and enzyme. For each sample's PCR multiplex, the desired fragments were recovered by cutting the region of interest (59–115 bp) out of the gel and processed using the Zymoclean Gel DNA Recovery Kit (Zymo Research, Irvine, USA). The resulting SNP amplicons were subsequently end-polished using the NEBNext End-Repair module (NEB, Ipswich, USA). The polished amplicons of each sample were mixed with 0.125 nmol of a sample-specific barcode sequence (Supplementary Table S2) and ligated for 45 min using the Blunt T/A Ligase Mastermix (M0367 NEB, Ipswich, USA), after which a cleanup with 1.8 X AMPure XP beads (Beckman Coulter, High Wycombe, UK) was performed. The quality of the ligation products was assessed using the Agilent High-Sensitivity DNA kit (Bioanalyser, Agilent Technologies, California, USA). For each sample, the purified amplicon-barcode ligation products were quantified fluorometrically using a Qubit fluorometer (Life Technologies, Paisley, UK). The material of all samples was pooled in equimolar quantities prior to library preparation.

In this study both currently available Oxford Nanopore sequencing methods (1D and $1D^2$) were used. The 1D approach only sequences one template DNA strand, whereas with the $1D^2$ method both complementary strands are sequenced and the combined information is used to create a higher quality consensus read. Both sequencing methods require the attachment of a specific leader sequence to the DNA. To add these adaptors, end-repair was performed on 1 µg of the pooled ligation products using the Ultra II End-Repair/dA-Tailing module (NEB, Ipswich, USA) according to the manufacturer's instructions. The resulting A-tailed DNA was cleaned-up using 60 µl of AMPure XP beads (Beckman Coulter, High Wycombe, UK) following the manufacturer's instructions. Subsequently, either 20 µl 1D-adaptor mix followed by 50 µl of Blunt/TA ligase master mix (NEB, Ipswich, USA) or 2.5 µl $1D^2$-adaptor mix in combination with 20 µl Blunt/TA ligase master mix was added to produce 1D and $1D^2$ libraries respectively. The reactions were incubated at room temperature for 10 min. The adaptor-ligated libraries were purified using AMPure XP beads (Beckman Coulter, High Wycombe, UK). The 1D protocol requires the removal of free adaptor using AMPure XP beads (Beckman Coulter, High Wycombe, UK) and the Adaptor Bead Binding buffer (ABB buffer). The 1D library is eluted using 15 µl elution buffer and quantified fluorometrically using a Qubit fluorometer (Life Technologies, Paisley, UK).

In order to generate $1D^2$ reads the samples are subjected to an additional ligation reaction in which 5 µl of the Barcode Adaptor Mix (BAM) is ligated to 45 µl DNA sample using 50 µl Blunt/TA ligase master mix. This ligation takes 10 min after which the $1D^2$ libraries are purified using AMPure XP beads (Beckman Coulter, High Wycombe, UK) and ABB-buffer. As with the 1D protocol, the $1D^2$ samples are eluted in 15 µl elution buffer and quantified fluorometrically using a Qubit fluorometer. Both the 1D and $1D^2$ libraries are further prepared for loading on the flow cell by adding 35 µl of running buffer and 25.5 µl of library loading buffer. The libraries were loaded dropwise on an R9.4 (1D) and R9.5 ($1D^2$) Spot-ON flow cell. Sequencing protocol 48-hour FLO-MIN106_SQK-LSK208 or FLO-MIN107_SQK-LSK308 was chosen to produce the 1D and $1D^2$ reads respectively.

