ABSTRACT

Forensic DNA examinations harness the high degree of repeat length variation characteristic of short tandem repeats (STRs) for human identification. Conventional approaches to STR profiling consist of PCR amplification followed by length-based separation and detection via capillary electrophoresis (CE). These well-established methods are used in forensic laboratories throughout the world to generate robust and reliable profiles that can discriminate between individuals based on differences in STR repeat length alone. The power of discrimination achieved with length-based allele designations across established panels of autosomal and Y-STRs is often sufficient for routine DNA examinations. However, nucleotide-level variation within and around STRs has been shown to increase resolution and facilitate interpretation in more challenging casework scenarios such as those involving partial and mixed DNA profiles.

The MinION is a DNA sequencer from Oxford Nanopore Technologies (ONT) that is small in both size and price tag. This portable device could provide an alternative for STR sequencing in forensic laboratories that cannot afford the initial investment or commitment of common next-generation sequencing (NGS) platforms. Despite this potential, the relatively high error rate and lack of STR analysis software have precluded accurate forensic profiling with nanopore sequencing in previous studies. This project aims to determine whether STRs amplified with a commercial kit can be sequenced and profiled on the ONT MinION device. To achieve our overall objective, we developed and tested a novel bioinformatic method known as STRspy that is designed to produce forensic STR profiles from third-generation sequencing data. The results presented herein demonstrate that STRspy can predict the correct sequence- and length-based allele designations across an entire panel of autosomal and Y-STRs using error-prone ONT reads as well as detect variation in the flanking regions with a high level of accuracy. Moreover, these data provide novel insight into how PCR-induced stutter and sample multiplexing impact STR profiling on the MinION. Ultimately, this work increases the feasibility of nanopore sequencing in forensic investigations and provides the foundation for future efforts that aim to harness the big potential of the small MinION device.

KEYWORDS

forensic DNA analysis, STR, SNP, nanopore sequencing, MinION

STRSPY-ING HIDDEN VARIATION IN FORENSIC DNA PROFILES WITH THE OXFORD NANOPORE TECHNOLOGIES MINION DEVICE

DISSERTATION

Presented to the Graduate Council of the

University of North Texas Health Science Center at Fort Worth

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

Courtney L. Hall, M.S.

November 11, 2022

TABLE OF CONTENTS

CHAPTER 1

HIDE & SEQ: FINDING NUCLEOTIDE-LEVEL VARIATION IN FORENSIC STR PROFILES **1**

Forensic STRs 2
Nanopore Sequencing 10
Project Overview 15

CHAPTER 2

ACCURATE PROFILING OF FORENSIC AUTOSOMAL STRS USING THE OXFORD NANOPORE TECHNOLOGIES MINION DEVICE **17**

Chapter Overview 18
Materials & Methods 19
Results 24
Discussion 34
Conclusion 38
Supplemental Figures 40
Supplemental Files 43
Supplemental Tables 43

CHAPTER 3

EXPANDING THE CAPABILITIES OF STRSPY TO PROFILE MARKERS ON THE Y **47**

Chapter Overview 48
Materials & Methods 49
Results 54
Discussion 58
Concluding Remarks 62
Supplemental Figures 62
Supplemental File 67
Supplemental Tables 67

CHAPTER 4

THE MINION DEVICE: A SMALL SEQUENCER WITH BIG POTENTIAL IN FORENSIC DNA EXAMINATIONS **70**

Chapter Overview 71
More Accurate Reads 71
More Information 73
More Streamlined Methods 73

CHAPTER 5

DISCUSSION & CONCLUSION **76**

LIST OF FIGURES

CHAPTER 1

Fig. 1 Chromosomal positions of the 22 autosomal loci profiled in this project 3
Fig. 2 Replication slippage in repeat sequences 4
Fig. 3 Basics of autosomal and Y-STR profiles 5
Fig. 4 Hidden variation in length-based STR profiles 8
Fig. 5 Allelic gains in core CODIS loci with NGS 10
Fig. 6 How nanopore sequencing works 12
Fig. 7 Overview of project aims 16

CHAPTER 2

Fig. 8 STR sequencing and profiling with STRspy 18
Fig. 9 STRspy benchmarking at different normalization thresholds 25
Fig. 10 Autosomal STR profile predictions 24
Fig. 11 STRspy genotyping errors 34

CHAPTER 3

Fig. 12 Y-STR profile predictions 55
Fig 13 STRspy can resolve isoalleles in 24-sample multiplexes 59

LIST OF TABLES

CHAPTER 1

Table 1 Cost comparison between portable ONT devices and the MiSeq Fgx system 14

CHAPTER 2

Table 2 STRspy can resolve autosomal isoalleles 28
Table 3 SNP benchmarking 31

CHAPTER 1

Hide & seq: Finding nucleotide-level variation in forensic STR profiles

FORENSIC STRS

DNA evidence is the gold standard for human identification in forensic investigations due to the significant amount of information contained within the genome [1]. Numerous genetic marker systems have been developed to accommodate the amount and condition of DNA encountered in different casework scenarios [2,3]. Short tandem repeats (STRs) are the most common genetic markers with millions of profiles generated in forensic laboratories throughout the world each year [4]. The high degree of repeat length variation observed across established panels of STRs enables individualization of evidence and identification of the respective source when enough intact DNA is available [5–7]. Other genetic markers have lower discriminatory power than STRs but often provide useful information when DNA evidence is degraded or low copy [2,3]. Although smaller versions of STRs (miniSTRs), single nucleotide polymorphisms (SNPs), and mitochondrial DNA (mtDNA) can generate critical investigative leads in some casework scenarios, STRs remain the marker of choice in forensic laboratories worldwide [2–4].

Forensic genetic analyses harness a small set of well-characterized STRs for human identification and databasing [5,7–9]. The Federal Bureau of Investigation (FBI) maintains the Combined DNA Index System (CODIS) to facilitate electronic sharing of standard genetic information between participating laboratories in the United States [8]. The 20 STR loci that comprise the core CODIS panel are depicted in **Fig. 1** [7,9,10]. CODIS-compliant profiles can be uploaded and searched against those in the National DNA Index System (NDIS) database. This allows analysts at the federal, state, and local levels to link crime scene evidence from an active investigation to cold cases and repeat offenders [8]. STR profiles and associated databases have therefore become a pillar of our criminal justice system.

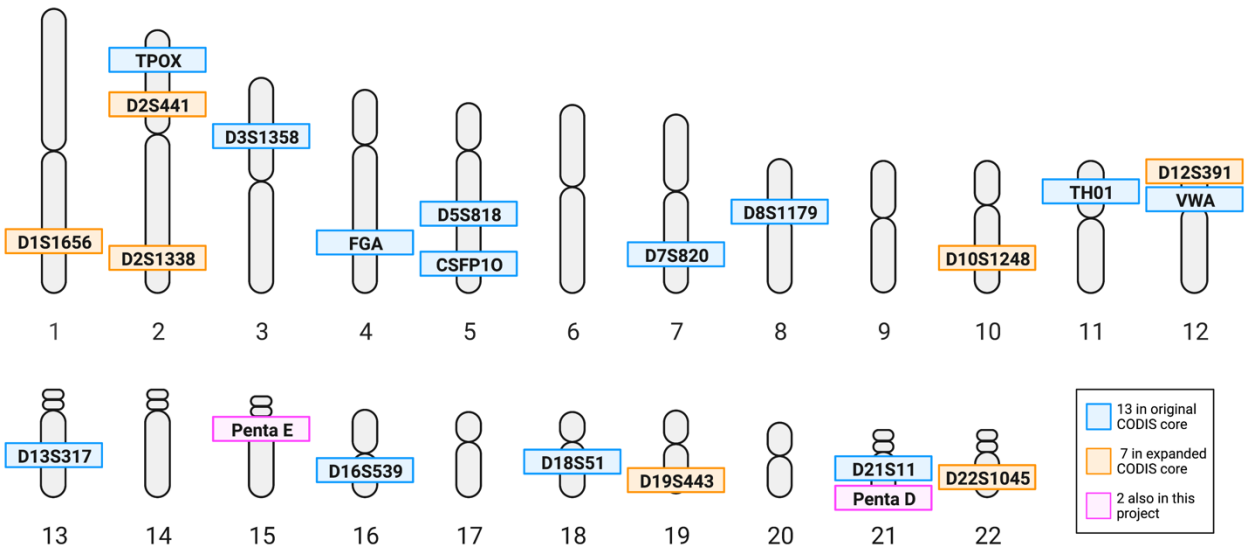


Fig. 1 Chromosomal positions of the 22 autosomal loci profiled in this project. The rapid growth of the NDIS database prompted the FBI to add 7 loci (orange) to the original 13 (blue) in the core CODIS panel. This expanded panel has a higher discriminatory power and lower likelihood of returning adventitious matches. The 2 other loci profiled in subsequent chapters are also shown in pink. Created with biorender.com.

BIOLOGICAL BASIS

STRs are defined as short segments of DNA (2 to 6bp) that repeat in tandem to form simple, compound, and complex patterns [5,6]. These low complexity sequences are characterized by mutation rates several orders of magnitude larger than unique, non-repetitive regions of the genome [11]. The high level of genetic instability observed at STRs results from polymerase slippage during DNA replication [12,13] (**Fig. 2**). This well-established mutation model maintains that repeat sequences in the template strand cause DNA polymerase to stall and dissociate from the replication complex [11–13]. Polymerase slippage in turn allows the nascent strand to dissociate from the template strand and reanneal to a repeat unit in either direction. Elongation then resumes producing a nascent strand that is expanded or contracted or expanded

relative to the template strand. Consequently, the number of repeat units observed at different loci varies between individuals, making STRs ideal genetic markers for human identification [5,6].

