International Journal of Legal Medicine (2021) 135:1685–1693
https://doi.org/10.1007/s00414-021-02604-0

ORIGINAL ARTICLE

Forensic nanopore sequencing of STRs and SNPs using Verogen's ForenSeq DNA Signature Prep Kit and MinION

Zi-Lin Ren[1] · Jia-Rong Zhang[2] · Xiao-Meng Zhang[2] · Xu Liu[3] · Yan-Feng Lin[1] · Hua Bai[4] · Meng-Chun Wang[2] · Feng Cheng[2] · Jin-Ding Liu[2] · Peng Li[1] · Lei Kong[4] · Xiao-Chen Bo[1] · Sheng-Qi Wang[1] · Ming Ni[1] · Jiang-Wei Yan[2]

Received: 9 December 2020 / Accepted: 14 April 2021 / Published online: 5 May 2021

Abstract
The MinION nanopore sequencing device (Oxford Nanopore Technologies, Oxford, UK) is the smallest commercially available sequencer and can be used outside of conventional laboratories. The use of the MinION for forensic applications, however, is hindered by the high error rate of nanopore sequencing. One approach to solving this problem is to identify forensic genetic markers that can consistently be typed correctly based on nanopore sequencing. In this pilot study, we explored the use of nanopore sequencing for single nucleotide polymorphism (SNP) and short tandem repeat (STR) profiling using Verogen's (San Diego, CA, USA) ForenSeq DNA Signature Prep Kit. Thirty single-contributor samples and DNA standard material 2800 M were genotyped using the Illumina (San Diego, CA, USA) MiSeq FGx and MinION (with R9.4.1 flow cells) devices. With an optimized cutoff for allelic imbalance, all 94 identity-informative SNP loci could be genotyped reliably using the MinION device, with an overall accuracy of 99.958% (1 error among 2926 genotypes). STR typing was notably error prone, and its accuracy was locus dependent. We developed a custom-made bioinformatics workflow, and finally selected 13 autosomal STRs, 14 Y-STRs, and 4 X-STRs showing high consistency between nanopore and Illumina sequencing among the tested samples. These SNP and STR loci could be candidates for panel design for forensic analysis based on nanopore sequencing.

Keywords ForenSeq DNA Signature Prep Kit · MiSeq FGx™ · Nanopore sequencing · STR analysis · SNP analysis

Introduction

Genetic marker–based profiling is essential for forensic applications such as human identity and kinship testing. Usually, collected samples are transported to laboratories for forensic DNA analysis. When turnaround times are limited, however, such as in situations requiring disaster victim identification, rapid DNA profiling is needed [1]. To address this challenge, fully automated short-tandem repeat (STR) profiling systems, such as the ANDE™ rapid DNA identification system (ANDE, Longmont, CO, USA) [2, 3], RapidHIT® ID system (Thermo Fisher Scientific, Waltham, MA, USA) [4], and ParaDNA® screening system (LGC Group, Teddington, UK) [5, 6], have been developed in the past decade. The ANDE system, for instance, can be used on site by non-technical operators

Zi-Lin Ren and Jia-Rong Zhang these authors wish it be known that in their opinion. Authors are regarded as joint first authors

✉ Sheng-Qi Wang
sqwang@bmi.ac.cn

✉ Ming Ni
niming@bmi.ac.cn

✉ Jiang-Wei Yan
yanjw@sxmu.edu.cn

Zi-Lin Ren
zilin.ren@outlook.com

Jia-Rong Zhang
zhangjiarong_9431@163.com

[1] Beijing Institute of Radiation Medicine, Beijing 100850, People's Republic of China

[2] School of Forensic Medicine, Shanxi Medical University, Taiyuan 030001, People's Republic of China

[3] Beijing Center for Physical and Chemical Analysis, Beijing 100089, People's Republic of China

[4] Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, People's Republic of China

and generates concordant STR allele results within 2 h [7]. However, these systems cannot be used to profile other forensic markers, such as single nucleotide polymorphisms (SNPs), and they are characterized by low-throughput sample processing.

An alternative for rapid on-site DNA profiling is the use of the pocket-sized MinION nanopore sequencing device (Oxford Nanopore Technologies, Oxford, UK). The MinION weighs less than 100 g and has a maximum throughput exceeding 10 gigabase pairs (Gbp), which is theoretically sufficient for the profiling of large numbers of samples. To date, the MinION has been used in many on-site applications, such as the recovery of pathogen genomes from clinical samples in epidemic areas [8–10], species identification in rainforests [11], and genome assembly on the International Space Station [12].

The use of the MinION for forensic genetic analysis, however, is challenging for several reasons. First, the MinION sequencer, whose applications depend on sample preparation and post-sequencing analysis, is not a sample-to-answer system designed for forensic analysis. Second, nanopore sequencing is more error prone than Sanger and massively parallel sequencing [13]. In the few pilot studies conducted to assess the accuracy of STR profiling with the MinION, in very limited samples, error rates were high [14–16]. Nonetheless, Tytgat et al. [15] reported that MinION-based STR typing errors were limited largely to specific loci; some loci were typed correctly in all three samples in their study. This finding reflects the characteristic systematic nature of some nanopore sequencing errors [17, 18], and the need to evaluate the MinION's performance for more STR loci and with more samples.

The genotyping of human SNPs by nanopore sequencing is more accurate than STR genotyping [19]. Identity-informative SNPs are considered to be supplementary markers of STRs in identity and kinship testing [20–24]. Nanopore sequencing has been used to profile a 52-SNP panel developed by the SNPforID consortium and a panel of 16 tri-allelic SNPs in a few DNA standards and individual samples [25, 26], with several problematic SNP loci characterized.