2.4. Data-analysis

The nanopore sequencing data was processed using the Albacore base caller (v3.1.2). The 1D sequencing experiment provides 1D reads whereas the $1D^2$ sequencing experiment provides both 1D and $1D^2$ reads. The FASTQ files generated by albacore, were processed as follows: (1) Sample demultiplexing was done using a custom python script to sort the reads by barcode into sample-specific FASTQ files, by means of a fuzzy regex allowing up to three mismatches to handle sequencing errors. (2) For each sample, the reads were aligned against the reference sequences of the 16 SNP loci using BWA (0.7.15) with default settings supplemented with the -x ont2d setting [27]. These reference sequences consist of the SNP position and 25 nucleotides of the left and the right flank, and were obtained from dbSNP build 150 [28]. A table containing the nucleotide variations at all positions was extracted from the alignment data by means of SAMtools (version 1.3.1) [29] and BCFtools (version 1.3.1) [30], allowing detection and quantification of the SNP variants in the reads. Documentation and scripts are available in the respective notebooks on the TRI-pore GitHub repository (https://github.com/SenneC1/TRI-pore.git). A heterozygous SNP should ideally result in a 50:50 ratio of reads from each allele. A homozygous SNP should result in a 100:0 ratio. However, using sequencing techniques in general, these ratios are rarely obtained because sequence errors get introduced due to PCR amplification of the STR amplicons, the physics of the sequencing technique, the basecaller algorithm and the read mapping algorithm. A more lenient threshold needs to be used to call homo- and heterozygous SNPs. A homozygous call was made when at least 75% of the reads corresponded to one of the possible alleles. This homozygosity threshold, which is in essence arbitrarily chosen, is in line with thresholds used to analyze Illumina reads and is the same threshold that is used in the proprietary Illumina CASAVA SNP calling pipeline. Finally, a visual inspection of the alignments was done using IGV [31] in an attempt to trace erroneous mapping.

2.5. Reference profile

A reference profile of the selected tri-allelic loci for all samples was obtained from Illumina sequencing data. An Illumina sequencing library was created for each sample starting from an aliquot of the SNP multiplex PCR products using the NEBNext® Ultra DNA Library Prep (New England Biolabs, Ipswich, USA), according to the manufacturer's instructions. The Illumina TruSeq adaptors (Illumina, San Diego, USA) were added to the ends of the DNA fragments, followed by a MinElute PCR Purification (Qiagen) procedure to remove buffer and enzyme. Library size selection was performed with the E-Gel iBase Power system (Invitrogen) using an E-gel EX 2% agarose gel and a 1 kb Plus DNA ladder (Thermo Fisher). Fragments with a size of approximately 180 to 300 bp (amplicon + adaptors) were cut from the gel and purified using the Zymoclean gel DNA recovery kit (Zymo Research, Irvine, USA). The recovered DNA fragments were then subjected to an Agilent Bioanalyzer chip analysis (Bioanalyser, Agilent Technologies, California, USA) to ensure that the adaptor ligation was successful. The exact amount of sequence-able library fragments was determined by qPCR using the Sequencing Library qPCR Quantification kit (Illumina, San Diego, USA). Finally, paired-end 150 bp sequencing was performed in a standard flow cell on a MiSeq (Illumina) sequencer. For each sample, the resulting sequencing reads were aligned against the reference sequences of the 16 SNP loci (consisting of the SNP and 25 nucleotides of flanking region on either side, extracted from dbSNP) [[32]] using the BWA (0.7.15) software [27] with default settings. Variant calling and determination of the SNP alleles was done as described for the nanopore sequencing.

3. Results

3.1. Reference profile from Illumina sequencing data

All samples were amplified in parallel with the above described PCR method. Aliquots of the same PCR product were sequenced with nanopore and Illumina sequencing. Profiles generated with Illumina sequencing were used as reference profiles and are shown in Table 1. Half of the 16 tested loci (*rs17287498, rs2032582, rs2069945, rs2307223, rs3091244, rs356167, rs433342, rs5030240*) displayed all three possible alleles in the limited sample (n = 5) population. The relative frequency of each nucleotide observed in the mapped reads at the SNP position is shown in the first stacked bar in Fig. 1, representing the data for the 5 samples.

3.2. Nanopore library preparation

The amplicon and barcode ligation resulted in DNA fragments with a median length between 1000–2000 bp (Supplementary Fig. S1 shows the size distribution of the ligated PCR products), which overcomes the minimum length requirement (100 bp) set by ONT's base calling software.

