Forensic Science International: Genetics 56 (2022) 102621

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Nanopore sequencing of a forensic combined STR and SNP multiplex

Olivier Tytgat [a,b], Sonja Škevin [a], Dieter Deforce [a], Filip Van Nieuwerburgh [a,*]

[a] Laboratory of Pharmaceutical Biotechnology, Ghent University, 9000 Gent, Belgium
[b] Imec, Kapeldreef 75, Leuven 3001, Belgium

ARTICLE INFO

Keywords:
Next-generation sequencing
Short Tandem Repeats
Single Nucleotide Polymorphisms
Nanopore sequencing

ABSTRACT

Nanopore sequencing for forensic purposes has gained attention, as it yields added discriminatory power compared to capillary electrophoresis (CE), without the need for a high up-front capital investment. Besides enabling the detection of iso-alleles, Massively Parallel Sequencing (MPS) facilitates the analysis of Short Tandem Repeats (STRs) and Single Nucleotide Polymorphisms (SNPs) in parallel. In this research, six single-contributor samples were amplified by such a combined multiplex of 58 STR and 94 SNP loci, followed by nanopore sequencing using an R10.3 flowcell. Basecalling was performed using two state-of-the-art basecallers, Guppy and Bonito. An advanced alignment-based analysis method was developed, which lowered the noise after alignment of the STR reads to a reference library. Although STR genotyping by nanopore sequencing is more challenging, correct genotyping was obtained for all autosomal and all but two non-autosomal STR loci. Moreover, genotyping of iso-alleles proved to be very accurate. SNP genotyping yielded an accuracy of 99% for both basecallers. The use of novel basecallers, in combination with the newly developed alignment-based analysis method, yields results with a pronouncedly higher STR genotyping accuracy compared to previous studies

1. Introduction

Massively Parallel Sequencing (MPS) technologies have become a well-established approach for forensic human identification [1]. The most commonly used markers for forensic DNA profiling are Short Tandem Repeats (STRs), which are short nucleotide sequences repeated multiple times in a head-to-tail fashion [2]. The repeat number for such loci varies between individuals, and can thus be used to generate a unique DNA fingerprint. STR sizing is realized mainly by performing a polymerase chain reaction (PCR) followed by capillary electrophoresis (CE) [3]. Although this generates highly informative profiles, sequence variants within the amplicons, such as iso-alleles and SNPs, are not detected by CE. MPS-based approaches yield additional discriminatory power for STR analysis, which is in particular useful for low-input samples. Moreover, MPS allows higher order SNP and STR multiplexing, enabling analysis of STR loci in parallel with Single Nucleotide Polymorphism (SNP) loci. In contrast to STRs, the amplicons targeting the SNP markers are much shorter, which is favorable for the analysis of highly degraded samples [4].

Although validated forensic MPS strategies are commercially available, e.g. the Verogen technology [5], the widespread implementation in forensic routine is hampered by both the high up-front capital investment and the high reagent costs. The MinION device, an affordable long-read sequencer commercialized by Oxford Nanopore Technologies, has gained importance in the forensic field [6]. Moreover, the device is handheld, enabling on-site analysis of samples, which could be of great use for disaster victim identification, or for extremely urgent crime scene samples.

Although the technology has improved considerably, nanopore sequencing still results in a higher level of sequencing error noise than Illumina sequencing [7]. Nevertheless, accurate data for sequencing of forensic bi- and multi-allelic SNPs were obtained by our group [8,9]. Nanopore sequencing of forensic STRs proved to be more cumbersome [10–13]. Some specific locus-dependent success-limiting factors hampering accurate STR genotyping could be identified, one of them being the presence of homopolymers in the repeat or flanking region [12]. In this research, a forensic STR- and SNP-multiplex is nanopore sequenced using an R10.3 flowcell. This novel type of flowcell was designed to resolve homopolymers with a higher accuracy. It features pores characterized by a dual constriction, which both modulate the raw signal obtained during sequencing [14]. Moreover, two state-of-the-art basecallers (Guppy and Bonito) are compared for this purpose, and an improved alignment-based analysis method is demonstrated.

* Correspondence to: Ottergemsesteenweg 460, 9000 Gent, Belgium.
E-mail address: Filip.VanNieuwerburgh@UGent.be (F. Van Nieuwerburgh).

https://doi.org/10.1016/j.fsigen.2021.102621
Received 7 July 2021; Received in revised form 22 October 2021; Accepted 22 October 2021
Available online 28 October 2021

2. Materials and methods

2.1. Samples

The results presented in this research were obtained from six samples: two commercially available reference DNA samples (9947a and 9948) from OriGene (Rockville, Maryland, USA), and four blood samples collected from anonymous donors. The ethical review board of Ghent University Hospital provided ethical approval, and all healthy volunteers signed the informed consent (BC-05557). The blood samples were obtained by a finger puncture using a 21G Minicollect® Lancelino safety lancet with a penetration depth of 2.4 mm (Greiner Bio-One, Kremsmünster, Austria) and collected in a K3E K3EDTA Minicollect® collection tube (Greiner Bio-One, Kremsmünster, Austria). DNA extraction of the blood samples was performed using the DNeasy® Blood and Tissue kit according to the manufacturer's instructions.

2.2. PCR amplification

All six samples were amplified using the ForenSeq DNA Signature Prep Kit (Verogen, San Diego, USA), using Primer Mix A, according to the manufacturer's instructions. Primer Mix A contains 149 primer pairs, targeting 27 autosomal STRs, 24 Y-STRs, 7 X-STRs, and 94 identity SNPs. It should be noted that primers designed for some other loci (e.g. SE33) are also included in the Primer Mix, but are not analysed by the Universal Analysis Software [7]. All primers consist of a target-specific region and a mutual overhang region. A first PCR reaction was performed using a SimpliAmp Thermal Cycler (ThermoFisher Scientific, Waltham, MA, USA), during which the target-specific regions anneal to their complement. In a second PCR step, the targets were enriched, along with the incorporation of indexes for sample de-multiplexing and sequencing adapters. After amplification, samples are purified using the Sample Purification beads provided in the DNA Signature Prep Kit, according to the manufacturer's instructions. Elution was performed in 52.5 µL Resuspension Buffer, aliquots of this eluate were subjected to both Verogen and nanopore sequencing.