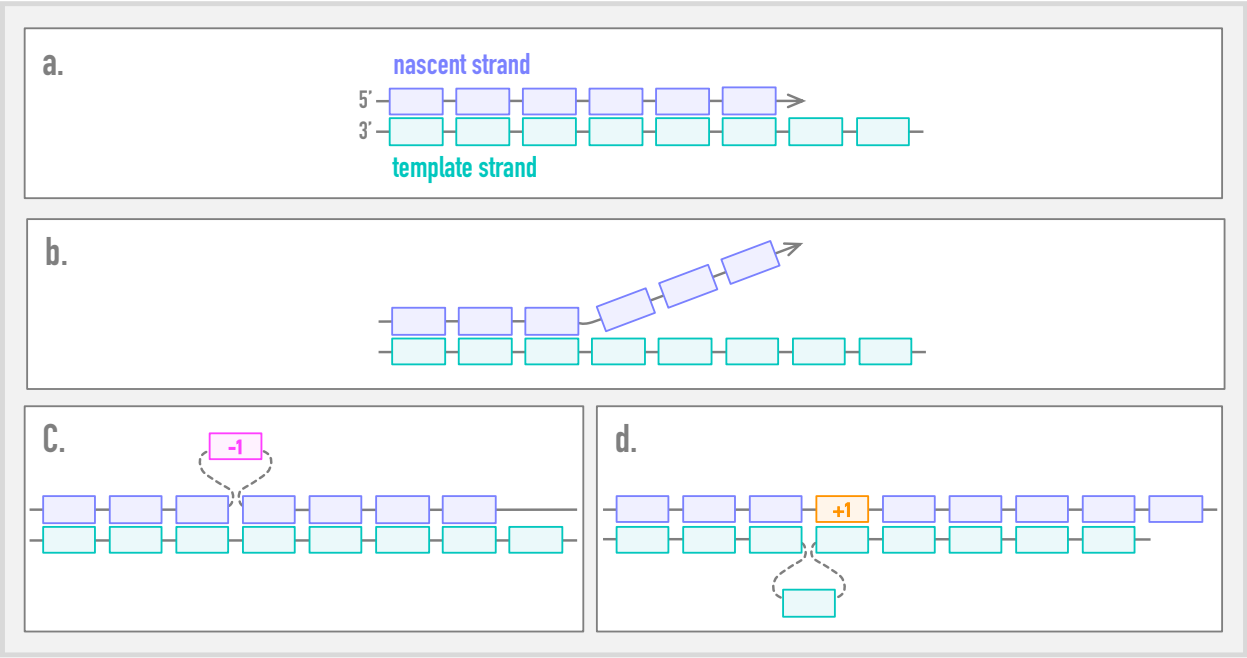


Fig. 2 Replication slippage in repeat sequences. (a) DNA polymerase replicates STR-containing sequences during elongation. (b) A high density of low complexity repeats in the template strand can cause polymerase to pause and dissociate from the replication complex, allowing the 3' end of the nascent strand to unpair from the template. The nascent strand then reanneals to a repeat unit in either direction resulting in DNA that is (c) contracted or expanded (d) relative to the template.

Panels of STRs on both the autosomal and sex chromosomes have been developed for forensic purposes. Autosomal STRs are the predominant and preferred genetic markers for human identification due to the high power of discrimination achieved using conventional length-based profiles [5]. The individualizing nature of autosomal STRs is attributable to the pattern of inheritance and codominant expression of alleles in forensic profiles [5]. Offspring inherit one autosomal chromosome from each parent at random and the alleles on both chromosomes can

be detected in autosomal STR profiles (**Fig. 3**). The unique allelic composition observed at autosomal STR loci in established panels can therefore be used to link DNA from a crime scene to a known source and confirm familial relationships.

Fig. 3 Basics of autosomal and Y-STR profiles. One of each autosomal chromosome is inherited from both parents at random (top). Parental alleles are represented as peaks in the resultant electropherogram and labeled with the number of complete and incomplete repeats separated by a decimal point. The maternal and paternal alleles for individual 1 contain the same number of repeats (5) and are depicted as a single peak whereas individual 2 is heterozygous (3.2, 5) at this locus. In contrast, the Y chromosome is passed from father to son in a linear manner (bottom). Paternal male relatives thus share the same haplotype in the absence of mutational events between generations.

STRs located on the Y chromosome (Y-STRs) are also profiled in routine forensic casework when male DNA is present [14,15]. In contrast to autosomal STRs, Y-STRs cannot be used for individualization because the Y chromosome is passed from father to son in a linear manner (**Fig. 3**). Paternal male relatives will therefore have identical Y-STR profiles in the absence of mutational events between generations. Although often unable to differentiate between descents of the same paternal lineage, Y-STRs often provide critical information in sexual assaults as well as father-son and sibling assessments.

CONVENTIONAL TECHNIQUES

Forensic laboratories harness two basic molecular methods to obtain length-based STR profiles from limited amounts of DNA evidence [4]. After DNA extraction and quantification, STRs of interest are targeted for amplification in the polymerase chain reaction (PCR) using locus-specific primers labeled with fluorescent dyes. This process results in an exponential increase in STR-containing DNA fragments and allows for subsequent separation and detection of alleles via capillary electrophoresis (CE). The fluorescent PCR amplicons are injected into and migrate through the CE system at speeds that correspond to length (shortest to longest). The migration times of laser-excited fragments are recorded relative to an internal size standard and compared to an STR allelic ladder to produce repeat length designations. Resultant data are visualized as peaks in an electropherogram labeled with the number of complete and incomplete repeat motifs separated by a decimal point (**Fig. 3**). These profiles can therefore resolve STRs of different repeat lengths but do not contain information regarding the sequence composition of each allele.

STR profiles are often generated using commercial kits that target the core CODIS loci. These standard DNA-to-profile workflows simplify sample processing and eliminate burdens associated with panel development and validation in individual forensic laboratories. The high

degree of genetic variation captured in length-based STR profiles provides the strong statistical support needed to individualize DNA evidence. The chance of observing the resultant profile within a given population is calculated as the product of individual genotype frequencies of all loci and thus increases with the number of STRs in the panel. The power of discrimination achieved with the 20 core CODIS loci is often sufficient for human identification and NDIS database searches in routine casework [9].

The primary disadvantages associated with length-based STR profiles (resolution, locus multiplexing, sample throughput) stem from the inherent limitations of CE instruments. The information obtained using conventional typing techniques is restricted to variation in repeat length rather than the underlying nucleotides. Resultant profiles are therefore unable to resolve STRs of the same length but distinct sequence composition or motif organization (isoalleles) and distinguish minor alleles from PCR-induced stutter. Researchers have demonstrated that length-based STR profiles may be inadequate for mixture deconvolution and complex kinship analyses even when additional STRs and other genetic markers are included alongside the 20 loci in the expanded core CODIS panel [16,17]. Larger panels of loci generally provide a higher power of discrimination, but the number of STRs that can be profiled in a single reaction is also limited by CE. The ability to multiplex STRs of interest depends on the spread of alleles at each locus as well as the dye channels in the instrument itself. Commercial kits have been strategically designed to capture up to 27 loci and available CE systems can process 8 to 24 samples during each run (depending on the number of capillaries) [4]. However, separate PCR and CE reactions must be performed for autosomal and Y panels. Male samples require additional DNA and analyst time and thus contribute to the persistent backlog problem faced by forensic laboratories across the nation. This is also true for other genetic markers including miniSTRs, SNPs, and mtDNA. Despite the relative ease and high discriminatory power of current typing approaches, CE is considered

low resolution and throughput compared to more recent advancements in DNA sequencing technologies.

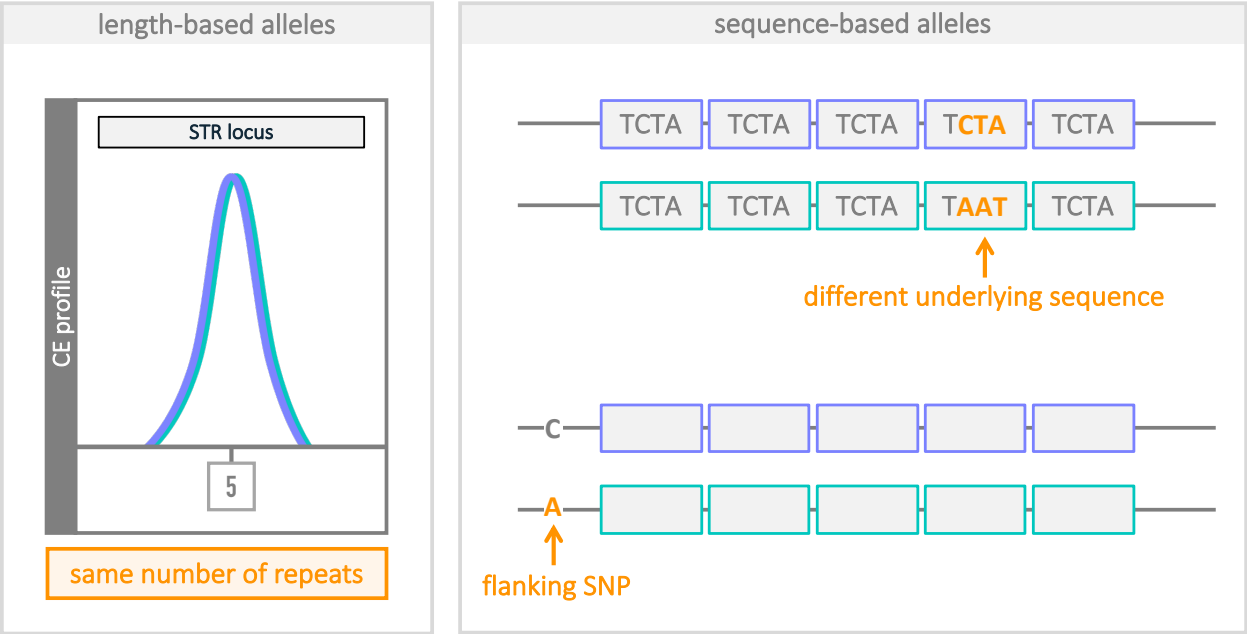


Fig. 4 Hidden variation in length-based STR profiles. The resolution of conventional STR profiles is limited to repeat length variation. Both alleles at this STR locus contain 5 repeat units and are thus represented by a single peak in the CE profile. These length-based homozygote alleles are however heterozygous in terms of sequence and harbor additional variation in the flanking region that cannot be detected using convention typing techniques.