3.3. 1D sequencing

Nanopore 1D sequencing was continued for 20 h and produced 334,327 reads of which 174,636 were filtered as 'pass' reads (52%). These 'pass' reads had a mean read length of 666 bp. All 'pass' reads had a Phred quality score above five with only 690 reads (0.4%) having a Phred quality score above ten. Sample demultiplexing using the barcode extraction protocol retrieved 155,473 reads with a least one barcode and yielded an average of 31,094 reads per sample (Supplementary Table S3). The sample-specific reads were uniquely mapped against the 51 bp long SNP reference sequences using the ont2d setting of the BWA MEM aligner. This less stringent setting handles a series of typical Oxford Nanopore sequencing errors. The average depth, taking all loci of all samples into account, was 661X (± 435X). The locus with the lowest sequencing depth (95X) across all samples was the *rs2069945* locus of sample 9948. A low coverage was observed for this locus throughout all five samples.

The second stacked bar in Fig. 1 shows the relative frequency of each nucleotide observed at the SNP position using the 1D reads resulting from the 1D sequencing experiment for the five samples. In theory, only one allele for a homozygous locus and two alleles for a heterozygous locus should be observed. However, due to sequencing errors, more than two alleles could be observed at the SNP position. To facilitate SNP profile generation, a homozygosity threshold (see Materials & Methods) is applied to deal with the additional noise of the nanopore sequencing. Based on this cut-off, 78 out of 80 SNP loci over all samples had a genotype corresponding to the reference genotype. The *rs2069945* locus of the 2800 sample and the *rs*9275*142* locus of the 9948 sample were not corresponding with the reference. The complete profiles for all five samples are shown in Table 1, alleles discordant with the reference sequence are indicated with suffix b.