2.3. Verogen sequencing

After performing bead-based normalization, the samples were pooled and denatured as specified in the DNA Signature Prep Kit protocol. Immediately after denaturation, the samples were loaded on the reagent cartridge. Paired-end sequencing was performed on the MiSeq FGx Sequencing System (Illumina, San Diego, USA). Data analysis was done using the ForenSeq Universal Analysis Software v1.3 (Verogen). For all loci, the analytical threshold, which is the lower limit of detection, was set at 1.5% of the reads for the specific locus, and the interpretation threshold at 4.5%. This implicates that each allele to which more than 4.5% of the reads is assigned, should be interpreted as a true allele or a stutter. The threshold for STR intra-locus imbalance was set at 60%. For SNPs this was set at 50%. The stutter filter settings for the STRs were locus dependent, default settings were used.

2.4. Nanopore sequencing

The purified amplicons obtained after the PCR2 amplification step were subjected to library preparation for nanopore sequencing as described in previous work [12]. An input of 75 ng was used for each sample. DNA repair and end preparation were performed using NEBNext FFPE DNA Repair Mix and NEBNext End Repair/dA-Tailing Module (NEB, Ipswich, MA, USA). After purification using a 1.8 × volume of AMPure XP beads (Beckman Coulter, High Wycombe, UK), barcode ligation was performed using the Native Barcoding Expansion 1–12 (EXP-NBD104) kit (ONT, Oxford, UK). This was realized by adding 25 µL NEB Blunt/TA Ligase Master Mix and 2.5 µL Native Barcode to the sample, followed by a 10 min incubation step at room temperature. Next, the barcoded amplicons were purified using a 1.8 × volume of AMPure XP beads and quantified using a Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher, Waltham, MA, USA). An equimolar pool, with a total input of 50 ng, was subjected to adapter ligation. To realize this, 5 µL of Adaptor mix II, 20 µL of NEBNext Quick Ligation Reaction Buffer, and 10 µL of Quick T4 DNA Ligase were added to the library, followed by a 10 min incubation step at room temperature. Again, a purification step was performed using a 1.8 × volume of AMPure XP beads, followed by quantifying the final library using a Qubit fluorimeter. 19 ng of DNA was loaded onto the SpotON flowcell, according to the manufacturer's instructions, using the SQK-LSK109 kit (ONT, Oxford, UK). Sequencing was performed using a GridION device, which accomodates the same flowcells as the portable MinION device. Sequencing was performed for 48 h, to obtain a maximal amount of data. However, as the flow cell quality deteriorates over time, most reads are obtained during the first hours of sequencing.

2.5. Data analysis

Basecalling was performed with both the fully supported basecaller Guppy (v.4.3.4, ONT) and the research basecaller Bonito (v.0.3.8, ONT). Sample de-multiplexing based on the barcode sequence was done in real-time by the MinKNOW control software. For autosomal STR genotyping, a reference library (Supplementary File 2) was constructed for all investigated STR-loci, encompassing all alleles occurring within the European population with a frequency > 1%. The population information was obtained using the pop.STR database [15], while the sequence data were obtained from STRbase [16] and STRSeq [17]. Moreover, the frequently occurring iso-alleles reported in the STRbase and STRSeq databases were included, as well as iso-alleles detected in the Verogen sequencing results. For most Y-STR loci, sequence information was obtained from STRBase 2.0 [16]. For Y-STR loci not included in the STRBase 2.0 database, the sequence of the repeat unit and the flanking regions were retrieved from the obtained Verogen reads within this study. Population data were retrieved from the YHRD database [18], as well as from data obtained by Kline et al. [19]. X-STR sequence information and population data were obtained from Borsuk et al. [20]. Alignment of the obtained reads against this reference library was performed by Burrow Wheelers Aligner (v0.7.17) with the -x ont2d option enabled [21]. To lower the noise in the genotyping data, caused by amplification and sequencing errors, an alignment score (AS) filter was applied. As the AS reflects how well the obtained read resembles the reference it aligned to, higher AS scores are expected to be found for true alleles and stutters. Reads affected by sequencing and basecalling errors might lead to mis-alignment, characterized by a lower AS. The maximal AS is read-specific and equals the read span. Based on the CIGAR string, the read span for each aligned read was calculated. All reads with an AS lower than 90% of the read span were discarded. The resulting read counts per allele were used for genotyping, using the same genotyping rule as used in previous research [12]: for each locus, the allele with the highest read count was called as present, as well as the second most abundant allele if the corresponding read count equals at least 50% of the maximal read count. For single-copy Y-STR loci, only the allele with the highest read count was called as present.

For SNP genotyping, all reads were aligned using Burrow Wheelers Aligner (v0.7.17) to a library of reference sequences, containing one reference per locus. These references are all 51 nucleotides long, with the SNP positioned centrally, and were retrieved from the Single Nucleotide Polymorphism database (dbSNP) [22]. Based on the obtained alignment data, the nucleotide variations at all positions were extracted. The read count corresponding to each possible allele at the SNP position was obtained by SAMtools (v1.11) [23] and BCFtools (v1.6–45-gdb2e2b6) [24], and was used for SNP genotyping. An arbitrary allelic imbalance cut-off should be set for heterozygous samples, to manage PCR and sequencing bias, as well as biological phenomena such as copy number variation and somatic mutations. All alleles

representing more than 20% of the total read count were called as present.