In this pilot study, we examined more STRs and SNPs as candidate markers suitable for forensic nanopore sequencing than have been assessed in previous studies [14–16, 25, 26]. We used ForenSeq DNA Signature Prep Kit mix A (Verogen, USA), which includes 54 STRs and 94 SNPs for amplification, and then performed nanopore sequencing using the MinION device. Although multiplex PCR amplification is used commonly in on-site nanopore sequencing [8–10], the ForenSeq DNA Signature Prep Kit is not suitable for on-site use. Further study is needed to establish an on-site protocol for forensic marker amplification and subsequent library preparation.

Materials and methods

Sample collection and DNA extraction

This study was approved by the ethics committee of Shanxi Medical University (no. 2020GLL031), and all participants provided informed consent. Blood samples were collected from 30 Han Chinese volunteers (15 females [F1–F15] and 15 males [M1–M15]). The samples were anonymized, and DNA was extracted from whole blood using the PureLink genomic DNA kit (Invitrogen, USA) according to the manufacturer's instructions. The DNA was stored at –80 °C and quantified using a Qubit Fluorometer (Invitrogen) before use. The control DNA sample of the ForenSeq DNA Signature Prep Kit (2800 M; Verogen) was used in this study.

SNP and STR amplification and sequencing

PCR amplification was performed with the ForenSeq kit's DNA primer mix A (Verogen) according to the manufacturer's instructions, except that the input was increased from the recommended 1 to 5 ng to yield sufficient PCR products for comparative MinION nanopore and Illumina (San Diego, CA, USA) MiSeq FGx sequencing. Approximately 500 ng of DNA was obtained after amplification and normalization. The final barcoded libraries were pooled and sequenced using the Illumina MiSeq FGx system.

Library preparation for MinION sequencing

Aliquots of normalized libraries generated with the ForenSeq DNA signature prep kit were used as input for further library preparation and sequencing with the MinION device (Oxford Nanopore Technologies). As the amplicons were short (400–600 bp), we adopted the Quick [27] protocol recommended by the manufacturer for the construction of short insert libraries (https://community.nanoporetech.com/protocols/pcr-tiling-SARS-CoV-2/v/PTC_9096_v109_revA_06Feb2020). This protocol is based on the ARTIC protocol for MinION sequencing of severe acute respiratory syndrome coronavirus 2, which has a comparable amplicon size. Briefly, 0.2 pmol cleaned PCR products per sample were used in the end-repair and dA-tailing steps. We calculated sample volumes for 0.2 pmol DNA using an online calculator provided by Oxford Nanopore Technologies (https://nanoporetech.ent.box.com/s/rzlbor9yl7jll9vkqdwgl892cesxeshp), with DNA concentrations quantified with the Qubit 3.0 fluorometer (Thermo Fisher Scientific) and a fragment length of 450 bp serving as inputs. Then, the DNA was barcoded with a 32-bp index using the native barcoding kit (Oxford Nanopore Technologies). The adapters for nanopore sequencing from the ligation sequencing kit (Oxford Nanopore Technologies) were ligated to the pooled and barcoded DNA using the NEBNext quick ligation module (New England Biolabs, USA). Finally, 15 ng of the final library was loaded onto an R9.4.1 flow cell for MinION sequencing. Guppy (v3.6.0; Oxford Nanopore Technologies) was employed for base calling and demultiplexing.

SNP and STR genotyping based on Illumina sequencing data

For SNP and STR typing, the MiSeq FGx data were analyzed using the ForenSeq universal analysis software (v1.3.6897; Verogen) with the default settings (estimated minimum analytical threshold ≥ 10 reads, minimum interpretation threshold > 30 reads). Some of software-reported SNP alleles that were reverse complements of those recorded in the National Center for Biotechnology Information's dbSNP database (build 141; https://www.ncbi.nlm.nih.gov/snp/) [28] were modified manually according to the dbSNP entries. A quality control indicator of a genotyped SNP or STR locus can be obtained in the sample report including "Imbalanced" (the number of reads supporting alleles are imbalanced), "Allele Count" (the number of alleles above analytical threshold than excepted), "Stutter"(the number of reads supporting allele at a stutter position is beyond the stutter filter percentage), "Interpretation threshold"(at least one allele that is not stutter is above the analytical threshold but below the interpretation threshold), "Not Detected"(No reads were detected for the locus), and/or "User Actions" (at least one allele was manually edited as typed or not typed, or a user comment was submitted).

SNP genotyping based on MinION sequencing data

The MinION sequencing reads were first aligned to the autosomes and X and Y chromosomes of the human reference genome (assembly GRCh37, hg19) using Minimap2 (v2.17r941) [29]. The output SAM files of the alignments were transformed to BAM format using SAMtools (v1.9) [30]. The scaffolds in the hg19 assembly were not included as reference sequences for alignment, as the unlocalized scaffold chr17_gl000204_random could induce misalignment of reads supporting allele C of and incorrect genotyping rs938283 (Fig. S1).

SNP genotyping was performed using the Medaka software (v1.0.3; Oxford Nanopore Technologies; https://github.com/nanoporetech/medaka) [31]. The medaka_variant program was used to perform small variant calling with the r941_min_high_g360 model. For SNP loci without Medaka variant output, the corresponding genotypes were considered to be homozygous for the reference allele.

In parallel, we developed a custom-made bioinformatics workflow (https://github.com/Ming-Ni-Lab/NanoForenSNP). Mpileup files were first generated for the BAM alignment files for the 94 SNP loci using SAMtools (v1.9) [30]. Then, the numbers of reads supporting the reference (n_r) and alternative (n_a) alleles were obtained for each locus. To evaluate allelic imbalance, we calculated the minor allele frequency (minAF; the relative frequency of the allele with second most supporting reads): $min\{n_r / (n_r + n_a), n_a /(n_r + n_a)\}$. This frequency was used to determine whether the SNP locus was homozygous or heterozygous.