HIDDEN VARIATION

The potential to harness all the information contained in STR amplicons has led to a significant amount of interest in DNA sequencing for human identification. Early studies involving Sanger sequencing revealed an abundance of nucleotide-level variation both in and around forensic STRs (**Fig. 4**) [18]. This first-generation sequencing method allowed researchers to characterize STR loci and provided valuable insight into how mutational events impact resultant CE profiles [19–25]. Variation that occurs within STR loci can alter the concept of allele sharing and trigger a complex process of evolution that impacts the diversity and distribution of alleles within a

population [26]. Although Sanger sequencing helped forensic researchers uncover hidden variation within length-based profiles, this method is time consuming, labor intensive, and low throughput.

The advent of next-generation sequencing (NGS) has enabled forensic researchers to access this information with increasing ease and speed in larger, more diverse populations. NGS is a class of high throughput techniques that can sequence many DNA fragments in parallel (and thus is also known as massively parallel sequencing or MPS). The most well-established NGS method for forensic STR profiling is the sequencing by synthesis (SBS) chemistry used in Illumina platforms. SBS involves reversible incorporation and fluorescent detection of terminator nucleotides in bridge-amplified DNA clusters to produce high accuracy read data [27]. The enhanced multiplex capabilities and throughput of NGS over CE and Sanger sequencing allow autosomal and Y-STRs to be profiled in a single run alongside other forensically relevant genetic markers. Verogen (the forensic branch of Illumina) has harnessed these features in the first and only STR sequencing workflow approved for upload into the NDIS database. This integrated solution can profile over 200 forensic genetic markers in a single run using 1ng of DNA. NGS can therefore provide more information in less time than the current PCR-CE method.

Illumina SBS data has been used to detect population-specific flanking region SNPs and differentiate between isoalleles. Although not all loci feature isoalleles, researchers have identified more than twice as many sequence-based alleles using NGS compared to CE at some STRs (**Fig. 5**) [28]. The high throughput nature of Illumina platforms has also enabled more samples to be sequenced in less time, thus revealing variation that had not been observed in prior studies [28]. The resultant increase in allelic diversity has been shown to facilitate complex kinship analyses and mixture deconvolution.

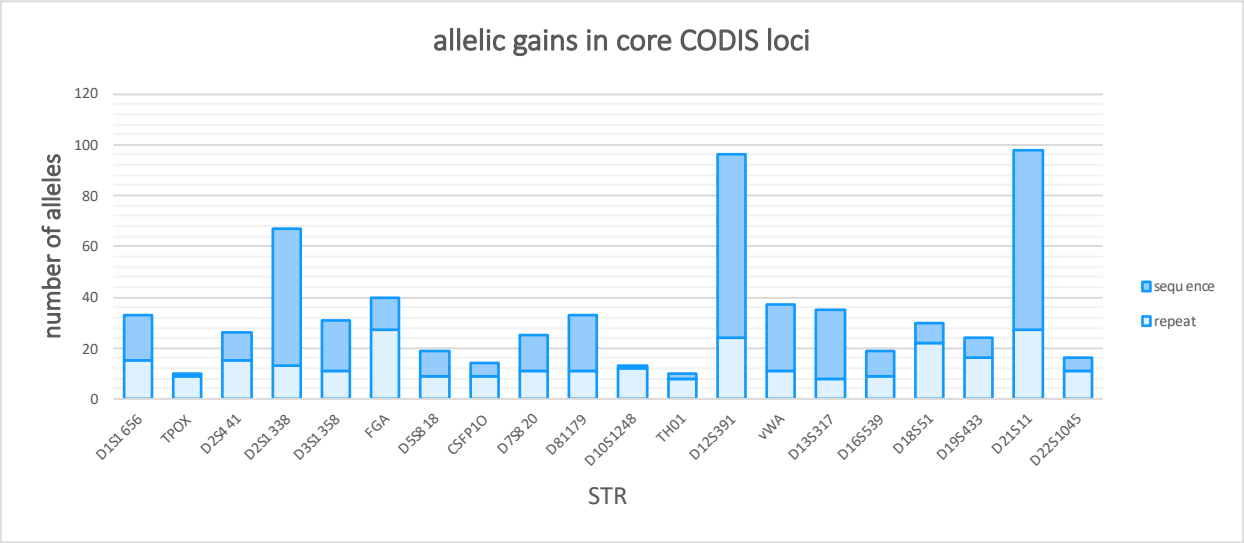


Fig. 5 Allelic gains in core CODIS loci with NGS. Stacked bar chart depicting the number of alleles gained at the 20 core CODIS loci in the NIST1036 dataset. The light and dark blue represent the number of alleles by length and sequence, respectively.

Significant efforts have been geared toward developing and validating forensic NGS workflows and data analysis software. Despite FBI approval, widespread adoption of sequence-based STR typing has been hindered by the high startup fees and steep learning curves associated with NGS. Most forensic laboratories would be unable to allocate funds to purchase and validate these platforms while maintaining conventional STR typing workflows. This would force analysts to outsource for NGS data as needed, further increasing backlog and decreasing the speed at which investigative hits are generated.

NANOPORE SEQUENCING

OVERVIEW

The recent development and commercialization of nanopore sequencing devices by Oxford Nanopore Technologies (ONT) has brought the potential to bypass some financial obstacles of NGS and could even support forensic field applications in the future. ONT sequencing relies on

the translocation of molecules through nanopore proteins to determine the composition of nucleotides in native strands of DNA (**Fig. 6**) [27]. Application of an electric voltage across a nanopore-containing membrane produces a constant ionic current through each of the pores within a given flow cell [29]. Flow cells contain hundreds to thousands of independent nanopore channels that are controlled and measured by an application-specific integrated circuit (ASIC) [30]. Disruptions in the baseline current occur as individual strands of DNA are unwound and passed through the pore by a motor protein. These current disruptions, which are unique to the motif of three to five bases present in the pore, are recorded and decoded to determine the sequence of nucleotides [31]. ONT platforms are therefore capable of directly sequencing reads of any length in a massively parallel fashion.

ONT manufactures various library preparation kits to accommodate different starting materials and applications. The ligation-based sequencing and barcoding kits used in this project provide a flexible workflow for preparing multiplexed amplicon libraries containing up to 24 samples [32]. After end-repair and dA-tailing, unique barcode adapters are attached to each sample via ligation and pooled in equimolar amounts. In the final steps of library preparation, nanopore sequencing adapters and tethers are sequentially ligated onto both ends of the DNA fragments to facilitate strand capture and processing. Prepared libraries are directly loaded onto the nanopore-containing sensor array within the disposable flow cell. The flow cell is then inserted into the respective ONT device for sequencing and data collection.

ONT IN FORENSICS

Nanopore sequencing offers numerous advantages over current CE and NGS approaches used in forensic DNA examinations. One of the most unique and appealing features of ONT sequencing is the scalability. The use of ionic current disruptions through nanoscopic pores allows DNA to be sequenced without the large, laboratory-confined equipment required for Illumina SBS. This

makes it possible to simultaneously profile STRs and other markers of forensic interest on platforms that are scalable to the output needs and financial restrictions of individual laboratories.