3. Results and discussion

3.1. Verogen sequencing and genotyping

Six samples were amplified with the ForenSeq DNA Signature Prep Kit and sequenced with a MiSeq FGx® device. An average of 283,778 ± 90,840 reads were sequenced per sample, of which 62.7% was assigned to an STR locus, and 37.3% to an SNP locus, on average. The genotypes obtained by the Universal Analysis Software are shown in Supplementary Table S1. On average, 93% of the Verogen reads corresponding to one of the targeted STR loci were assigned to a true allele, ranging from 86% to 99%. This is shown in Table 1 and Supplementary Table S2, for the autosomal and the non-autosomal STR loci, respectively. Non-true-allele assignment occurs due to amplification errors and artifacts, e.g. stutter, and sequencing error. SNP genotyping could be performed for all but three loci of sample E. For these three loci, there was insufficient sequencing depth. For all STR loci, a sufficient number of reads was obtained to allow genotyping.

3.2. Nanopore sequencing, basecalling, alignment, and alignment filtering

Nanopore sequencing using an R10.3 flowcell, resulted in 132,268 ± 56,719 Guppy-basecalled reads per sample that aligned to an STR locus. After AS filtering, 54.7% of these reads were retained, on average. After Bonito basecalling, 71,906 ± 38,373 reads per sample aligned to an STR locus, of which 85.4% was retained after AS filtering. SNP variant calling could be performed using 60,049 ± 27,124 Guppy basecalled reads per sample, whereas 50,441 ± 27,985 Bonito basecalled reads per sample could be used for SNP genotyping. An overview of the sequencing depth per sample is shown in Table 2. The read counts obtained after alignment and AS filtering are shown in Supplementary Files 3 and 4, for Guppy and Bonito, respectively.

Table 1 and Supplementary Table S2 show the percentage of the nanopore reads aligning to the true-allele(s) both before and after AS filtering. Discarding the reads with an insufficiently high alignment score resulted in an increase of the true-allele alignment for almost all loci, with a maximum of 10 percentage point. However, for two loci, the true-allele alignment decreased. On average, AS filtering resulted in an increase of true-allele alignment of about 4 percentage point for Guppy reads, whereas for the Bonito reads, this metric only increased with 2 percentage point. A two-tailed, paired T-test indicated no statistical significant difference in true-allele alignment between both basecallers ($p = 0.49$).

A violin plot illustrating the distribution of alignment scores after Guppy basecalling, normalized for read span, is shown in Fig. 1 for all autosomal STR loci, and in Supplementary Fig. S1 for the non-autosomal STR loci. Fig. 1 shows that for locus D7S820, an insufficient number of reads were retained after AS filtering. Therefore, this step was not applied for the D7S820 locus. The cut-off for AS filtering of 90% was chosen arbitrarily, and should preferably be optimized for each locus separately. Nevertheless, these findings clearly show that a substantial part of the obtained noise can be filtered out bioinformatically.

3.3. SNP genotyping after nanopore sequencing

A heterozygous sample is theoretically expected to result in a 50:50 ratio of reads for both alleles. However, the amplification, sequencing, basecalling, and variant calling process causes deviations from this theoretical ratio. Therefore, an allele was called as present when more than 20% of the reads were assigned to this allele. Fig. 2 shows an overview of the SNP genotyping results after both Guppy and Bonito basecalling. Variant calling of both datasets resulted in an accuracy of 99%, as 555 out of 561 SNP loci were genotyped correctly, taken all SNP

Table 1
True-allele alignment after Verogen sequencing and nanopore sequencing combined with both Guppy and Bonito basecalling, for all autosomal STR loci.

Locus	True-allele alignment (Verogen) (%)	Guppy			Bonito		
		True-allele alignment before AS filter (%)	True-allele alignment after AS filter (%)	Difference (Percentage point)	True-allele alignment before AS filter (%)	True-allele alignment after AS filter (%)	Difference (Percentage point)
D2S441	97	80	90	10	89	92	3
PentaD	98	51	60	9	73	77	4
D18S51	88	50	58	8	41	41	0
PentaE	99	67	75	8	82	83	1
D1S656	91	73	81	8	82	84	2
D9S1122	92	73	80	7	83	85	2
CSF1PO	97	71	77	6	76	77	1
D5S818	96	71	77	6	81	82	1
D3S1358	92	74	79	5	82	84	2
D4S2408	97	75	80	5	84	87	3
D8S1179	90	71	76	5	79	82	3
D13S317	96	69	74	5	77	81	4
FGA	88	45	49	4	34	34	0
vWA	93	73	77	4	81	82	1
D16S539	93	73	77	4	78	73	-5
D17S1301	91	73	77	4	81	83	2
D22S1045	92	68	72	4	77	79	2
D2S1338	87	72	76	4	78	79	1
D6S1043	92	66	70	4	74	76	2
D12S391	86	58	61	3	67	68	1
D19S433	95	77	80	3	80	82	2
TH01	95	87	90	3	90	91	1
TPOX	97	90	93	3	94	96	2
D20S482	92	67	70	3	75	77	1
D10S1248	91	80	82	2	82	83	1
D21S11	93	78	71	-7	84	85	2
D7S820	96	79	N/A	N/A	85	N/A	N/A
Average	**93**	**71**	**75**	**4**	**77**	**79**	**2**
Standard deviation	**3**	**10**	**10**	**3**	**13**	**13**	**2**

Table 2
Sequencing depth after nanopore sequencing.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Average	Standard deviation
Guppy basecalled STR reads	105,211	103,499	60,699	139,864	225,107	159,228	132,268	56,719
% retained after AS filtering	52.2	56.1	56.8	53.7	53.6	57.0	54.7	2.0
Bonito basecalled STR reads	59,645	58,127	14,051	80,653	129,164	89,796	71,906	38,373
% retained after AS filtering	84.1	85.5	84.2	86.4	85.8	85.1	85.4	0.9
Guppy basecalled SNP reads	34,713	47,612	26,038	85,818	81,930	84,181	60,049	27,124
Bonito basecalled SNP reads	28,611	42,729	9270	77,255	71,305	73,474	50,441	27,985