STR genotyping based on MinION sequencing data

STR genotyping was performed using the repeatHMM software (v2.0.2) [32] with default parameters and a custom-made bioinformatics workflow written in Python3 (https://github.com/Ming-Ni-Lab/NanoForenSTR). In the custom-made workflow, reads generated by MinION were aligned to the autosomes and X and Y chromosomes (assembly GRCh37, hg19) using Minimap2 (v2.17r941) [29] and transformed to BAM format using SAMtools (v1.9) [30]. The Pysam Python module (https://github.com/pysam-developers/pysam) was used to find reads covering the repeat regions of the STR loci and upstream and downstream flanking regions of at least 20 bp. The flanking regions were trimmed, and only the repeat regions were extracted for the identification of repeat units. Repeat units that exactly matched the expectation were identified first. For the other STR core regions, a Smith–Waterman algorithm was used to further distinguish repeat units based on the requirements of aligned length > 60% of the unit repeat and > 60% identity. Finally, the counts of supporting reads for STR alleles were summarized and reported. Incomplete repeats (alleles with decimal points) could not be reported with the custom-made workflow.

SNP validation by Sanger sequencing

We used Sanger sequencing for the genotyping of SNP loci rs1493232, rs338882, and rs1357617. Primer3 (v0.4.0) [33, 34] was used to design the PCR primer pairs for rs1493232 and rs338882. The PCR primers for rs1357617 were from Zhang et al. [35]. The primer pair sequences are provided in Table S1. PCR was performed with TaKaPa Taq hot start version (Takara Bio Company, Japan) and the following thermal cycling regimen: 180 s at 98 °C, followed by 35 cycles of 30 s at 98 °C, 30 s at 62 °C, and 45 s at 72 °C. Sanger sequencing was conducted by Sangon Biotech (Shanghai, China).

Statistical analysis

Linear correlations of locus depths based on MiSeq FGx and MinION sequencing were determined by Pearson correlation (r). The distributions of repeat count deviations from true alleles at STR loci were illustrated as bubble plots, in which the size is indicated in the respective figures (R: ggplot2), and in which the bubble size represents the number of samples with the corresponding deviation.

Results

Multiplex PCR and sequencing using MiSeq FGx and MinION

We collected whole blood samples from 30 individuals in this study. DNA was extracted as single-contributor samples and multiplex PCR was performed using the ForenSeq DNA primer mix A kit, which resulted in the amplification of genomic regions harboring 94 identity-informative SNPs, 27 autosomal STRs, 24 Y-STRs, and 7 X-STRs. Together with the three replicates of 2800 M DNA standard material 2800 M with 3 replicates, a total of 33 samples was sequenced in a batch with the MiSeq FGx system and with two R9.4.1 flow cells on a MinION sequencer.

MiSeq FGx sequencing yielded a total of 1.55 Gbp data and 34.45–65.86 mega base pairs (Mbp) per sample. The amount of data per sample obtained with MinION sequencing ranged from 108.7 to 188.15 Mbp. For each sample, we obtained the normalized sequencing depths at the SNP loci by dividing the locus-specific depth by the mean depth of the sample. The sample depth profiles based on MiSeq FGx and nanopore sequencing were correlated (Pearson correlation, $r = 0.87$; Fig. S2). Details of the sequencing are provided in Table S2.

MinION SNP profiling

Based on the MiSeq FGx sequencing data, 176 SNPs in the 33 samples had insufficient coverage and were labeled with an "interpretation threshold" by the ForenSeq universal analysis software. Fifty-five of these SNPs had reported genotypes; examination of them suggested that the software's filter was strict, leading to the disregarding of an allele with a depth ≤ 30 when the depth of the other allele was > 30. This led to incorrect genotyping at some heterozygous loci with insufficient coverage. For instance, sample M3 had 30 reads at rs1357617 supporting allele C and 35 reads supporting allele T, but a homozygous TT genotype was reported. With Sanger sequencing, we validated 16 genotypes (rs1357617 in five samples, rs1493232 in four samples, and rs338882 in seven samples) which were reported as homozygous by the ForenSeq universal analysis software, and they were all validated as heterozygous (Table S3). Thus, we excluded SNPs with the "interpretation threshold" label and used the remaining 2926 SNP genotypes from MiSeq FGx sequencing to assess MinION-based SNP profiling.

SNP genotyping based on MinION sequencing was first conducted using the widely used Medaka software (v1.0.3) [31]. However, we found that only 53–55% of SNP loci in the three 2800 M replicates had genotypes corresponding to the reference profile. The errors were attributed to allelic imbalance, with incorrect classification of genotypes as homozygous or heterozygous, indicating the need to optimize the allelic imbalance cutoff for MinION sequencing. We evaluated allelic imbalance by calculating minAFs for loci, which should be 0 for homozygous and 0.5 for heterozygous genotypes. Although the MinION data exhibited more allelic imbalance than did the MiSeq FGx data, homozygous and heterozygous genotypes were distinguishable by minAFs based on the MinION data. All 1183 heterozygous SNPs had minAFs > 0.15, and all but 1 (rs914165 in sample M10) of the 1743 homozygous SNPs had minAFs < 0.1 (Fig. 1). The application of a minAF cutoff of 0.15 yielded a MinION SNP profiling accuracy rate of 99.958% (2925 of 2926 SNPs). The single typing error for rs914165 in sample M10 may be related to the low sequencing depth; 14 non-allelic bases (As) were found, with a total depth of 83-fold. We also observed false-negative typing of an allele at a locus with low coverage (rs1357617 in M15 [Table S3]) labeled with the "interpretation threshold" and validated as heterozygous by Sanger sequencing.