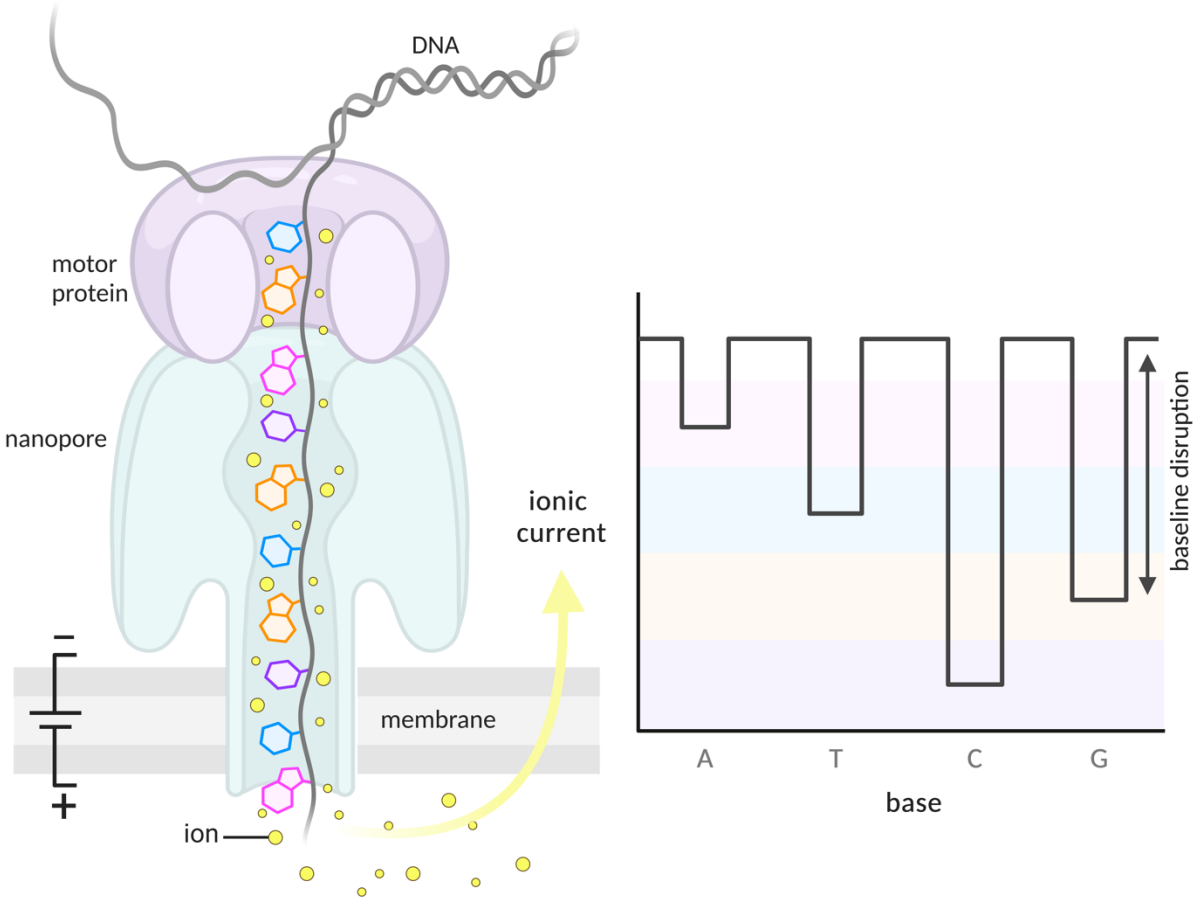


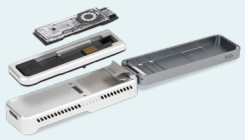
Fig. 6 How nanopore sequencing works. ONT sequencing platforms determine nucleotide composition by decoding the motif-specific disruptions in ionic current as DNA is translocated through nanopore proteins (green). During this process, an electric voltage is applied across the nanopore-containing membrane (grey) causing ions (yellow) to flow through the nanoscopic hole. Double-stranded DNA in prepared libraries are directed to available nanopores by the motor protein (purple). This helicase then unzips and pushes individual strands of DNA through the pore at a given speed. Each nucleotide motif causes a unique current disruption that can be decoded to produce the sequence of the DNA with available basecallers. Created with biorender.com.

The scalability of nanopore sequencing has given rise to a class of devices that could enable forensic genetic analyses to be performed at crime scenes and police stations. The MinION device used in this project is the smallest DNA sequencer available at the time of writing. This handheld platform weighs about 90g and can be controlled on a personal laptop via USB connection. ONT has implemented a degree of scalability in the MinION device itself for lower throughput experiments [33]. The Flongle adapter and flow cells provide a cost-effective alternative to standard MinION flow cells for single sample sequencing but are less stable and more sensitive to contaminates [32]. In further support of forensic field applications, ONT developed a portable device for automated library preparation known as the VolTRAX. Although the forensic potential has yet to be explored, the VolTRAX could be used alongside the MinION to facilitate development of a streamlined workflow for on-site DNA examinations with minimal human intervention.

In terms of cost, the pocket-sized MinION is a small fraction of the initial investment required for implementation of other NGS platforms. While both the Flongle and VolTRAX are priced higher than the MinION, the combined cost is still less than the Illumina MiSeq FGx Sequencing System (**Table 1**). It is the current high price of disposable MinION flow cells and reagents (as opposed to device startup fees) that would prohibit use in routine casework at present. This however will likely decrease with increasing commercial competition and improvements to the more affordable Flongle flow cells in the future. Further development of nanopore sequencing could make the ONT MinION device an efficient and cost-effective alternative to mainstream NGS platforms for forensic DNA examinations. The MinION has the potential to achieve the most comprehensive representation of genetic variation in DNA evidence at the site of collection. However, the relatively high error rate (particularly in

homopolymers and low complexity repeats) and lack of STR analysis software are significant obstacles to implementation of nanopore sequencing in forensics.

Table 1 Cost comparison between portable ONT devices and the MiSeq Fgx system.

	Flongle	MinION	VolTRAX	MiSeq Fgx
				
starter pack	$1.46k	$1k	$8.15k	$213.88k
flow cell	$90	$475–$900*	–	$1858k

*cost depends on pack size (96 flow cells to 1 flow cell)

Despite increasing interest in nanopore sequencing for forensic DNA examinations, studies focused on STRs are limited. The first published attempt to assess forensic STRs on the MinION device dates to 2018 in which only partial profiles could be extracted from error-prone nanopore reads [34]. Significant improvements in available basecalling and mapping algorithms have enabled researchers to achieve more accurate genetic profiles in recent years. While SNPs and mtDNA have been successfully typed in numerous forensic applications, STRs continue to be a challenge in terms of data analysis. Researchers have relied on a combination of in-house bioinformatic pipelines and tools developed for the broader class of tandem repeats (rather than forensic STRs). Although Ren et al. demonstrated that their STR-specific pipeline outperformed an available tandem repeat tool (repeatHMM), only 14 of the 27 autosomal loci were correctly typed across all samples [35]. These and other researchers have attributed the inability to obtain complete and accurate STR profiles to the high error rate of ONT platforms [34–37]. The results obtained were used to identify locus- and allele-specific features (repeat number, motif complexity, presence of homopolymers) that prevent successful genotyping using nanopore

sequencing data and provide guidelines for developing panels of ONT-compatible STR loci. A streamlined data analysis method capable of resolving length- and sequence-based STRs amplified by commercial kits from nanopore reads is critical for forensic ONT applications but has yet to be developed.

PROJECT OVERVIEW

PROBLEM

The power of discrimination achieved with established autosomal and Y-STR panels is limited by the length-based profiles generated using conventional typing techniques. Nucleotide-level variation within and around STRs increases resolution and facilitates interpretation in challenging casework scenarios. The MinION is a novel NGS platform that is small in both size and price tag. This portable device could provide an alternative for STR sequencing in forensic laboratories that cannot afford the initial investment or commitment of larger, laboratory-confined platforms. However, the relatively high error rate and lack of STR analysis software have precluded accurate forensic profiling with nanopore sequencing in previous studies.

HYPOTHESIS

Forensic autosomal and Y-STRs can be sequenced on the MinION device but will require a novel bioinformatic method to produce profiles consistent with CODIS databases from error-prone ONT read data.

SPECIFIC AIMS

This project will evaluate the application of nanopore sequencing platforms in forensic DNA examinations through the following specific aims:

1. Determine whether forensic STRs amplified with a commercial kit can be sequenced on the ONT MinION device.
2. Develop and test a bioinformatic pipeline capable of generating forensic STR profiles that capture sequence-based variation and are compatible with length-based CODIS databases.
3. Assess how PCR cycle number and sample multiplexing impact resultant STR profiles.

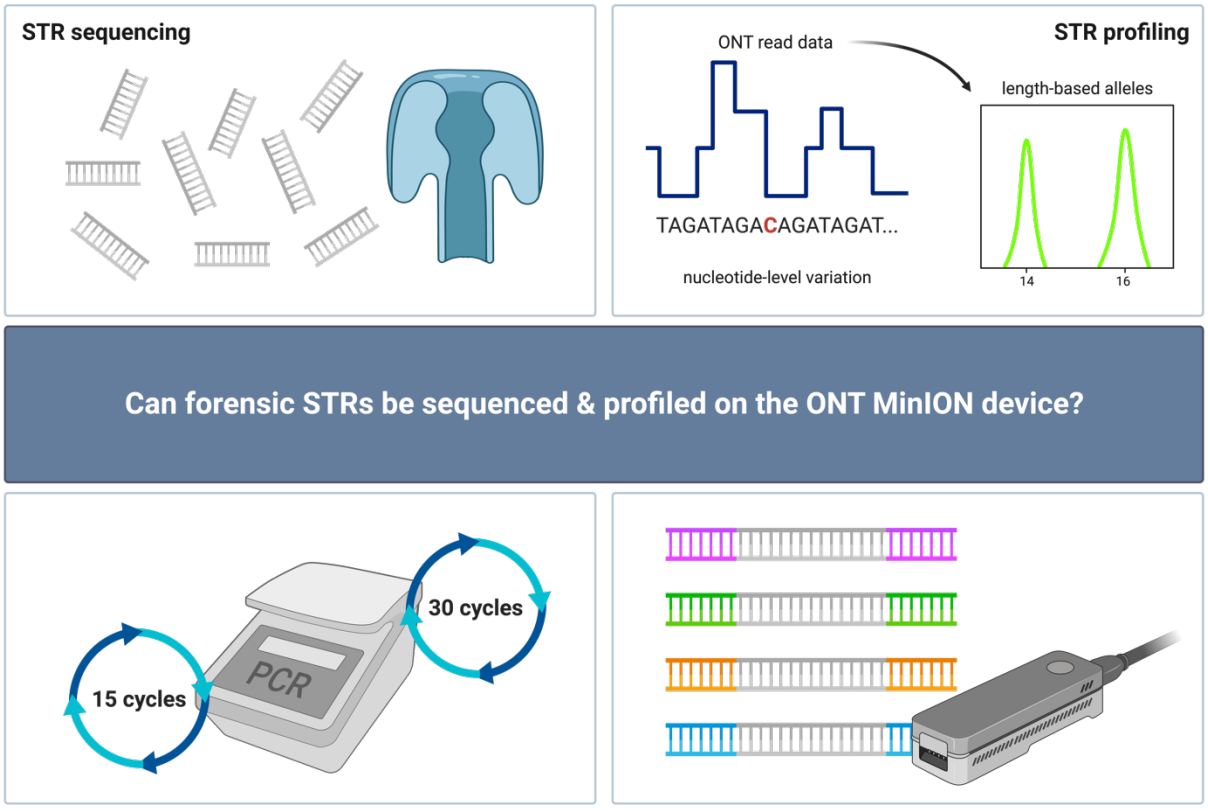


Fig. 7 Overview of project aims. Created with biorender.com.