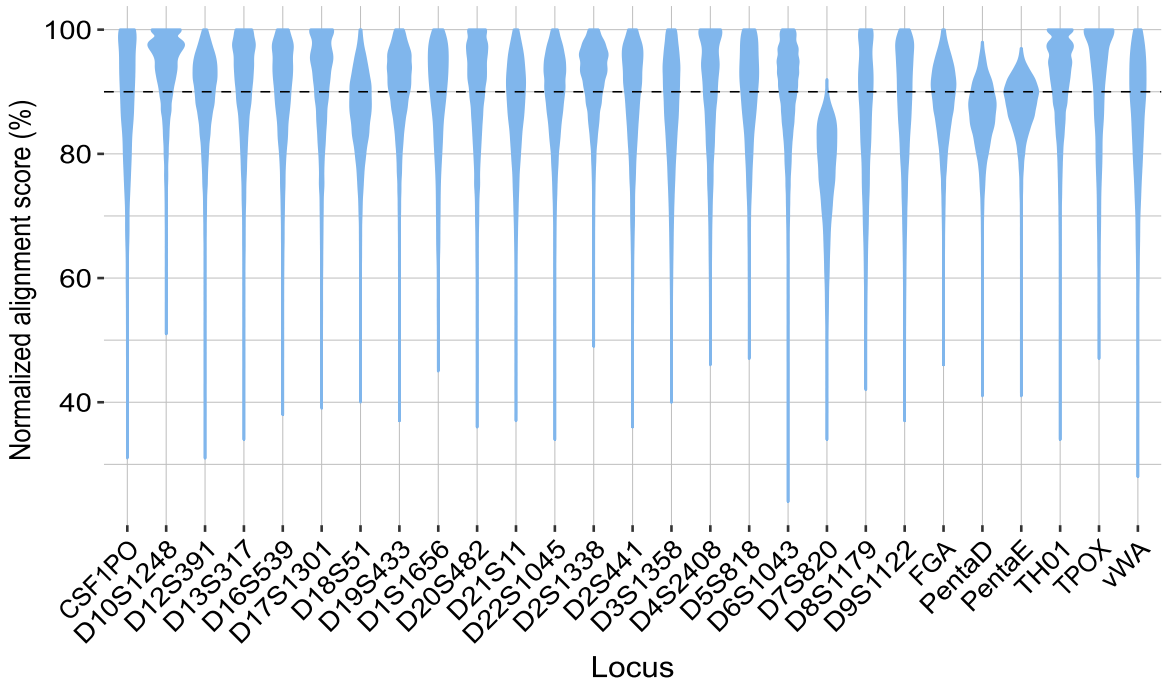

Fig. 1. Violin plot showing the distribution of alignment scores after Guppy basecalling, normalized by read span, for all autosomal STR loci.

loci of all samples together. Due to low read depth, allelic drop-out was observed for locus rs4606077 in three samples, and locus rs907100 in one sample. Two samples were genotyped incorrectly for locus rs6955448, due to an allelic imbalance which was also present in the Verogen data and thus most probably originated during PCR. Locus rs1031825, which is characterized by the presence of a homopolymer next to the SNP position, was genotyped incorrectly for sample F. Three SNP loci could not be determined by Verogen, due to insufficient read depth. As a consequence, the nanopore sequencing data of these loci could not be compared to the Verogen data.

3.4. STR genotyping after nanopore sequencing

The AS filtered alignment results obtained after nanopore sequencing were used for STR genotyping. The concordance between nanopore and Verogen results for all samples was assessed by comparing the obtained length-based genotypes, for both Guppy basecalling and Bonito basecalling. An overview is shown in Fig. 3. In general, for both basecallers, most autosomal STR loci were called correctly for all samples. Nevertheless, some remarkable differences between both datasets should be pointed out. Genotyping using Guppy basecalled reads was correct for all loci, except for two genotypes for locus PentaD. Interestingly, these genotypes were called correctly using Bonito basecalled reads, implicating that all loci were genotyped correctly by at least one basecaller. However, all obtained genotypes for loci D18S51 and FGA were incorrect after Bonito basecalling, as well as locus D16S539 for sample A. Moreover, locus CSF1PO for sample C was also genotyped incorrectly after Bonito basecalling. Although the average true-allele alignment for this locus is relatively high, the insufficient read depth of 24 leads to incorrect genotpying.

Fig. 3 shows that some loci, indicated in blue, were characterized by allelic imbalance which was also observed in the Verogen data. This imbalance thus originates from PCR bias or biological phenomena, such as copy number variations and somatic mutations. Due to slippage of the polymerase during amplification, stutter + 1 and stutter − 1 PCR artifacts are present in the amplified sample. Unfortunately, the signal corresponding to these artifacts is increased due to nanopore sequencing and alignment errors. For some loci, this often results in other highly represented alleles. This makes differentiation between imbalance and sequencing or alignment noise challenging. Currently, a peak should be at least 50% of the highest peak to be called as a true allele, which is an arbitrary cut-off. Making this rule less stringent would allow correct genotyping despite the occurrence of such allelic imbalance, but would lead to drop-ins.

For the non-autosomal STR-loci, a similar pattern is observed. All samples were genotyped correctly using Guppy, except for locus DXS10135 (three genotypes) and locus DXS10103 (one genotype), as shown in Supplementary Fig. S2. Bonito basecalled reads resulted in incorrect profiles for 7 loci, for most samples. This indicates that, although the average true-allele alignment is slightly higher after Bonito basecalling, genotyping fails consistently for a specific subset of STR loci using this basecaller.