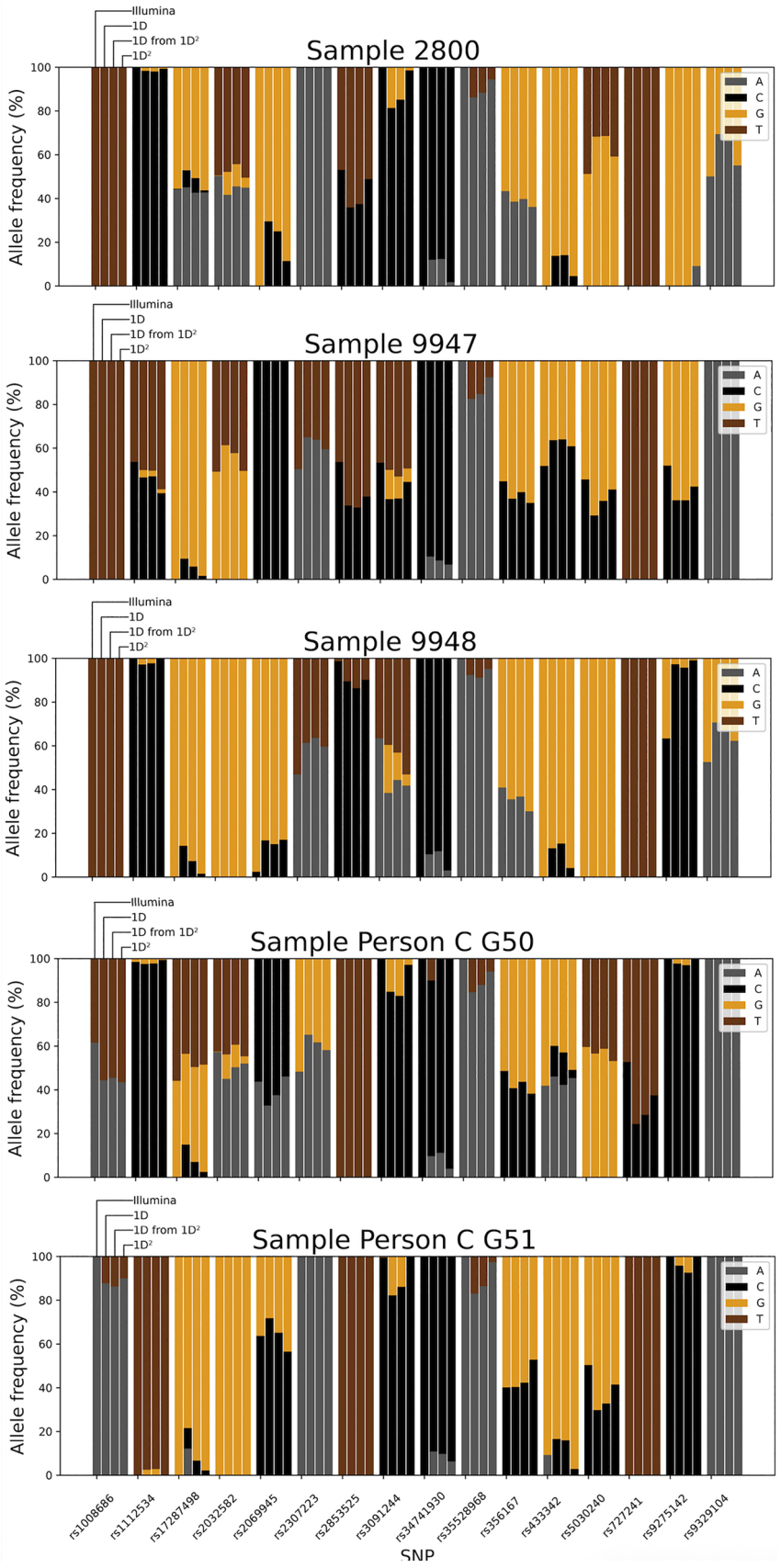

Fig. 1. Overview of the relative frequency of each nucleotide observed at the SNP position of each locus for the 9947, 9948, 2800, GEDNAP 50 Person C and GEDNAP 51 Person C sample. For each SNP, the first stacked bar (1) represents the data for Illumina sequencing, the second one (2) for 1D nanopore sequencing, the third (3) one for 1D data after $1D^2$ nanopore sequencing, and the fourth one (4) for $1D^2$ nanopore sequencing.

3.4. $1D^2$ sequencing

Nanopore $1D^2$ sequencing was continued for 48 h and produced 1,119,765 reads of which only 4407 were filtered as 'pass' $1D^2$ reads (Phred quality score > 11). When lowering the cut-off Phred quality score to 10, 12,812 $1D^2$ reads could be retrieved. These 'pass' reads have a mean read length of 1382 bp. Besides recovering $1D^2$ reads from this sequencing run, it is also possible to extract the 1D reads. Of the 1,119,765 reads, 806,463 (72%) were classified as 'pass' 1D reads. These 1D reads are inherently less qualitative as only 0.1% had a Phred quality score higher than 10. However, the sheer number of 1D versus $1D^2$ reads justifies to use both datasets to generate a SNP profile. Sample demultiplexing using the barcode extraction protocol retrieved 595,556 1D reads and 8862 $1D^2$ with a least one barcode, yielding an average of 119,311 (± 64,456) 1D reads and 1772 (± 1306) $1D^2$ reads per sample. The exact read count per barcode for both 1D and $1D^2$ can be found in Supplementary Table S3.

The sample-specific reads were uniquely mapped against the 51 bp long SNP reference sequences using the ont2d setting of the BWA MEM aligner. The average read depth, taking all loci of all samples into account, was 250X (± 184X) and 5375X (± 2890X) using the $1D^2$ and the 1D reads respectively. The locus with the lowest sequencing depth across all samples was the *rs2069945* locus with 23X using the $1D^2$ reads (GEDNAP G51 Person C) and 1380X when using the 1D reads (Sample 9948). A low coverage was observed for this locus throughout all five samples.

The third and fourth stacked bar of Fig. 1 show the relative frequency of each nucleotide observed at the SNP position using respectively the 1D and $1D^2$ reads resulting from the $1D^2$ sequencing experiment for all five samples. From this figure it is clear that the more accurate $1D^2$ reads generate the least noise. However, occasionally multiple incorrect nucleotides are observed at the SNP position. Hence, application of the homozygosity threshold (see Materials & Methods) is required to determine the SNP profile. The $1D^2$ reads resulted in correct genotyping of 79 out of 80 SNPs over all samples. The only incorrectly genotyped locus was *rs*9275*142* of the 9948 sample. For the 1D reads, 78 of the 80 SNP loci over all samples had a genotype corresponding to the reference genotype. Identical to the results of the 1D sequencing run, the *rs2069945* locus of the 2800 sample and the *rs*9275*142* locus of the 9948 sample were not corresponding with the reference. The latter locus was misinterpreted throughout all nanopore sequencing experiments. The complete profiles for all five samples inferred from 1D and $1D^2$ reads from the $1D^2$ run are shown in Table 1, alleles discordant with the reference sequence are indicated with suffix c and d respectively.