CHAPTER 2

Accurate profiling of forensic autosomal STRs using the Oxford Nanopore Technologies MinION device

Published in Forensic Science International: Genetics (October 2022)

C.L. Hall
R.K. Kesharwani
N.R. Phillips
J.V. Planz
F.J. Sedlazeck
R.R. Zascavage

HIGHLIGHTS

- STRs can be sequenced on the Oxford Nanopore Technologies MinION device.
- STRspy correctly profiled 22 STRs amplified at 30 PCR cycles across all samples.
- STRspy produces accurate autosomal STR profiles from long-read sequencing data.
- SNPs in flanking regions were detected with > 90% accuracy for the 15-cycle dataset.
- Isoalleles can be resolved in nanopore sequencing reads when analyzed with STRspy.

CHAPTER OVERVIEW

The high variability characteristic of STR markers is harnessed for human identification in forensic genetic analyses. Despite the power and reliability of current techniques, nucleotide-level information both within and around STRs are masked in the length-based profiles generated. Forensic STR typing using NGS has therefore gained attention as an alternative to traditional CE approaches. This chapter aims to evaluate the forensic applicability of the newest and smallest NGS platform available – the ONT MinION device. Although nanopore sequencing on the handheld MinION offers numerous advantages, including low startup cost and on-site sample processing, the relatively high error rate and lack of forensic-specific analysis software have prevented accurate profiling across STR panels in previous studies. Here we present STRspy, a streamlined method capable of producing length- and sequence-based STR allele designations from noisy, error-prone third-generation sequencing reads. To assess the capabilities of STRspy, seven reference samples (female: n = 2; male: n = 5) were amplified at 15 and 30 PCR cycles with the Promega PowerSeq 46GY System and sequenced on the ONT MinION device in triplicate (**Fig. 8a**). Basecalled reads were then processed with STRspy using a custom database containing alleles reported in the STRSeq BioProject NIST 1036 dataset (**Fig. 8b**). Resultant STR allele designations and flanking region SNP calls were compared to the manufacturer-validated genotypes for each sample. STRspy predicted the correct genotypes across all autosomal STR loci amplified with 30 PCR cycles, achieving 100% concordance based on both length and sequence. Furthermore, we were able to identify flanking region SNPs in the 15-cycle dataset with >90% accuracy. These results demonstrate that ONT reads can reveal additional variation in and around STR loci depending on read coverage when analyzed with STRspy. As the first and only third-generation sequencing platform-specific method to

successfully profile the entire panel of autosomal STRs amplified by a commercially available multiplex, STRspy significantly increases the feasibility of nanopore sequencing in forensic applications.

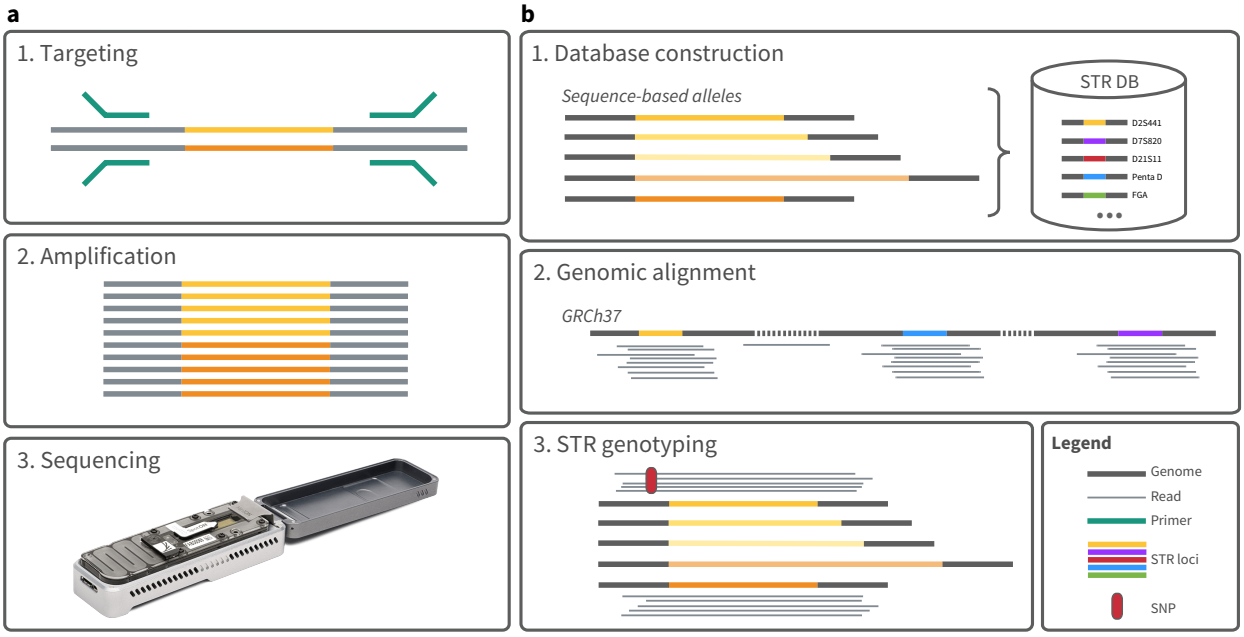


Fig. 8 STR sequencing and profiling with STRspy. (a) Lab workflow. STR loci are targeted and amplified via multiplex PCR. Amplicon libraries are then prepared and sequenced on the ONT MinION device to generate nucleotide-level data. (b) Data analysis pipeline. STRspy relies on a user-generated STR database (DB) containing sequence-based alleles for each locus of interest. Reads are first aligned to the human reference genome. Reads overlapping STR loci are then extracted and mapped to the custom STR DB. STRspy uses the normalized read counts to rank the STR alleles and predict the genotype at each locus. Sequencing data produced and analyzed as described can resolve alleles of the same length but different underlying sequence (dark yellow and orange) and identify SNPs in the flanking region (red). See figure legend for more details.

MATERIALS & METHODS

SAMPLES

The results presented in this chapter are based on sequencing data from six NIST traceable standards and one Promega control (female n = 2; male n = 5). Extracted DNA along with

validated length- and sequence-based genotype information for these reference samples were obtained directly from the respective manufacturers. Promega single-source male DNA 2800M for human STR analysis was normalized to 0.1ng/μL based on the manufacturer-specified quantification value. Components A, B, and C of NIST Standard Reference Material (SRM) versions 2391c and 2391d were quantified on the Qubit 2.0 Fluorometer using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific) and diluted to a concentration of 0.1ng/μL. The same methods were used to verify the final concentration of all samples prior to downstream applications.

STR AMPLIFICATION

Six full PCR reactions per sample were prepared with the Promega PowerSeq 46GY System (PS4600) according to the manufacturer's technical manual with 0.5ng input DNA. Amplification was performed in triplicate on an Eppendorf Mastercycler pro S using the recommended thermal cycling conditions at either 15 or 30 cycles. Resultant amplicons were then subject to an Agencourt AMPure XP bead (Beckman Coulter) cleanup (2.5:1 ratio based on sample volume) as previously described [38] to remove remaining primers and PCR reaction components. DNA was eluted in 48μL of nuclease-free water, which is the input volume required for the ONT library preparation protocol used herein.

ONT LIBRARY PREP & SEQUENCING

Purified PCR products were multiplexed and prepared for nanopore sequencing using the ONT Ligation Sequencing Kit (SQK-LSK109) with Native Barcoding Expansion 1-12 (EXP-NBD104). Library preparation was performed with the following modifications to the standard Native Barcoding Amplicons protocol (NBA_9093_v109_revC_12Nov2019). Amplicon DNA input (48μL from above) used for library preparation fell below the recommended 1μg for all samples.

Quantification steps were conducted on the Agilent TapeStation 4200 with D1000 ScreenTape for samples amplified at 30 cycles but were completely bypassed for the 15-cycle. Following DNA repair and end-prep, unique barcodes were ligated onto each bead-purified amplicon library to be sequenced together. Details regarding sample pooling per MinION flow cell are provided in **Supplemental Table S1**. To reduce potential sample loss from bead purification, pooled barcodes exceeding the volume required for subsequent steps (>65µL) were concentrated in an Eppendorf 5301 Vacufuge System. After ligation of ONT sequencing adapters, samples were subject to a final bead cleanup and washed with short fragment buffer (SFB, ONT). To minimize pore clogging and maximize yield of the short amplicon libraries, no more than 75ng was loaded onto each individual flow cell (based on previous optimization studies; data not shown). The 30-cycle pooled barcodes were therefore quantified and diluted to 75ng in elution buffer (EB, ONT) before preparing the loading library if necessary. Again, the 15-cycle amplicon libraries were not quantified, and the entire volume was used in the final reaction.

Prepared libraries were loaded in a drop-wise fashion into the SpotON port of primed vR9.4D flow cells (FLO-MIN106D, ONT). Flow cells were placed in the MinION device and sequenced until exhaustion (up to 72hrs) using the ONT MinKNOW software. Raw signal data were then processed as described in the Data Analysis section to obtain the base called reads.

BIOINFORMATICS PIPELINE & ALGORITHM

Implementation. STRspy is designed to predict forensic STR genotypes from third-generation sequencing data. STRspy requires a minimum of one thread and is executed at the command line. We implemented and tested this framework in a Unix/Linux environment. STRspy is under MIT license (open source) and can be downloaded from the GitHub page along with associated documentation, step-by-step instructions, and a small test set to verify successful installation.

STRspy relies on a user-generated reference database to produce allele designations consistent with the established forensic naming system [39]. The same STR database can be used to analyze any samples of interest, and thus users are only required to build it once. We constructed the database for this study using STR sequencing data for 1036 samples published under the STRSeq BioProject (NIST 1036) [28]. Our STR database includes all reported sequence-based alleles for the 22 PowerSeq autosomal loci along with 500bp flanks from the human reference genome (GRCh37/hg19). Each entry is labeled with the locus name, bracketed repeat motif, and length-based allele designation used in standard STR profiling (**Supplemental Fig. S1**). The custom STR database is available at https://github.com/unique379r/strspy.