In general, both the alignment and genotyping data show that the presence of homopolymers, high repeat numbers, complex repeat patterns, and a high similarity between repeat region and the flanking regions proved to hamper the accuracy of STR genotyping after nanopore

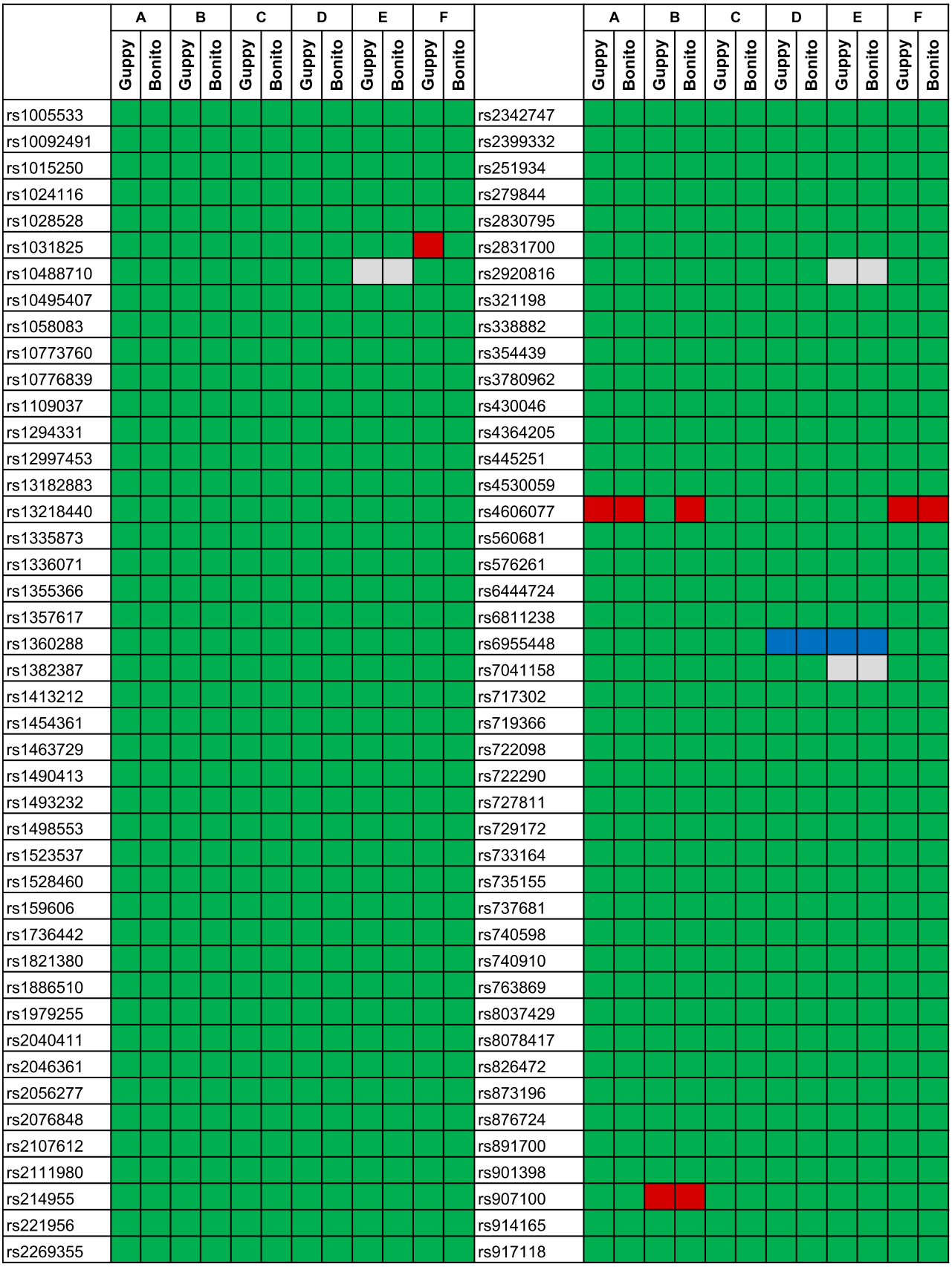

Fig. 2. Overview of genotyping results after both Guppy and Bonito basecalling, for all SNP loci. Green indicates correct genotyping; red indicates incorrect genotyping; blue indicates incorrect genotyping due to allelic imbalance which is also present in the Verogen data. For the loci indicated in grey, no Verogen data were obtained due to insufficient read depth. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article)