3.5. Sequence data deposition

All sequencing data was deposited in the European Nucleotide Archive (ENA) database under project accession number PRJEB25630 (http://www.ebi.ac.uk/ena/data/view/ PRJEB25630).

4. Discussion

4.1. Nanopore library preparation & barcoding

The amplicon and barcode ligation protocol resulted in sequenceable ligation products, overcoming the minimum length requirement for current nanopore base callers. The barcoding was combined with amplicon ligation to avoid an extra step in the library preparation, even though Oxford Nanopore provides a Native Barcoding kit to allow sequencing of multiplexed samples. Multiplex sequencing is essential to further reduce the costs per sample. However, the total cost per sample remains considerably higher compared to a conventional SNaPshot analysis [33].

4.2. Tri-allelic multiplex

In contrast to a SNaPshot analysis, the quality of the result generated with nanopore sequencing is influenced by the nature of the amplicon sequences in the PCR multiplex. Nanopore sequencing is especially prone to sequencing errors linked to short homopolymeric tracts of four or more bases which tend to produce false inserts of deletions [34,35]. Therefore, the SNP multiplex of choice should contain amplicons lacking such regions. Even with those precautions taken, $1D^2$ nanopore sequencing currently achieves a maximal base calling accuracy of only 95% [36]. One way to cope with this error rate is to obtain a high sequencing depth so that the true SNP alleles are well represented above the noise of sequencing errors and it becomes possible to acquire a reliable consensus sequence. A lower number of analyzed SNP loci improves the coverage per SNP and simplifies the PCR multiplex optimization, clearly demonstrating the advantage of tri-allelic over bi-allelic SNPs. Furthermore, in contrast to a standard bi-allelic SNP multiplex, multiple contributors can be identified when using a tri-allelic multiplex.

4.3. Sequencing

The 1D and $1D^2$ sequencing runs were allowed to proceed until saturation, i.e. up to the moment where no considerable amount of newly produced reads per hour was observed. For the 1D run, new read production was negligible after 20 h, while the $1D^2$ run still produced some new reads when it was stopped after 48 h. The discrepancy between the runs could be attributed to the differences in flow cell type, as well as to the somewhat lower number of active pores in the 1D run (890 vs 1293 for the $1D^2$ run). The latter could have led to faster flow cell saturation in the 1D run and resulted in a lower read yield (334,327 vs 1,119,765 for the $1D^2$ run).

Although the $1D^2$ run produced a satisfactory number of raw 1D reads, the base caller software only managed to extract a small fraction of 'pass' $1D^2$ reads out of them. The high fail rate is most likely a direct result of the characteristics of the sequenced amplicons and the way $1D^2$ reads are generated. The $1D^2$ base calling relies on an alignment of the template and the complement strand to generate a consensus read with better overall quality than the original reads. When sequencing a library of fragments with similar sequences and length, base calling software fails to identify the correct complementary reads producing low quality $1D^2$ reads that will not pass quality filtering (Phred quality score > 11). The incorrect pairing of $1D^2$ reads is caused by reads of the same length getting into the same pore one after the other without being from the same double stranded DNA. Unlike the now discontinued 2D chemistry, where the template and complement strand were covalently linked together through a hairpin, the $1D^2$ chemistry does not physically join the two strands. An adapter is attached to the complementary strand allowing it to be tethered to the membrane while the template strand is being sequenced. Shortly after the template strand leaves the pore, the complement strand can be pulled in and sequenced [37]. Hence, only 4407 qualitative $1D^2$ reads were generated resulting in a sequencing depth too low for reliable SNP profiling.