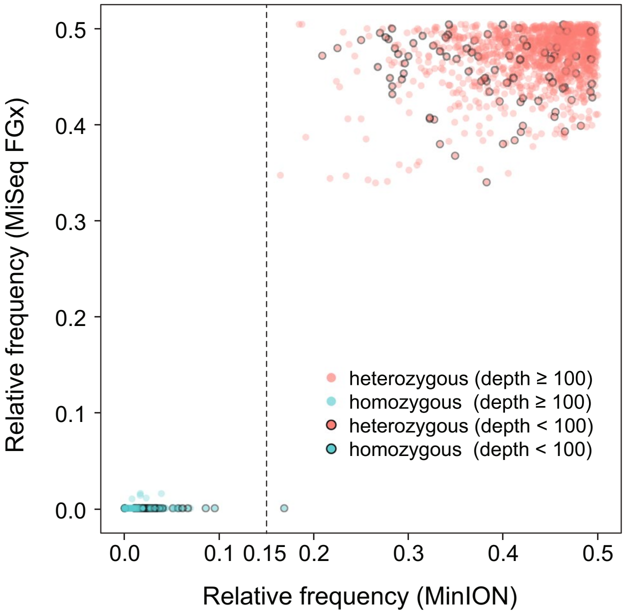

Fig. 1 Allelic imbalance of SNPs based on MinION and Illumina MiSeq FGx sequencing. Relative frequencies of alleles with fewer supporting reads at each of the 94 biallelic SNP loci in the 30 samples and 2800 M replicates are shown. Heterozygous (red) and homozygous (blue) SNPs were identified based on MiSeq FGx sequencing. SNPs labeled with the "interpretation threshold" by the ForenSeq universal analysis software were excluded

The frequencies of non-allelic bases (misincorporation rates) reflect baseline noise during SNP profiling. Among the 2573 SNPs with ≥ 100-fold MinION sequencing depths, 930 (36.2%) had non-allelic bases, with misincorporation rates ranging from 0.00053 to 0.07826 (mean, 0.0076; standard deviation, 0.0083; median, 0.0050). Two loci (rs873196 and rs13218440) had maximum misincorporation rates > 5%. The MinION-based misincorporation rates are summarized by SNP locus in Table S4. Compared with the MinION data, the MiSeq FGx data had much less noise; non-allelic bases were observed at only five (0.19%) SNPs, with misincorporation rates ranging from 0.0097 to 0.015.

Read-level noise of STR typing using MinION

To assess MinION STR sequencing, MiSeq FGx–based STR profiles were used as a reference, and STRs labeled by the ForenSeq universal analysis software with "imbalanced," "allele count," "stutter," and/or "interpretation threshold" were excluded. DY385a and DY385b were also excluded, as the MiSeq FGx universal software did not distinguish them. Another three Y-STRs with no exact genomic coordinates (DYF387S1, DYS389I, and DYS389II) were excluded because our STR typing analysis required the alignment of nanopore sequencing reads to the human reference genome with coordinates.

We investigated the characteristics of MinION sequencing noise at the read level. The homozygous autosomal STRs identified by MiSeq FGx sequencing were extracted, enabling clear definition of the deviation of repeat counts between observed and true alleles. As expected, the extent of deviation depended on the STR locus. The 27 autosomal STR loci could be divided into 2 groups (A and B) based on the read-level noise patterns (Fig. S3). At the loci assigned to group A, true STR alleles were identified with the highest coverage, and the deviating repeats represent stutters and other artifacts due to errors in sequencing and analysis. Group A contained 16 autosomal STR loci (D2S441, D3S1358, D4S2408, D5S818, D6S1043, D7S820, D8S1179, D10S1248, D13S317, D16S539, D17S1301, D18S51, D20S482, D22S1045, TPOX, and CSF1PO). Among them, D18S51 and especially D10S1248 had much higher stutter rates than did the other loci (Fig. S3). At autosomal loci assigned to group B, the alleles with the highest coverage were false positives, deviating from the true alleles with –9 to 2 repeat counts. Among the X and Y STR loci, also grouped according to the above-described criteria, 14 of the 21 Y-STRs (DYS391, DYS392, DYS438, DYS439, DYS460, DYS522, DYS533, DYS549, DYS570, DYS635, YGATAH4, DYS437, DYS612, and DYS505) and 4 of the 7 X-STRs (DXS8378, HPRTB, DXS10135, and DXS7132) were assigned to group A (Fig. S4 and S5). The remaining STRs were assigned to group B.

MinION STR profiling

We adopt the following criteria to type STRs at the locus level in diploid chromosomes. For each such locus, the two alleles with the most abundant supporting reads were extracted, and all other observed alleles and stutters were disregarded. We calculated minAFs for the two alleles to quantify allelic imbalance, as for SNP typing. Then, we employed an optimized minAF cutoff of 0.36 (Fig. S6) to determine whether the locus was homozygous or heterozygous. When reads supporting the less-covered repeat exceeded 36% of all reads for the two most frequent repeats, the STR was determined to be heterozygous; otherwise, it was considered to be homozygous. For loci on haploid chromosomes, STR typing yielded the alleles with the highest coverage. As we obtained a minimum depth of 214-fold for the STRs by MinION sequencing, an arbitrary requirement of ≥ 200-fold depth was applied to STR profiling.

With MinION sequencing data, both alleles in 11 of the 16 autosomal STR loci in group A were typed correctly in all samples (Fig. 2a). Three other loci in group A (D6S1043, D20S482, and D13S317) had high accuracy rates, but false-positive alleles in one or two samples. Due to the high stutter rate, the typing of D18S51 and D10S1248 was incorrect for one allele in 13 and 32 samples, respectively. No autosomal STR assigned to group A had incorrect typing of both alleles. In contrast, most (277 of 348, 79.6%) autosomal STR genotypes for loci in group B were incorrect for both alleles, and the remaining genotypes were incorrect for one allele. The distributions of deviation at autosomal STRs among these samples are shown in Fig. 2b. The deviations for FGA, vWA, and D9S1122 were identical in most samples. As no software designed specifically for forensic STR typing based on nanopore sequencing is currently available, we conducted locus-level typing using repeatHMM (v2.0.2) [32], a computational tool developed to detect microsatellites from long reads, for comparison. Our custom-made workflow was more accurate than repeatHMM for STR typing (Fig. S7).