STRspy accepts basecalled reads in the form of either fastq or bam files to accommodate both ONT and PacBio data. Users are also required to provide bed and fasta files for the STR database (see below). STRspy executes the following three steps in a per sample manner (**Supplemental Fig. S2**):

1. Basecalled reads are first aligned to the human reference genome (GRCh37/hg19) with minimap2 (v2.18-r1015) [40]. STRspy includes predefined parameters to adapt minimap2 to either ONT or PacBio read data. Subsequently, the mapped reads are automatically converted and sorted into a bam file using samtools (v1.12) [41].
2. The genome-wide bam file is processed with bedtools intersect (v2.30.0) [42] to extract reads that overlap STR loci of interest based on the locations specified in the user-provided bed file. The extracted locus-specific reads are then mapped to the predefined collection of alleles contained within the custom STR database using minimap2 (v2.18-r1015) [40]. As in the previous step, STRspy generates sorted bam files containing the mapped reads.

3. STRspy computes the number of reads (with mapping quality greater than 1) mapped to each sequence-based STR allele in the sorted bam files with samtools (v1.12) [41]. This part of the pipeline can be implemented in a multi-threaded manner to increase the speed of analysis. STRspy calculates locus-specific normalized read counts by dividing the number of reads per allele across the highest number of reads mapping to a single allele at each STR. Both the raw and normalized read counts are stored for subsequent filtering and assessment of the results. STRspy uses the normalized read counts to rank the STR alleles at each locus and reports either a single allele (homozygous) or the top two alleles (heterozygous) based on the user-defined normalization threshold. By default, this threshold is set to 0.4.

SNP detection. STRspy uses xAtlas [43] to detect SNPs within the flanking regions of each autosomal locus contained within the STR database and region bed file. SNP calls produced by xAtlas are output in vcf file format which is compatible with various available bioinformatic tools for downstream data analysis. We filtered resultant vcf files to keep SNP calls with "PASS" flags and p-values of 0.8 or higher. To prevent the accumulation of incorrect SNP calls due to differences in sequencing depth [44], samples amplified at 30 PCR cycles were uniformly subsampled to 1% of total mapped reads with samtools view -s 0.01 (v1.12) [41]. The randomly subsampled datasets were then used for SNP calling and benchmarking of the 30-cycle dataset.

DATA ANALYSIS

Raw signal data collected on the MinION device were basecalled and separated by barcode with the standalone GPU version of Guppy (v3.4.2). Reads with a q-score greater than 7 (those in the “pass” folder output by Guppy) were then merged by barcode using the concatenate command. These fastq files can be downloaded from the NCBI Sequence Read Archive (SRA, BioProject

accession #: PRJNA757759). Merged fastq files from the seven samples amplified at 15 and 30 PCR cycles in triplicate were processed using the STRspy command line interface to obtain normalized read counts, length- and sequence-based allele designations, and SNP calls in the flanking regions. The utility scripts available on the STRspy GitHub repository (https://github.com/unique379r/strspy) were implemented to assess the overall performance of STRspy, evaluate concordance between predicted and known genotypes, identify stutter artifacts, and visualize results as heatmaps and line plots.

Manufacturer-validated genotypes obtained via CE and NGS served as the ground truth for assessing STRspy performance based on error rate calculations. Correct allele predictions produced by STRspy were classified as true positives, incorrect as false positives, and drop out as false negatives. Precision and recall were calculated as the correct STR allele designations (true positive) out of total alleles reported by STRspy (true positive + false positive) or the ground truth dataset (true positive + false negative), respectively. F1 score, which provides a measure of overall test accuracy, was determined by taking the harmonic mean of precision and recall. These metrics were calculated with normalization cutoffs ranging from 0.1 to 0.9 to identify the optimal threshold at both cycle numbers (**Fig. 9**, **Supplemental Fig. S3**). STRspy achieved the highest recall, precision, and F1 score at a normalization threshold of 0.4 (see Results). Allele designations obtained at this cutoff (0.4) were therefore used as the STRspy predictions for overall performance assessments.

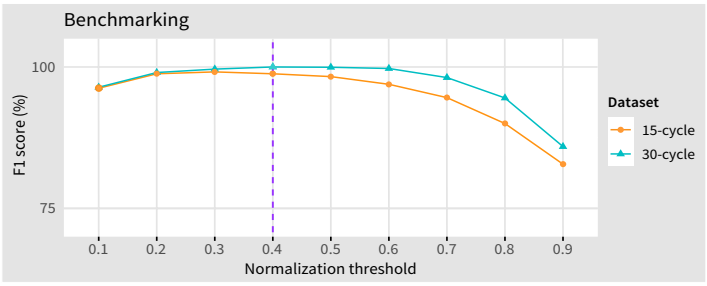
RESULTS

ASSESSING FORENSIC STRS ON THE ONT MINION DEVICE

As a relatively new sequencing platform, the ONT MinION device has undergone limited testing for forensic DNA analyses. To assess the capabilities of this device in the context of human

identification, 22 autosomal STRs were amplified at 15 and 30 PCR cycles using the Promega PowerSeq 46GY System and successfully sequenced on the MinION (see methods). Processing each of the seven reference samples in triplicate at both cycle numbers allowed us to evaluate on-target efficiency and depth of coverage between runs. As expected, the number of reads produced for each sample varied based on PCR cycle number (**Supplemental Table S2**). The percent of total reads that mapped to STR loci for samples in the 30-cycle dataset ranged from 87.76% to 92.87% with an average of 90.76%. More variability was observed across the 15-cycle dataset, in which on-target efficiencies fell between 50.96% and 71.67% and averaged 65.09%. Nonetheless, the raw read counts mapped to STR loci were comparable across 15-cycle samples. Similarly, depth of coverage per locus was impacted by PCR cycle number, resulting in a mean of 246,002.27 and 321.56 reads in the 30- and 15-cycle datasets, respectively. We also observed PCR amplification bias that resulted in reduced – and sometimes insufficient – coverage over several loci, particularly D22S1045 (**Supplemental Table S2**). The effect of amplification bias on genotype determination was overcome with increased PCR cycles. Overall, these data suggest that PCR amplification followed by nanopore sequencing results in high on-target rates and enables in-depth analysis of allelic content.

a



b

	Normalization threshold: 0.4					
	TP	FP	FN	Recall	Precision	F1 score
30-cycle	924	0	0	100%	100%	100%
15-cycle	902	6	16	98.3%	99.3%	98.8%

Fig. 9 STRspy benchmarking at different normalization thresholds. (a) Plot of F1 score across different normalization thresholds. (b) Table showing the number of true positive (TP), false positive (FP), and false negative (FN) predictions produced by STRspy as well as associated benchmarking metrics at the normalization threshold used in this study (0.4).

A bioinformatic pipeline capable of producing complete and accurate STR profiles from third-generation sequencing data has yet to be established. We, therefore, developed STRspy, a novel method for the detection and characterization of forensic STR loci using ONT and PacBio reads (see methods). STRspy can identify different STR alleles in a phased manner and detect SNPs present in the flanking region, thus leveraging all information contained within the amplicons. Our method employs a user-defined threshold to predict if a locus is heterozygous (reporting the top two alleles) or homozygous (reporting the top allele) based on the normalized coverage supporting each STR allele. Thus, we first determined the optimal cutoff value by evaluating STRspy performance at different normalization thresholds in the 15- and 30-cycle datasets (**Fig. 9**, **Supplemental Fig. S3**). Recall, precision, and F1 score were 100% for samples amplified at 30 cycles when this threshold was set to 0.4. Decreasing (0.3) or increasing (0.5) the normalization threshold cutoff resulted in lower benchmarking values for the 30-cycle dataset. As the only normalization threshold at which all samples were correctly typed, 0.4 was considered the optimal cutoff value.

To determine how depth of coverage impacts profiling speed, we measured the runtime of STRspy using a single thread for each sample. The average runtime across samples in the 30-cycle dataset was 571 minutes (9.51 hours) due to the high depth of coverage (mean: 246,002.27). We observed a significant reduction in STRspy runtime for the lower coverage 15-cycle dataset (mean: 321.56), which averaged 3.54 minutes per sample. STRspy is implemented to support multithreading and thus runtimes can be improved using multiple CPU cores to increase analysis speed. By sequencing and analyzing triplicate samples amplified at two distinct cycle numbers, our results provide novel insight into how coverage impacts genotype determination, reproducibility, and processing time.

LENGTH- & SEQUENCE-BASED GENOTYPE DETERMINATIONS

We assessed the true positive (correct STR allele), false positive (incorrect STR allele or additional STR allele at known homozygous loci), and false negative (missing STR allele at heterozygous loci) rates for each autosomal STR compared to the manufacturer-validated genotypes. Using these metrics, we were able to determine if STRspy fully recovered known STR genotypes and correctly assigned allele designations for each individual sample.