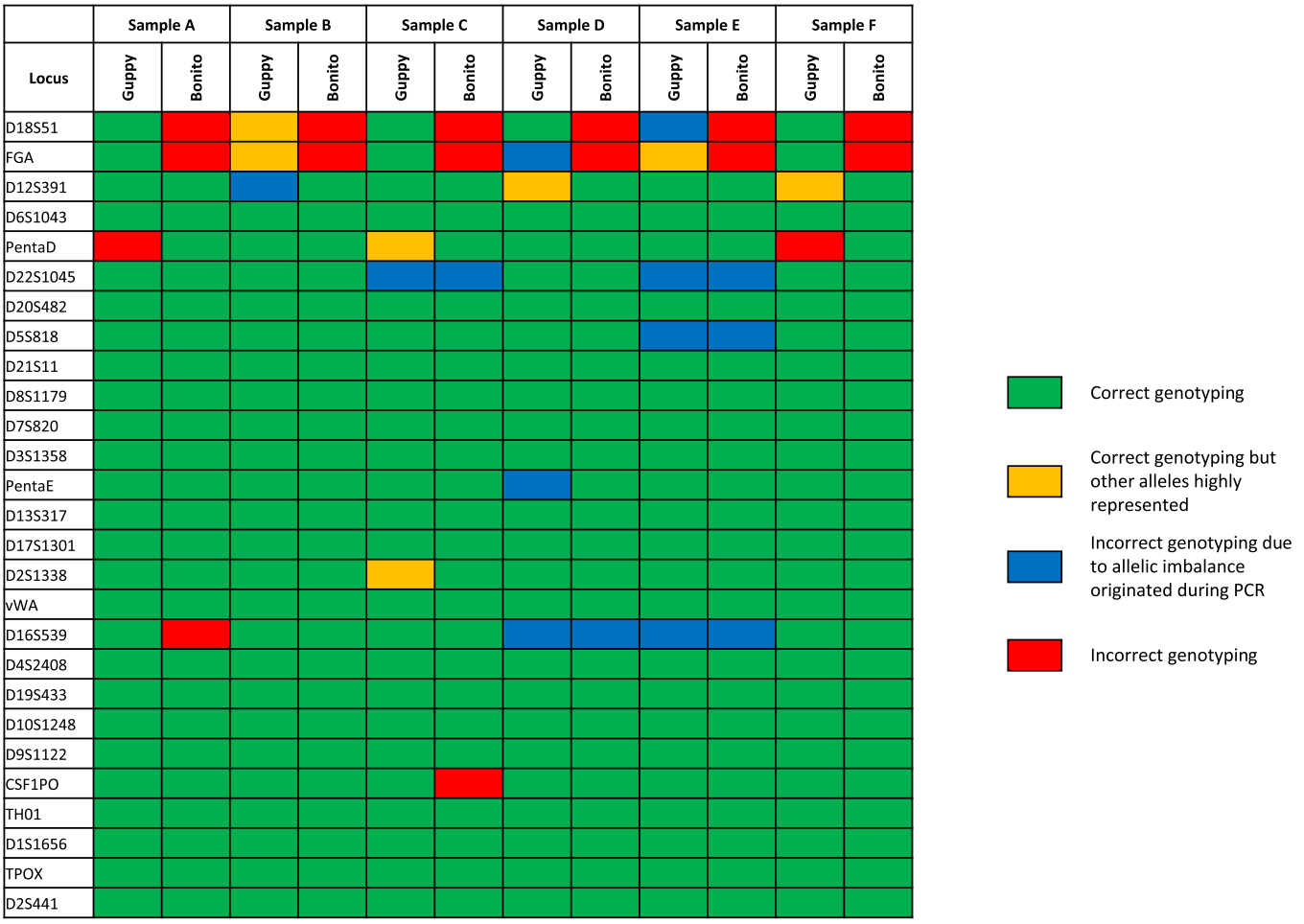

Fig. 3. Overview of genotyping results after both Guppy and Bonito basecalling, for all autosomal STR loci. Green indicates correct genotyping; yellow indicates correct genotyping, but hampered due to other highly represented alleles; red indicates incorrect genotyping; and blue indicates incorrect genotyping due to allelic imbalance originated during PCR, and thus is also present in the Verogen data. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article)

sequencing. These findings correspond well to our previous study [12], where we identified these success-limiting locus-dependent characteristics. Loci that proved to be troublesome for nanopore sequencing were vWA, SE33, FGA, D21S11, D18S51, and D3S1358. Although the genotyping accuracy improved for most of these loci by using improved basecallers and AS filtering, loci FGA and D18S51 remain challenging for this purpose. Nevertheless, the data obtained in this research show a substantial improvement compared to previous studies.

3.5. Genotyping of STR iso-alleles

Sequencing-based genotyping of STR alleles has the added advantage of yielding information on iso-alleles, thereby increasing the discriminatory power of the assay. Frequently occurring iso-alleles were included in the reference allele library for alignment. Table 3 shows that nanopore sequencing is capable of accurately genotyping iso-alleles, as on average 96% of the reads aligning to one of the alleles with the correct repeat number, aligned to the truly present iso-allele, whereas only 4% of the reads aligned to non-true iso-alleles with the same repeat number. Moreover, as shown in Fig. 4, four genotypes could correctly be called as heterozygous, although both alleles share the same repeat number. This added discriminatory power is of great value for genotyping of low-input samples, which are prone to allelic drop-out, or mixture samples.

Table 3
Iso-allele genotyping accuracy: number of Guppy basecalled reads that aligned to a true iso-allele and the number of reads that aligned to the other iso-alleles with the same repeat number.

Locus	True iso-allele alignment (number of reads)	Non-true iso-allele alignment (number of reads)	True iso-allele alignment (%)
DYS448	441	0	100
D3S1358	6055	130	98
D8S1179	1998	38	98
vWA	309	7	98
D9S1122	3131	98	97
D5S818	418	18	96
D21S11	2620	179	93
DYS389II	1604	114	93
Total	**16,576**	**584**	**96**

3.6. Data analysis strategy

Multiple tools have been described to perform STR genotyping using NGS data, e.g. STRinNGS [25], STRait Razor [26], MyFLq [27], RepeatHMM [28], and FDSTools [29]. Ren and colleagues used RepeatHMM to perform STR genotyping on a dataset obtained by nanopore sequencing of a Verogen ForenSeq DNA Signature Prep Kit library with an R9.4 flowcell and Guppy basecalling (v4.2.2) [13]. This resulted in a low accuracy overall. Moreover, Ren and colleagues