Similar to the autosomal STR results, typing of the 14 Y-STRs and 4 X-STRs assigned to group A was very consistent with the reference profiles (Fig. 3a). All X- and Y-STRs in group A were typed correctly, except for one error at DXS7132 in sample F01. Typing of the 6 Y-STRs and 3 X-STRs in group B was error prone, whereas all of these STRs exhibited identical deviations in most of the samples (Fig. 3b).

Conclusion and discussion

In this pilot study, we evaluated the performance of nanopore sequencing for forensic SNP and STR profiling in samples from 30 individuals and the 2800 M DNA standard. The

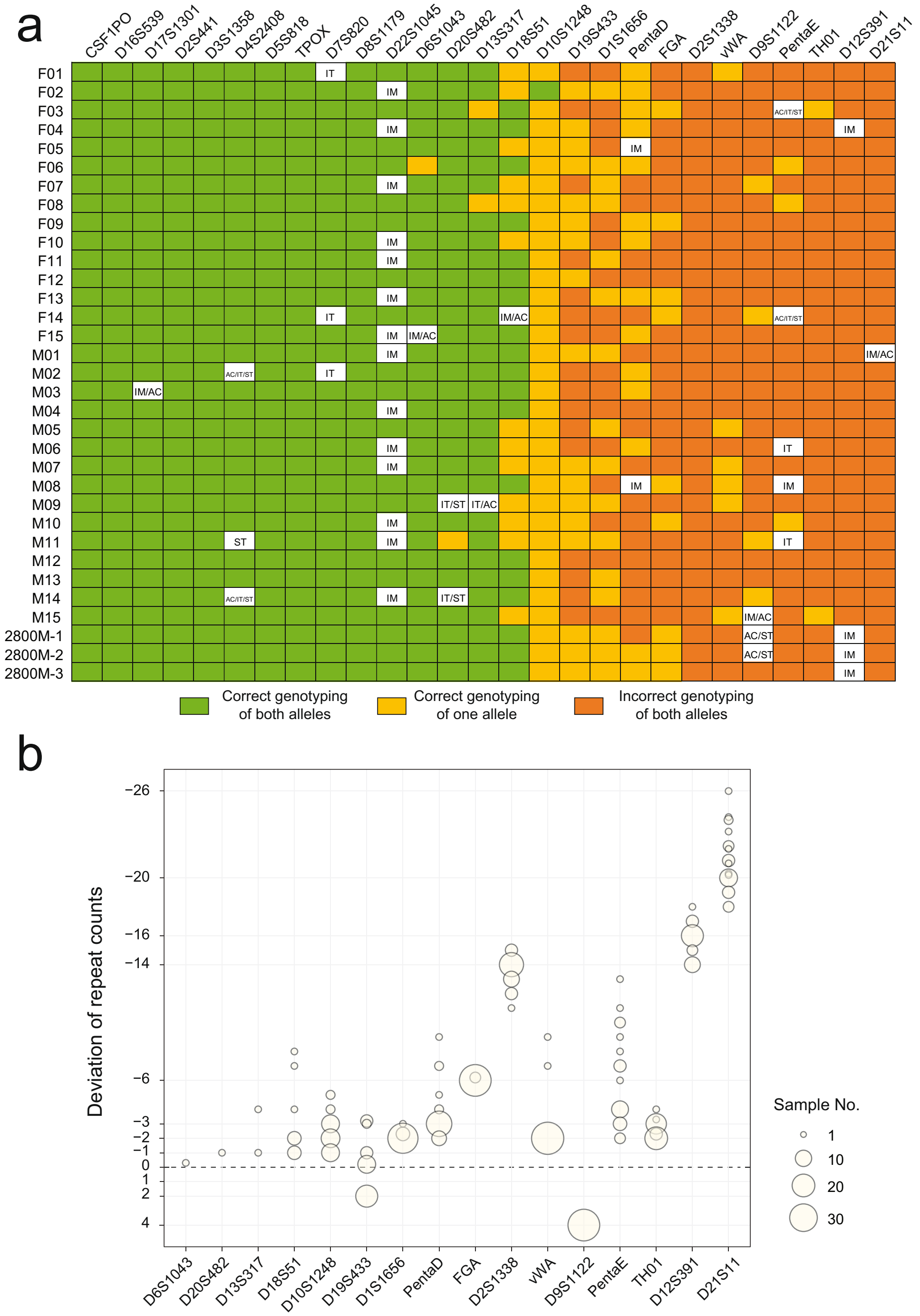

Fig. 2 Genotyping of autosomal STRs based on MinION sequencing. (**a**) Consistency of STR genotyping of 30 samples and 2800 M replicates between MinION and MiSeq FGx sequencing. Imbalanced (IM), allele count (AC), stutter (ST), and interpretation threshold (IT) are warning labels based on MiSeq FGx sequencing applied by the ForenSeq universal analysis software. (**b**) Deviations of repeat counts from true alleles in the 30 samples and three 2800 M replicates based on MinION sequencing. For heterozygous STRs, the sum of deviation for both alleles was calculated

SNPs and STRs targeted by primer mix A of the ForenSeq DNA Signature Prep Kit (Verogen) were amplified and sequenced using MinION and MiSeq FGx. Custom-made scripts were developed for SNP and STR profiling based on MinION sequencing, and the profiles generated by MiSeq FGx sequencing were used as reference for the assessment. Despite the high error rate in MinION sequencing and analysis, we determined that all 94 SNP loci, 14 autosomal STRs (D2S441, D3S1358, D4S2408, D5S818, D6S1043, D7S820, D8S1179, D13S317, D16S539, D17S1301, D20S482, D22S1045, TPOX, and CSF1PO), 14 Y-STRs (DYS391, DYS392, DYS438, DYS439, DYS460, DYS522, DYS533, DYS549, DYS570, DYS635, YGATAH4, DYS437, DYS612, and DYS505), and 4 X-STRs (DXS8378, HPRTB, DXS10135, and DXS7132) could be typed accurately in these samples.