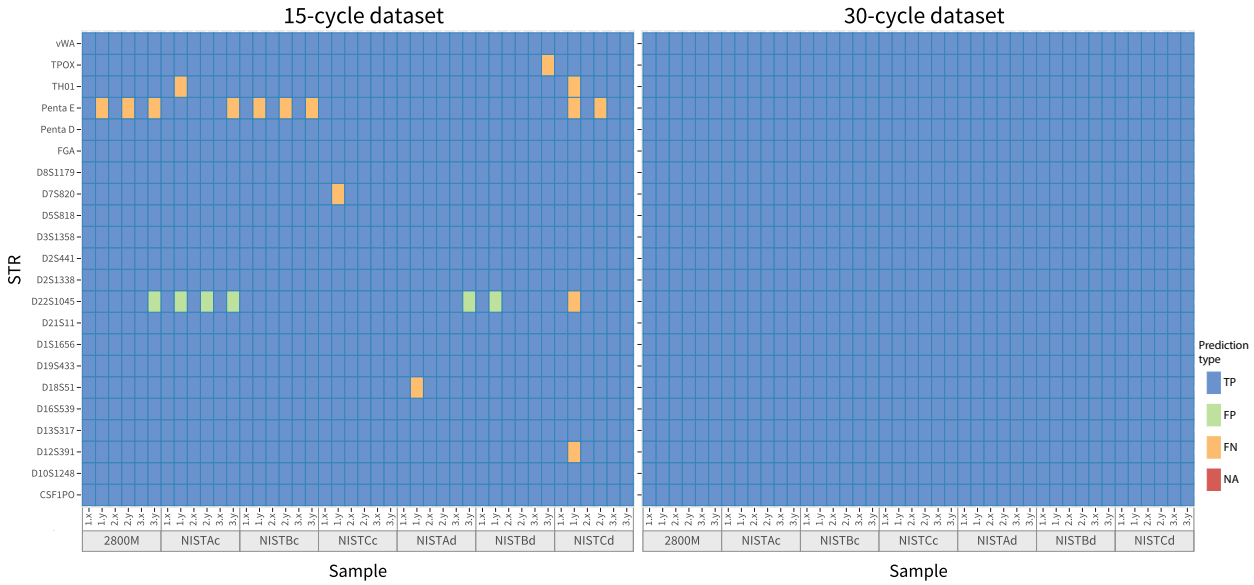


Fig. 10 Autosomal STR profile predictions. Heatmap comparison of STRspy predictions to manufacturer-verified length- and sequence-based genotypes across the 15-cycle dataset and 30-cycle dataset. True positive (TP) predictions are depicted in blue, false positives (FP) in green, and false negatives (FN) in orange. Reference samples (grey boxes) are labeled by triplicate (1, 2, 3) and haplotype (x, y).

STRspy was able to consistently predict the correct allele designations based on both length and sequence for all 22 autosomal loci amplified at 30 PCR cycles (**Fig. 10**). The utility of ONT sequencing data analyzed with STRspy is demonstrated by the 30-cycle triplicates for NIST A from SRM 2391c (NISTAc). STRspy successfully identified repeats characterized by simple motifs such as the D2S441 tetranucleotide [TCTA]10 allele. Further, our method was able to

resolve the length-based homozygous 10 alleles observed at this locus to produce heterozygous calls consisting of the simple [TCTA]10 and compound [TCTA]8 TCTG [TCTA]1 repeats (**Table 2**). Similar results were achieved for NIST B from SRM 2391d (NISTBd), which possesses isoalleles at DS2441 (11, 11). These data also enabled differentiation of isoalleles between samples, further increasing profile resolution (**Table 2**).

Table 2 STRspy can resolve autosomal isoalleles. Normalized read counts, raw read counts (parentheses), and STRspy predictions (bold) for isoalleles at D2S441 loci with repeat lengths of 10 or 11. Reported values are for triplicate 1 in the 30- and 15-cycle datasets. TP = true positive.

Repeat length		30-cycle			15-cycle		
10		[TCTA]10	[TCTA]8 TCTG TCTA	Prediction	[TCTA]10	[TCTA]8 TCTG TCTA	Prediction
	2800M	**0.681 (40203)**	0.015 (861)	TP	**0.825 (70)**	0.012 (1)	TP
	NISTAc	**0.862 (72250)**	**1.0 (83848)**	TP	**0.984 (60)**	**1.0 (61)**	TP
	NISTBc	**0.737 (88101)**	0.057 (6768)	TP	**0.853 (64)**	0.027 (2)	TP
	NISTCc	0.035 (7280)	**1.0 (206237)**	TP	0.032 (3)	**1.0 (95)**	TP
11		[TCTA]11	[TCTA]9 TCTG TCTA	Prediction	[TCTA]11	[TCTA]9 TCTG TCTA	Prediction
	NISTAd	**1.0 (83390)**	0.022 (1851)	TP	**1.0 (123)**	0.016 (2)	TP
	NISTBd	**1.0 (57604)**	**0.912 (52511)**	TP	**1.0 (80)**	**0.975 (78)**	TP
	NISTCd	**0.993 (47865)**	0.036 (1710)	TP	**0.906 (106)**	0.009 (1)	TP

Despite variation in raw and normalized read counts, STRspy was able to resolve sequence-based heterozygous alleles of the same length using ONT reads across the 22 loci. Complete concordance was achieved for all samples amplified at 30 PCR cycles, resulting in 100% recall, precision, and F1 score (**Fig. 9**). These observations demonstrate the ability of our method to (1) differentiate alleles of the same length but different sequence and (2) accurately genotype simple, compound, and complex repeat motifs using ONT sequencing data.

Next, we evaluated the ability of STRspy to profile the same seven samples at 15 PCR cycles (**Fig. 10**). STR loci amplified with a lower number of PCR cycles had less coverage

compared to those in the 30-cycle dataset (**Supplemental Table S2**). Nevertheless, STRspy distinguished between the length-based homozygous 10 and 11 alleles at D2S441, predicting the correct heterozygote genotypes across all three 15-cycle triplicates for samples in this dataset (**Table 2**). Other repeat motifs are composed of homopolymers and intervening sequences that are not counted toward the length-based genotypes produced via CE. Even with this low level of coverage, STRspy predicted the correct allele designations at the homopolymer-containing Penta D and complex FGA loci. Allele designations concordant with CE were also obtained for D19S433 and D21S11 despite the presence of intervening sequences that complicate length-based profiling from sequencing data.

Precision, recall, and F1 score for the 15-cycle datasets were 99.34%, 98.26%, and 98.80%, respectively (**Fig. 9**). The 22 incorrect genotypes (out of 924) produced by STRspy fall into two distinct categories of errors: false positives (two alleles predicted at a known homozygous locus, or the incorrect allele predicted) and false negatives (one allele predicted at a known heterozygous loci). All six of the false positive allele designations were observed at D22S1045 due to the relatively low coverage over this locus (**Supplemental Table S3**) and the presence of stutter artifacts (see below). The other 16 errors were false negatives, an overwhelming majority (9) of which were at Penta E (**Fig. 11a**). False negatives at this locus across all samples were characterized by allele dropout of the longer repeating unit. For instance, in one NISTAc triplicate, STRspy correctly predicted [AAAGA]5 but not [AAAGA]10 at Penta E (5, 10). Although a greater number of raw reads supported the 10 allele for the false negative genotype (15-cycle.3: 99 reads) compared to a true positive genotype (15-cycle.1: 38 reads), the normalized read count for the 15-cycle.3 NISTAc triplicate fell below the 0.4 threshold.

In contrast to Penta E, STRspy correctly predicted the longer [AATG]6 ATG [AATG]3 but not the shorter [AAGT]8 for TH01 in one of the NISTAc triplicates. Further examination of the

individual NISTAc datasets in which these loci were problematic revealed a minor allele normalized count of 0.38 and 0.31 for Penta E and TH01, respectively (**Supplemental File S1**). Consequently, decreasing the normalization cutoff value to 0.3 increased the 15-cycle F1 score from 98.80% to 99.13% by preventing minor allele dropout (**Supplemental Fig. S3**). These observations ultimately suggest that the prevalence of false negatives is due to amplification bias and lack of locus coverage rather than inherent limitations of STRspy itself.

FLANKING REGION VARIATION

Single nucleotide polymorphisms (SNPs) as well as insertions and deletions (indels) have been observed in sequences around forensic STRs. These variants further increase the discriminatory power of current STR panels but cannot be detected in the length-based profiles generated via CE. We therefore examined the ability of STRspy to detect known flanking region SNPs in the NIST SRMc/d samples. Detailed benchmarking results are provided in **Table 3**.

We first assessed the SNP calls at samples amplified with 30 PCR cycles. Poor performance was observed for the non-subsampled 30-cycle dataset, indicating that excessive coverage (mean: 246,002.27) hinders SNP calling due to the accumulation of sequencing errors. For this reason, reads were subsampled and reanalyzed as in previous publications [44]. The recall and precision achieved by the subsampled 30-cycle dataset were 92.06% and 74.05%, respectively. The reduced coverage (mean: 321.56) obtained with fewer amplification cycles in the 15-cycle dataset eliminated the need for subsampling. We recovered known SNPs across all but one of the samples amplified with 15 PCR cycles (rs1728369 in NISTB.1 from SRM 2391c). The overall recall and precision for the 15-cycle dataset were 98.41% and 84.05%, respectively. These results show that lower coverage improves our ability to identify SNPs in terms of both precision and recall.

Table 3 SNP benchmarking. Comparison of filtered calls generated by xAtlas to known flanking region SNPs in the 30- and 15-cycle NIST triplicates. DB position is the SNP position in our STR database with respect to the associated STR allele and 500bp flanks. 30-cycle data were subsampled as described in the main text.