Adjusting the quality score filtering settings by lowering the minimum Phred quality score from 11 to 10, increased the amount of 'pass' $1D^2$ reads to 12,812. This enabled robust SNP profiling with only the *rs9275142* locus of sample 9948 being incorrectly genotyped. As an alternative, we also used the individual 1D reads, which are less qualitative as no consensus read is created. Nevertheless, a large number of 1D reads can be obtained (806,463), providing in a higher coverage per locus. The use of these 1D reads originating from the $1D^2$ experiment resulted in one additional erroneous SNP locus (*rs2069945*, sample 2800). Similar results were obtained using the 1D reads originating from the 1D experiment, even though less than a third of the reads where produced during this sequencing run. In both cases the *rs2069945* locus of sample 2800 and the *rs9275142* locus of sample

9948 were genotyped incorrectly. The latter SNP was incorrectly genotyped for sample 9948 throughout all nanopore sequencing experiments. Interestingly, ambiguous sequencing results for this SNP have been reported in literature, resulting in the proposition to exclude this locus from the multiplex panel [11]. The incorrectly called *rs2069945* locus of the 2800 sample was only misinterpreted using the 1D reads, but was genotyped correctly when using high quality $1D^2$ reads. The *rs2069945* locus showed the lowest coverage of the entire dataset and is therefore the most prone to sporadic sequencing errors. With a Phred quality score ranging from 5 to 10, the 1D reads have an error rate probability between 10% and 32%. In case of the *rs2069945* locus, this results in an erroneous cytosine nucleotide identification producing a CG heterozygous call. The higher quality $1D^2$ reads, which have a minimal Phred quality score of 10 (error rate better or equal to 10%), only identify a cytosine in 11% of the reads, hence remaining below the heterozygous allelic balance threshold. The difference in quality can clearly be observed by comparing the relative allele frequencies at the SNP location for the 1D and $1D^2$ reads versus the reads generated by Illumina (Fig. 1). In general, the $1D^2$ allele frequencies are closer to the reference Illumina results, with heterozygous balances close to the ideal 50-50 equilibrium and less erroneous nucleotides being detected. However, when only using the original $1D^2$ reads with a quality score above 11 no reliable profile could be generated due to insufficient coverage. Clearly a balance has to be found between the coverage per locus and the quality of these reads. It is believed that significant improvements resulting in higher quality reads are still possible in the software pipelines translating the nanopore's electrical signals into a DNA sequence [40]. However, using the current technology and analysis tools, this study shows that it is possible to correctly genotype 4 of the 5 samples, of which two were casework samples (GEDNAP G50 and G51).

5. Conclusion

The applicability and current state-of-the-art of ONT's MinION nanopore sequencer for forensic tri-allelic SNP profiling was investigated. Forensic tri-allelic SNP profiles were generated using a 16-plex SNP locus PCR, followed by a ligation mediated sample-specific barcoding step, a nanopore library preparation on the equimolarly pooled ligation products of five different reference and casework samples, and finally a single nanopore sequencing run. Data analysis methods for multiplexed Oxford Nanopore Technologies' 1D and $1D^2$ sequencing were developed that provide correct genotyping of almost all SNP loci across all samples. Loci that are problematic for nanopore sequencing were identified. When such loci are avoided, nanopore sequencing of forensic tri-allelic SNPs seems technically feasible.

Conflict of interest

The authors declare no conflict of interests.

Funding

This research was mainly funded by a PhD grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen), awarded to Senne Cornelis (Grant nr° 141456), a Special Research Fund (BOF) from the Ghent University awarded to Jana Weymaere (BOF17/DOC/265) and a project grant from the Research Foundation Flanders (FWO) (G013916N) financing Sander Willems.

Acknowledgments

We would like to thank Sarah De Keulenaer and Ellen De Meester from NXTGNT Belgium for their invaluable practical expertise and assistance in the experiments of this study.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2018.11.012.