The characterization of locus-dependent errors deriving from nanopore sequencing is fundamental. As current

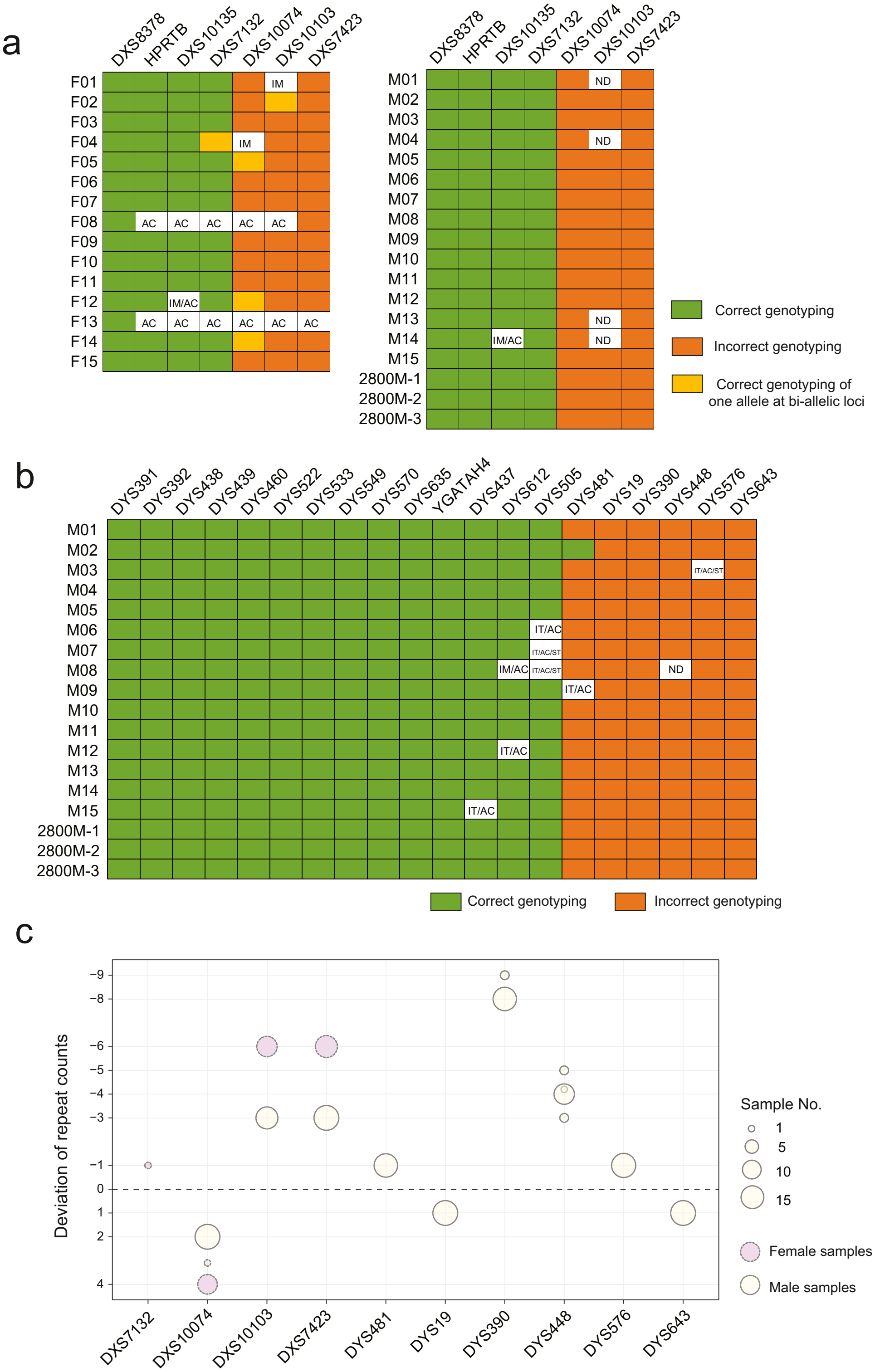

Fig. 3 X-STR and Y-STR genotyping based on MinION sequencing. (**a**, **b**) Consistency of X-STR and Y-STR genotyping, respectively, between MinION and MiSeq FGx sequencing. Female and male samples are shown in the left and right panels, respectively, of (**a**). Imbalanced (IM), allele count (AC), stutter (ST), interpretation threshold (IT), and not detected (ND) are warning labels based on MiSeq FGx sequencing applied by the ForenSeq universal analysis software. (**c**) Deviations of repeat counts from true alleles in the female and male samples

software tools for nanopore sequencing–based variant calling, such as Medaka, are not optimized for forensic SNP or STR analysis, we employed custom-made scripts for genotyping and investigated noise at the read and locus levels. For MinION sequencing–based SNP profiling, the use of an optimal cutoff for allelic imbalance was critical. MinION sequencing generated more serious allelic imbalance than did MiSeq FGx sequencing. About one-third of SNPs among samples had non-allelic bases, with an average misincorporation rate of 0.76% and a maximum rate of 7.8% (when the depth threshold was set to > 100-fold). However, we identified a minAF cutoff of 0.15 that distinguished heterozygous and homozygous genotypes. Only 1 of the 2926 SNPs assessed was typed incorrectly with MinION sequencing. This error occurred at rs914165 in sample M10, which had an abnormally high non-allelic rate (16.9%, 14 of 83); the maximum non-allelic rates of rs914165 in other samples and of other SNP loci in M10 were low (< 2%). The application of an ad hoc > 100-fold requirement for depth in genotyping eliminated this error. Nevertheless, further investigation with larger samples is needed to determine appropriate criteria for nanopore sequencing for forensic SNP profiling.