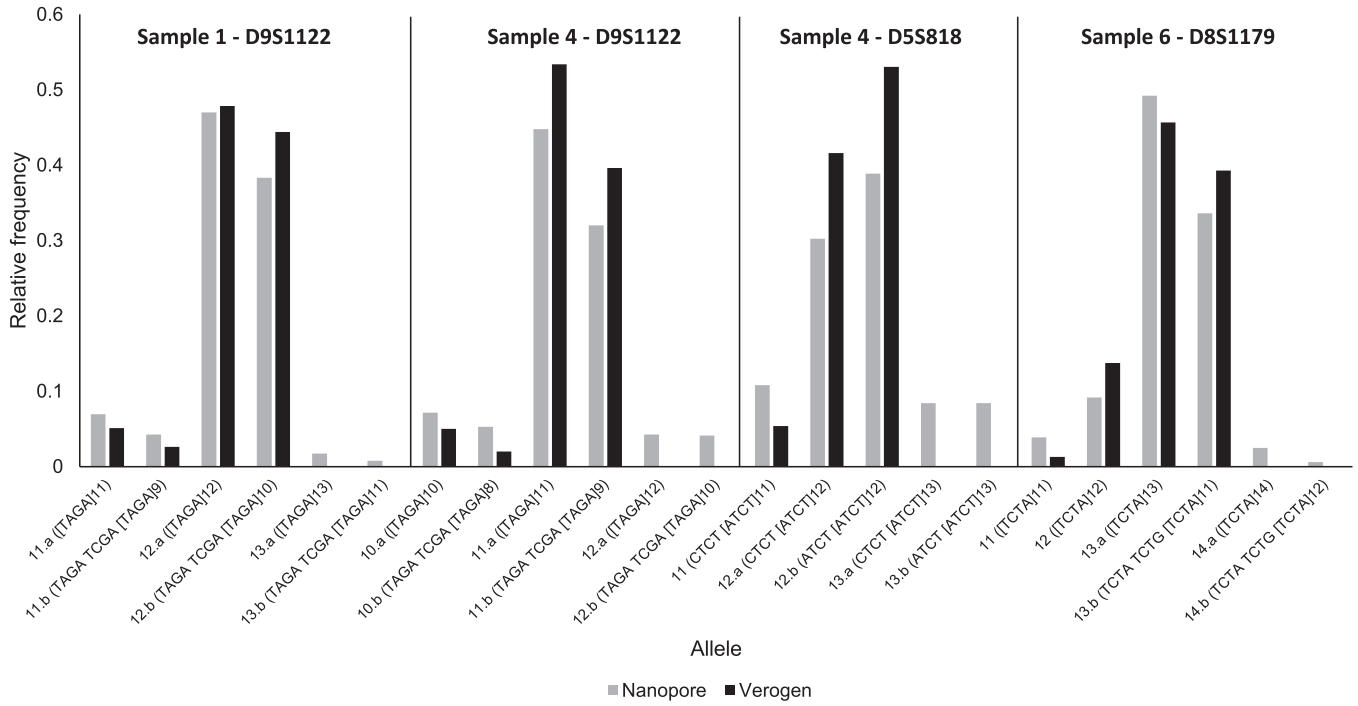

Fig. 4. Iso-allele genotyping reveals four additional heterozygous samples. The Y-axis shows the relative frequency of reads per allele, the X-axis shows the alleles. Sample number and locus are indicated above the bar chart.

analyzed the same dataset with a custom tool that aims to extract individual repeats from each read. However, this tool is not able to genotype partial repeats, and genotyping failed for about 50% of the autosomal STR loci. The results obtained in our research show that alignment of the reads to a reference allele database yields more accurate results, as all autosomal STR loci were genotyped correctly by at least one of both used basecallers. Moreover, we re-analyzed the reads obtained by Ren et al. [dataset] [30] with our workflow for the 2800 M positive control sample, which was sequenced in triplicate. The read counts obtained after alignment can be found in Supplementary File 5. A genotyping accuracy of 100% was obtained for all three triplicates. These findings suggest that the use of the R10.3 flowcell may not increase the genotyping accuracy. Moreover, this clearly indicates the importance of selecting a suited analysis method. The major drawback to our alignment strategy is the fact that the presence of all possible (iso-) alleles in the library is crucial. Absence of the true allele in the reference library might lead to incorrect genotyping for this specific locus. The library should thus constantly be improved based on population data gathered by sequencing. As sequencing is becoming more important in the field of forensic genotyping, expanding the existing databases (e.g. pop.STR [15] and STRSeq [17]) with data gathered by the community will be crucial. Nevertheless, nanopore sequencing has become a method capable of accurately genotyping forensic samples, and further improvements might enable implementation of this genotyping method in forensic routine.

4. Conclusion

Nanopore sequencing of a forensic STR and SNP multiplex was compared to Illumina sequencing for six single-contributor samples. Basecalling was performed with two state-of-the-art basecallers, Guppy and Bonito. Both datasets resulted in a 99% SNP genotyping accuracy. All autosomal STR genotypes were accurately called with at least one of both basecallers. A slightly higher fraction of the reads aligned to a true allele after Bonito basecalling, yet genotyping accuracy was lower for this basecaller, as a specific subset of loci failed consistently. Our analysis method, based on alignment of STR reads to a reference library with subsequent filtering based on the alignment score, was capable of accurately genotyping iso-alleles. The STR profiling after nanopore sequencing presented in this research is much more accurate compared to previous studies. These findings are an important step towards on-site sequencing of forensic samples using an affordable, handheld MinION device.

CRediT authorship contribution statement

Olivier Tytgat: Conceptualization, Investigation, Writing – original draft. **Sonja Škevin:** Formal analysis, Writing – review & editing. **Dieter Deforce:** Writing – review & editing, Supervision. **Filip Van Nieuwerburgh:** Conceptualization, Writing – review & editing, Supervision.

Funding

This work was supported by a PhD grant from the Special Research Fund (BOF) from the Ghent University [Grant BOF18/DOC/200 to O. T.].

Competing interest

None declared.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2021.102621.

References

[1] B. Bruijns, R. Tiggelaar, H. Gardeniers, Massively parallel sequencing techniques for forensics: a review, Electrophoresis 39 (21) (2018) 2642–2654.
[2] J.M. Butler, Forensic DNA Typing: Biology, Technology, and Genetics of STR markers, second ed., Elsevier,, Berlington, USA, 2005.
[3] J. Butler, B. McCord, J. Jung, R. Allen, Rapid analysis of the short tandem repeat HUMTH01 by capillary electrophoresis, BioTechniques 17 (6) (1994) 1062–1064, 1066, 1068 passim.