References

[1] John M. Butler, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Academic Press, 2005.
[2] Peter Gill, An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes, Int. J. Legal Med. 114 (April (4–5)) (2001) 204–210, https://doi.org/10.1007/s004149900117.
[3] Tim Senge, et al., STRs, mini STRs and SNPs – a comparative study for typing degraded DNA, Leg. Med. 13 (March (2)) (2011) 68–74, https://doi.org/10.1016/j.legalmed.2010.12.001.
[4] John M. Butler, Michael D. Coble, Peter M. Vallone, STRs vs. SNPs: thoughts on the future of forensic DNA testing, Forensic Sci. Med. Pathol. 3 (November (3)) (2007) 200–205, https://doi.org/10.1007/s12024-007-0018-1.
[5] Juan J. Sanchez, et al., A multiplex assay with 52 single nucleotide polymorphisms for human identification, ELECTROPHORESIS 27 (May (9)) (2006) 1713–1724, https://doi.org/10.1002/elps.200500671.
[6] Ranajit Chakraborty, et al., The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems, Electrophoresis 20 (January (8)) (1999) 1682–1696, https://doi.org/10.1002/(SICI)1522-2683(19990101)20:8<1682::AID-ELPS1682>3.0.CO;2-Z.
[7] Bruce Budowle, Angelavan Daal, Forensically Relevant SNP Classes, BioTechniques 44 (Supplement April (4)) (2008) 603–610, https://doi.org/10.2144/000112806.
[8] Bhavik Mehta, et al., Forensically relevant SNaPshot? Assays for human DNA SNP analysis: a review, Int. J. Legal Med. 131 (January (1)) (2017) 21–37, https://doi.org/10.1007/s00414-016-1490-5.
[9] Z.H. Li, et al., Validation of a multiplex system with 20 tri-allelic SNP loci for forensic identification purposes, Forensic Sci. Int. Genet. Suppl. Ser. 4 (1) (2013), https://doi.org/10.1016/j.fsigss.2013.10.165 e324–25.
[10] C. Phillips, et al., Tetra-allelic SNPs: informative forensic markers compiled from public whole-genome sequence data, Forensic Sci. Int. Genet. 19 (November) (2015) 100–106, https://doi.org/10.1016/j.fsigen.2015.06.011.
[11] Antoinette A. Westen, et al., Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples, Forensic Sci. Int. Genet. 3 (September (4)) (2009) 233–241, https://doi.org/10.1016/j.fsigen.2009.02.003.
[12] "Thermofisher.Com, Precision ID NGS System for Human Identification [Online]," accessed June 16, 2016, https://www.thermofisher.com/be/en/home/industrial/forensics/human-identification/forensic-dna-analysis/dna-analysis/next-generation-sequencing-ngs-forensics.html.
[13] "Illumina.com, (2016). ForenSeq DNA Signature Prep Kit. [Online]," accessed August 2, 2016, http://www.illumina.com/products/forenseq-dna-signature-kit.html.
[14] Xiangpei Zeng, et al., High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing, Forensic Sci. Int. Genet. 16 (May) (2015) 38–47, https://doi.org/10.1016/j.fsigen.2014.11.022.
[15] Christophe Van Neste, et al., Forensic STR analysis using massive parallel sequencing, Forensic Sci. Int. Genet. 6 (December (6)) (2012) 810–818, https://doi.org/10.1016/j.fsigen.2012.03.004.
[16] Katherine Butler Gettings, Kevin M. Kiesler, Peter M. Vallone, Performance of a next generation sequencing SNP assay on degraded DNA, Forensic Sci. Int. Genet. 19 (November) (2015) 1–9, https://doi.org/10.1016/j.fsigen.2015.04.010.
[17] Claus Børsting, Niels Morling, Next generation sequencing and its applications in forensic genetics, Forensic Sci. Int. Genet. 18 (September) (2015) 78–89, https://doi.org/10.1016/j.fsigen.2015.02.002.
[18] R. Daniel, et al., A SNaPshot of next generation sequencing for forensic SNP analysis, Forensic Sci. Int. Genet. 14 (January) (2015) 50–60, https://doi.org/10.1016/j.fsigen.2014.08.013.
[19] Antonio Alonso, et al., European survey on forensic applications of massively parallel sequencing, Forensic Sci. Int. Genet. 29 (July) (2017), https://doi.org/10.1016/j.fsigen.2017.04.017 e23–e25.
[20] Alexander S. Mikheyev, Mandy M.Y. Tin, A first look at the oxford nanopore MinION sequencer, Mol. Ecol. Resour. 14 (November (6)) (2014) 1097–1102, https://doi.org/10.1111/1755-0998.12324.
[21] Hagan Bayley, Nanopore Sequencing: From Imagination to Reality, Clin. Chem. 61 (January (1)) (2015) 25–31, https://doi.org/10.1373/clinchem.2014.223016.
[22] Yanxiao Feng, et al., Nanopore-based fourth-generation DNA sequencing technology, Genomics Proteomics Bioinformatics 13 (February (1)) (2015) 4–16, https://doi.org/10.1016/j.gpb.2015.01.009.
[23] J.J. Kasianowicz, et al., Characterization of individual polynucleotide molecules using a membrane channel, Proc. Natl. Acad. Sci. U.S.A. 93 (November (24)) (1996) 13770–13773.
[24] Sophie Zaaijer, et al., Rapid DNA Re-Identification for Cell Line Authentication and Forensics, April (2017), https://doi.org/10.1101/132381.
[25] Senne Cornelis, et al., Forensic SNP genotyping using nanopore MinION sequencing, Sci. Rep. 7 (February) (2017) 41759, https://doi.org/10.1038/srep41759.
[26] Kirsty Phillips, Nicola McCallum, Lindsey Welch, A Comparison of Methods for Forensic DNA Extraction: Chelex-100® and the QIAGEN DNA Investigator Kit (Manual and Automated), Forensic Sci. Int. Genet. 6 (March (2)) (2012) 282–285, https://doi.org/10.1016/j.fsigen.2011.04.018.
[27] H. Li, R. Durbin, Fast and accurate short read alignment with burrows-wheeler