The nanopore sequencing–based profiling of STRs was notably more error prone than was SNP profiling. Short insertion and deletion errors were prevalent in the reads generated by the MinION, which hampered the recognition of STR repeat units. With a custom-made bioinformatics workflow based on the Smith–Waterman algorithm, we investigated read-level noise and classified the STRs into two groups accordingly. At group A loci except D18S51 and D10S1248, the true repeats were the most frequent. Consequently, the nanopore sequencing–based profiling of group A STRs was highly accurate. In contrast, the loci assigned to group B were typed incorrectly in most samples. More STRs with compound and complex repeats were assigned to group B (6 of 11 autosomal STRs) than to group A (4 of 16 STRs). However, compound and complex repeats were not the only determinants of group allocation [36]. Simple repeats D9S1122 (repeat region sequence, [TGAG]n), D13S31 ([TATC]n), and D20S482 ([AGAT]n) were assigned to group B, indicating that further improvement of this type of bioinformatics analysis is needed. One approach to such improvement would be to explore locus-dependent cutoffs for repeat recognition, which requires a large sample size. On the other hand, the high typing error rates at Penta D ([AAAGA]n) and Penta E ([TTTTC]n) might be ascribed to the presence of 5-mer and 4-mer homopolymers in the repeat region. The base calling of homopolymers is a major issue in nanopore sequencing [37].

This study has several limitations. First, the individual samples were only collected from Han Chinese individuals. The sample representativeness could influence the assessment of forensic profiling of the SNPs and STRs. Second, although we included more samples than were included in relevant previous studies, the sample was still small and reproducibility was investigated only for the triplicates of the 2800 M DNA standard. This factor limited the establishment of bioinformatics criteria and may have biased the estimation of locus-dependent error rates. Third, we conducted the study in conventional forensic and molecular biology laboratories, and did not use the MinION field sequencing kit manufactured by Oxford Nanopore Technologies. Further research is needed to establish an on-site protocol for multiplex PCR and library construction with limited experimental equipment, which was beyond the scope of this study. Fourth, we did not include degraded or minute DNA samples in this study. Although Verogen's ForenSeq DNA Signature Prep Kit can be used with these types of sample, the effects of this practice on nanopore sequencing–based profiling are unclear.

This pilot study explored the efficacy of forensic marker profiling based on nanopore sequencing. With custom-made bioinformatic analyses, we identified 94 SNP and 32 STR loci that could be typed reliably based on MinION sequencing. The results reflect the potential of nanopore sequencing for forensic applications, such as human identity and paternity testing. These loci may be candidates for panel design for such analyses. However, further improvement of the bioinformatic workflow and validation with large samples are needed.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00414-021-02604-0.

Funding This work was supported by the National Natural Science Foundation of China (82030058). M.N was supported by the Beijing Nova Program (Z181100006218114).

Data Availability The nanopore sequencing data (fastq) is available at https://bigd.big.ac.cn/gsa-human/browse/HRA000534.

Code Availability Custom-made scripts for nanopore sequencing-based SNP genotyping are available at https://github.com/Ming-Ni-Lab/NanoForenSNP. Custom-made scripts for nanopore sequencing-based STR genotyping are available at https://github.com/Ming-Ni-Lab/NanoForenSTR.

Declarations

Conflict of interest The authors declare no competing interests.

Ethics approval The project and informed consent were approved by the Ethical Committee of Shanxi Medical University (Issued Number: 2020GLL031) and the ethical principles of the 2000 Helsinki Declaration of the World Medical Association (http://www.uma.net/e/policy/b3.htm) were followed.

References

1. Morrison J, Watts G, Hobbs G, Dawnay N (2018) Field-based detection of biological samples for forensic analysis: established techniques, novel tools, and future innovations. Forensic Sci Int 285:147–160
2. French JL, Turingan RS, Hogan C, Selden RF (2016) Developmental validation of the DNAscan^{TM} Rapid DNA Analysis^{TM} instrument and expert system for reference sample processing. Forensic Sci Int Genet 25:145–156. https://doi.org/10.1016/j.fsigen.2016.08.008
3. Turingan RS, Tan E, Jiang H et al (2020) Developmental validation of the ANDE 6C system for rapid DNA analysis of forensic casework and DVI samples. J Forensic Sci 65:1056–1071. https://doi.org/10.1111/1556-4029.14286