[4] C. Børsting, H.S. Mogensen, N. Morling, Forensic genetic SNP typing of low-template DNA and highly degraded DNA from crime case samples, Forensic Sci. Int. Genet. 7 (3) (2013) 345–352.
[5] J.D. Churchill, S.E. Schmedes, J.L. King, B. Budowle, Evaluation of the Illumina® beta version ForenSeqTM DNA signature prep kit for use in genetic profiling, Forensic Sci. Int. Genet. 20 (2016) 20–29.
[6] D. Plesivkova, R. Richards, S. Harbison, A review of the potential of the MinIONTM single-molecule sequencing system for forensic applications, Wiley Interdiscip. Rev. Forensic Sci. 1 (1) (2019), e1323.
[7] R.R. Wick, L.M. Judd, K.E. Holt, Performance of neural network basecalling tools for Oxford Nanopore sequencing, Genome Biol. 20 (1) (2019) 1–10.
[8] S. Cornelis, Y. Gansemans, L. Deleye, D. Deforce, F. Van Nieuwerburgh, Forensic SNP genotyping using nanopore MinION sequencing, Sci. Rep. 7 (2017) 41759.
[9] S. Cornelis, Y. Gansemans, A.-S. Vander Plaetsen, J. Weymaere, S. Willems, D. Deforce, F. Van, Nieuwerburgh, Forensic tri-allelic SNP genotyping using nanopore sequencing, Forensic Sci. Int. Genet. 38 (2019) 204–210.
[10] M. Asogawa, A. Ohno, S. Nakagawa, E. Ochiai, Y. Katahira, M. Sudo, M. Osawa, M. Sugisawa, T. Imanishi, Human short tandem repeat identification using a nanopore-based DNA sequencer: a pilot study, J. Hum. Genet. (2019) 1–4.
[11] S. Cornelis, S. Willems, C. Van Neste, O. Tytgat, J. Weymaere, A.-S. Vander Plaetsen, D. Deforce, F. Van Nieuwerburgh, Forensic STR profiling using Oxford Nanopore Technologies' MinION sequencer, bioRxiv (2018), 433151.
[12] O. Tytgat, Y. Gansemans, J. Weymaere, K. Rubben, D. Deforce, F. Van Nieuwerburgh, Nanopore sequencing of a forensic STR multiplex reveals Loci suitable for single-contributor STR profiling, Genes 11 (4) (2020) 381.
[13] Z.-L. Ren, J.-R. Zhang, X.-M. Zhang, X. Liu, Y.-F. Lin, H. Bai, M.-C. Wang, F. Cheng, J.-D. Liu, P. Li, Forensic nanopore sequencing of STRs and SNPs using Verogen's ForenSeq DNA signature prep kit and MinION, Int. J. Leg. Med. (2021) 1–9.
[14] S.E. Van der Verren, N. Van Gerven, W. Jonckheere, R. Hambley, P. Singh, J. Kilgour, M. Jordan, E.J. Wallace, L. Jayasinghe, H. Remaut, A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity, Nat. Biotechnol. 38 (12) (2020) 1415–1420.
[15] J. Amigo, C. Phillips, T. Salas, L.F. Formoso, Á. Carracedo, M. Lareu, pop. STR—an online population frequency browser for established and new forensic STRs, Forensic Sci. Int. Genet. Suppl. Ser. 2 (1) (2009) 361–362.
[16] C.M. Ruitberg, D.J. Reeder, J.M. Butler, STRBase: a short tandem repeat DNA database for the human identity testing community, Nucleic Acids Res. 29 (1) (2001) 320–322.
[17] K.B. Gettings, L.A. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, J. King, W. Parson, C. Phillips, P.M. Vallone, STRSeq: a catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci. Int. Genet. 31 (2017) 111–117.
[18] S. Willuweit, L. Roewer, International Forensic Y Chromosome User Group, Y chromosome haplotype reference database (YHRD): update, Forensic Sci. Int. Genet. 1 (2) (2007) 83–87.
[19] J.M. Butler, A.E. Decker, P.M. Vallone, M.C. Kline, Allele frequencies for 27 Y-STR loci with US Caucasian, African American, and Hispanic samples, Forensic Sci. Int. 156 (2–3) (2006) 250–260.
[20] L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, K.B. Gettings, Sequence-based US population data for 7 X-STR loci, Forensic Sci. Int. Rep. 2 (2020), 100160.
[21] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv Prepr. arXiv 1303 (2013) 3997.
[22] 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 (1) (2001) 308–311.
[23] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079.
[24] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A. Whitwham, T. Keane, S.A. McCarthy, R.M. Davies, Twelve years of SAMtools and BCFtools, Gigascience 10 (2) (2021) giab008.
[25] C.G. Jønck, X. Qian, H. Simayijiang, C. Børsting, STRinNGS v2. 0: improved tool for analysis and reporting of STR sequencing data, Forensic Sci. Int. Genet. 48 (2020), 102331.
[26] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. LaRue, J.L. King, B. Budowle, STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data, Forensic Sci. Int. Genet. 7 (4) (2013) 409–417.
[27] C. Van Neste, M. Vandewoestyne, W. Van Criekinge, D. Deforce, F. Van Nieuwerburgh, My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing, Forensic Sci. Int. Genet. 9 (2014) 1–8.
[28] Q. Liu, P. Zhang, D. Wang, W. Gu, K. Wang, Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing, Genome Med. 9 (1) (2017) 1–16.
[29] J. Hoogenboom, K.J. van der Gaag, R.H. de Leeuw, T. Sijen, P. de Knijff, J.F. Laros, FDSTools: a software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise, Forensic Sci. Int. Genet. 27 (2017) 27–40.
[30] Zi-Lin Ren, Jia-Rong Zhang, Xiao-Meng Zhang, Yan-Feng Lin, Hua Bai, Meng-Chun Wang, Feng Cheng, Jin-Ding Liu, Peng Li, Lei Kong, Xiao Chen, Sheng-Qi Wang, Ming Ni, Jiang-Wei Yan, Forensic nanopore sequencing of STRs and SNPs using Verogen's ForenSeq DNA Signature Prep Kit and MinION, Int J Legal Med . 135 (2021).