transform, Bioinformatics 25 (July (14)) (2009) 1754–1760, https://doi.org/10.1093/bioinformatics/btp324.
[28] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, DbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (January (1)) (2001) 308–311 n.d..
[29] H. Li, et al., The sequence Alignment/Map format and SAMtools, Bioinformatics 25 (August (16)) (2009) 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.
[30] H. Li, A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data, Bioinformatics 27 (November (21)) (2011) 2987–2993, https://doi.org/10.1093/bioinformatics/btr509.
[31] James T. Robinson, et al., Integrative genomics viewer, Nat Biotech 29 (January (1)) (2011) 24–26, https://doi.org/10.1038/nbt.1754.
[32] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, DbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (January (1)) (2001) 308–311.
[33] Andrei-Tudor Cernomaz, et al., Comparison of next generation sequencing, SNaPshot assay and real-time polymerase chain reaction for lung adenocarcinoma EGFR mutation assessment, BMC Pulm. Med. 16 (December (1)) (2016), https://doi.org/10.1186/s12890-016-0250-0.
[34] Nicholas J. Loman, Joshua Quick, Jared T. Simpson, A complete bacterial genome assembled de novo using only nanopore sequencing data, Nat. Methods 12 (June (8)) (2015) 733–735, https://doi.org/10.1038/nmeth.3444.
[35] Peter Sarkozy, Ákos Jobbágy, Peter Antal, et al., Calling homopolymer stretches from raw nanopore reads by analyzing K-Mer Dwell times, in: Hannu Eskola (Ed.), EMBEC & NBC 2017, vol. 65, Springer Singapore, Singapore, 2018, pp. 241–244, , https://doi.org/10.1007/978-981-10-5122-7_61.
[36] “Thar She Blows! Ultra Long Read Method for Nanopore Sequencing - Assessed 2018-02-05, Http://Lab.Loman.Net/2017/03/09/Ultrareads-for-Nanopore/,” n.d.
[37] Carlos Victor de Lannoy, Dick de Ridder, Judith Risse, A Sequencer Coming Of Age: De Novo Genome Assembly Using MinION Reads, May 26 (2017), https://doi.org/10.1101/142711.