4. Salceda S, Barican A, Buscaino J et al (2017) Validation of a rapid DNA process with the RapidHIT® ID system using GlobalFiler® Express chemistry, a platform optimized for decentralized testing environments. Forensic Sci Int Genet 28:21–34. https://doi.org/10.1016/j.fsigen.2017.01.005
5. Dawnay N, Stafford-Allen B, Moore D et al (2014) Developmental validation of the ParaDNA® screening system - a presumptive test for the detection of DNA on forensic evidence items. Forensic Sci Int Genet 11:73–79. https://doi.org/10.1016/j.fsigen.2014.02.004
6. Ball G, Dawnay N, Stafford-Allen B et al (2015) Concordance study between the ParaDNA® intelligence test, a rapid DNA profiling assay, and a conventional STR typing kit (AmpFlSTR® SGM Plus®). Forensic Sci Int Genet 16:48–51. https://doi.org/10.1016/j.fsigen.2014.12.006
7. Carney C, Whitney S, Vaidyanathan J et al (2019) Developmental validation of the ANDE^{TM} rapid DNA system with FlexPlex^{TM} assay for arrestee and reference buccal swab processing and database searching. Forensic Sci Int Genet 40:120–130. https://doi.org/10.1016/j.fsigen.2019.02.016
8. Quick J, Grubaugh ND, Pullan ST et al (2017) Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc 12:1261–1276. https://doi.org/10.1038/nprot.2017.066
9. Quick J, Loman NJ, Duraffour S et al (2016) Real-time, portable genome sequencing for Ebola surveillance. Nature 530:228–232. https://doi.org/10.1038/nature16996
10. Russell JA, Campos B, Stone J et al (2018) Unbiased strain-typing of arbovirus directly from mosquitoes using nanopore sequencing: a field-forward biosurveillance protocol. Sci Rep 8:5417. https://doi.org/10.1038/s41598-018-23641-7
11. Menegon M, Cantaloni C, Rodriguez-Prieto A et al (2017) On site DNA barcoding by nanopore sequencing. PLoS ONE 12:e0184741. https://doi.org/10.1371/journal.pone.0184741
12. Castro-Wallace SL, Chiu CY, John KK et al (2017) Nanopore DNA sequencing and genome assembly on the International Space Station. Sci Rep 7:18022. https://doi.org/10.1038/s41598-017-18364-0
13. Laver T, Harrison J, O'Neill PA et al (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif 3:1–8. https://doi.org/10.1016/j.bdq.2015.02.001
14. Cornelis S, Willems S, Van Neste C et al (2018) Forensic STR profiling using Oxford Nanopore Technologies' MinION sequencer. bioRxiv 433151. https://doi.org/https://doi.org/10.1101/433151
15. Tytgat O, Gansemans Y, Weymaere J et al (2020) Nanopore sequencing of a forensic STR multiplex reveals loci suitable for single-contributor STR profiling. Genes (Basel) 11:381. https://doi.org/10.3390/genes11040381
16. Asogawa M, Ohno A, Nakagawa S et al (2020) Human short tandem repeat identification using a nanopore-based DNA sequencer: a pilot study. J Hum Genet 65:21–24. https://doi.org/10.1038/s10038-019-0688-z
17. Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17:239. https://doi.org/10.1186/s13059-016-1103-0
18. Kono N, Arakawa K (2019) Nanopore sequencing: review of potential applications in functional genomics. Dev Growth Differ 61:316–326
19. Jain M, Koren S, Miga KH et al (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36:338–345. https://doi.org/10.1038/nbt.4060
20. Pakstis AJ, Speed WC, Fang R et al (2010) SNPs for a universal individual identification panel. Hum Genet 127:315–324. https://doi.org/10.1007/s00439-009-0771-1
21. Sanchez JJ, Phillips C, Børsting C et al (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis 27:1713–1724. https://doi.org/10.1002/elps.200500671
22. Børsting C, Fordyce SL, Olofsson J et al (2014) Evaluation of the Ion Torrent^{TM} HID SNP 169-plex: a SNP typing assay developed for human identification by second generation sequencing. Forensic Sci Int Genet 12:144–154. https://doi.org/10.1016/j.fsigen.2014.06.004
23. Mo SK, Ren ZL, Yang YR et al (2018) A 472-SNP panel for pairwise kinship testing of second-degree relatives. Forensic Sci Int Genet 34:178–185. https://doi.org/10.1016/j.fsigen.2018.02.019
24. Mo SK, Liu YC, Wang SQ et al (2016) Exploring the efficacy of paternity and kinship testing based on single nucleotide polymorphisms. Forensic Sci Int Genet 22:161–168. https://doi.org/10.1016/j.fsigen.2016.02.012
25. Cornelis S, Gansemans Y, Deleye L et al (2017) Forensic SNP genotyping using nanopore MinION sequencing. Sci Rep 7:41759. https://doi.org/10.1038/srep41759
26. Cornelis S, Gansemans Y, Vander Plaetsen AS et al (2019) Forensic tri-allelic SNP genotyping using nanopore sequencing. Forensic Sci Int Genet 38:204–210. https://doi.org/10.1016/j.fsigen.2018.11.012
27. Quick J (2020) nCoV-2019 sequencing protocol v2 (GunIt). protocols.io. https://doi.org/10.17504/protocols.io.bdp7i5rn
28. Sherry ST, Ward MH, Kholodov M et al (2001) DbSNP: The NCBI database of genetic variation. Nucleic Acids Res 29:308–311. https://doi.org/10.1093/nar/29.1.308
29. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191
30. Li H, Handsaker B, Wysoker A et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
31. nanoporetech/medaka: sequence correction provided by ONT Research. https://github.com/nanoporetech/medaka. Accessed 4 Jan 2021
32. Liu Q, Zhang P, Wang D et al (2017) Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. Genome Med 9:1–16. https://doi.org/10.1186/s13073-017-0456-7
33. Koressaar T, Remm M (2007) Enhancements and modifications of primer design program Primer3. Bioinformatics 23:1289–1291. https://doi.org/10.1093/bioinformatics/btm091
34. Untergasser A, Cutcutache I, Koressaar T et al (2012) Primer3-new capabilities and interfaces. Nucleic Acids Res 40:1–12. https://doi.org/10.1093/nar/gks596
35. Zhang S, Bian Y, Zhang Z et al (2015) Parallel analysis of 124 universal SNPs for human identification by targeted semiconductor sequencing. Sci Rep 5:1–9. https://doi.org/10.1038/srep18683
36. Gettings KB, Aponte RA, Vallone PM, Butler JM (2015) STR allele sequence variation: current knowledge and future issues. Forensic Sci Int Genet 18:118–130. https://doi.org/10.1016/j.fsigen.2015.06.005
37. Lu H, Giordano F, Ning Z (2016) Oxford nanopore MinION sequencing and genome assembly. Genomics Proteomics Bioinformatics 14:265–279. https://doi.org/10.1016/j.gpb.2016.05.004