Sample					30-cycle			15-cycle		
Locus	STR	SNP	DB position	dbSNP ID	Recall	Precision	F1 score	Recall	Precision	F1 score
NISTCc										
D13S317	11	T	545	rs9546005	66.67%	100%	80%	100%	50%	66.67%
NISTAd										
D16S539	13	C	406	rs1728369	100%	42.86%	60%	100%	50%	66.67%
D1S1656	15.3	T	569	rs4847015	100%	100%	100%	100%	100%	100%
D1S1656	18.3	T	581	rs4847015	100%	100%	100%	100%	100%	100%
D5S818	11	G	557	rs25768	100%	42.86%	60%	100%	100%	100%
D7S820	8	A	479	rs7789995	100%	100%	100%	100%	100%	100%
NISTBd										
D13S317	11	T	545	rs9546005	33.33%	16.67%	22.22%	100%	100%	100%
D16S539	9	C	552	rs11642858	100%	100%	100%	100%	100%	100%
D16S539	11	C	406	rs1728369	66.67%	50%	57.14%	66.67%	40%	50%
D1S1656	15.3	T	569	rs4847015	100%	100%	100%	100%	100%	100%
D2S1338	17	A	466	rs6736691	100%	100%	100%	100%	75%	85.71%
D5S818	12	A	497	rs73801920	100%	37.50%	54.55%	100%	50%	66.67%
		G	561	rs25768						
D5S818	12	G	561	rs25768	100%	27.27%	42.86%	100%	50%	66.67%
D7S820	10	A	479	rs7789995	100%	100%	100%	100%	100%	100%
NISTCd										
D13S317	14	T	557	rs9546005	100%	42.86%	60%	100%	100%	100%
D16S539	9	C	552	rs11642858	66.67%	100%	80%	100%	100%	100%
D5S818	13	G	565	rs25768	100%	75%	85.71%	100%	100%	100%
D5S818	15	G	573	rs25768	100%	60%	75%	100%	100%	100%
D7S820	9	A	479	rs7789995	100%	60%	75%	100%	50%	66.67%
		A	545	rs16887642						
D7S820	10	A	479	rs7789995	100%	100%	100%	100%	100%	100%
TPOX	8	A	405	rs145426142	100%	100%	100%	100%	100%	100%
				Overall	**92.06%**	**74.05%**	**82.08%**	**98.41%**	**84.05%**	**90.66%**

Unlike SNPs, indels in the flanking region impact the length-based allele designations used in forensics. Flanking region indels were therefore incorporated into the STR database itself and can be identified by inspecting the bracketed repeat motif reported by STRspy. For instance, a subset of alleles observed at D13S317 are characterized by a rare 4bp deletion in the flanking

region [28]. Consequently, the length-based 11 allele can correspond to [TATC]11 or [TATC]12. The latter repeat motif, in which the 4bp deletion occurs within the 3' flank, is identical to that of a 12 length-based allele but is identified as an 11 via CE. Despite these complexities, STRspy was able to distinguish between the [TATC]12 with the 4bp flanking region deletion (2800M), [TATC]12 (NISTBc, NISTAd, NISTCd), and [TATC]11 (NISTCc, NISTBd) to produce the correct sequence- and length-based genotypes across all samples at this locus.

Polymerase slippage during amplification of low complexity repeats can lead to stutter artifacts in resultant datasets [13]. In contrast to the sequence-by-synthesis technique harnessed by Illumina platforms, nanopore sequencing relies on the direct detection of nucleotides in each strand of DNA [27]. This unique capability provides novel insight into PCR-induced bias. To assess the impact of amplification cycle number on stutter artifact formation we examined the prevalence of reads one repeat unit smaller and larger than the true allele (n±1) with the STRspy utility scripts (**Supplemental File S1**). Previous STR genotyping attempts have been complicated by the presence of stutter artifacts at D18S51 ([AGAA]n) in ONT sequencing data [35]. Consistent with the notion that stutter percentage increases with the number of PCR cycles, we observed higher normalized read counts for n±1 stutter at D18S51 when NISTAc (12, 15) was amplified with 30 cycles (0.36, 0.38, 0.41) compared to 15 cycles (0.26, 0.28, 0.27). Nevertheless, STRspy was able to identify the correct alleles at both cycle numbers even when the normalized read count for stutter exceeded 0.4 (as in **Fig. 11c**: 30-cycle.3 and 15-cycle.2).

We also investigated how stutter artifacts contribute to the false positive results in the 15-cycle dataset. As mentioned, all six false positive allele designations were observed at D22S1045. Although notable, this observation is unsurprising because D22S1045 is known to have higher stutter values than other loci [45–47]. The raw count data revealed that a relatively low number of reads mapped to this locus across samples in both datasets, which is indicative of amplification

bias (**Supplemental Table S3**). Consequently, STRspy called an additional allele at D22S1045 for one of the 2800M samples amplified with 15 PCR cycles (**Fig. 11b**). Most reads mapped to the true homozygous allele (16) at this locus. However, the presence of stutter in the minus direction (15) exceeded the normalization threshold of 0.4 and was therefore called by STRspy. Similar observations were made at one of the NISTAd triplicates (**Fig. 11c**). In contrast to the allele drop-in for 2800M (which is homozygous at D22S1045), STRspy produced the incorrect designations for the shorter alleles in the NISTAd (14, 16) heterozygote. In the 30-cycle.3 and 15-cycle.2, the 15 allele (representing the overlap of minus stutter for the 16 allele and plus stutter for the 14 allele) exceeds the normalization threshold but falls below the true alleles in rank and thus is not reported by STRspy. Interestingly, the incorrect allele prediction for 15-cycle.3 was minus stutter (13) associated with the minor allele (14) rather than the overlap (15).

We observed higher levels of both amplification bias and stutter compared to PowerSeq 46GY amplicon sequencing data produced on the Illumina MiSeq FGx [48]. Despite the use of different amplification kits, these observations are consistent with other ONT-based STR studies [35]. Additionally, the loci identified as artifact-prone herein, namely D18S51 and D22S1045, were also noted in both of these studies [35,48]. Collectively, these observations highlight the stochastic nature of PCR-induced artifacts as well as the impact of amplification bias on genotyping errors.

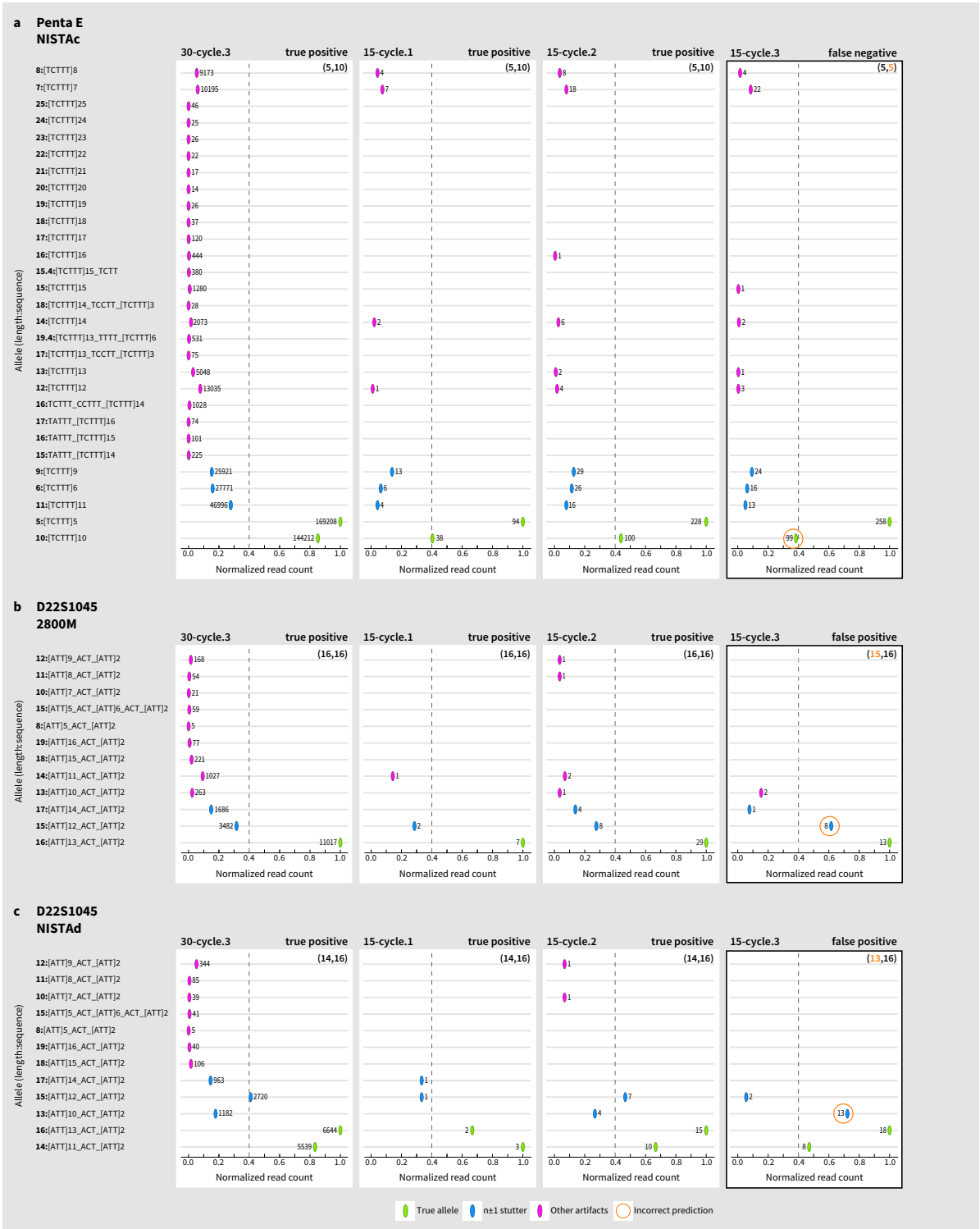


Fig. 11 STRspy genotyping errors. (a) False negative genotype due to allele drop out at Penta E for NISTAc. False positive genotype due to stutter artifacts at D22S1045 for the (b) 2800M homozygote and (c) NISTAd heterozygote. Raw read counts are included next to each point. The incorrectly typed triplicate in each set is denoted by a black box and the incorrect allele prediction is circled in orange. One 30-cycle (left) and all three 15-cycle triplicates are shown for comparison purposes.

DISCUSSION

In this chapter, we report the first STR analysis for accurate predictions of allele designations along with identification of flanking region SNPs specific to third-generation sequencing platforms. Using the Promega PowerSeq Kit and the ONT MinION device we produced robust sequencing data across all targeted loci that enabled us to investigate the impact of PCR cycle number. Although a higher number of reads mapped to PCR artifacts in the 30-cycle dataset, the normalized read counts for the true alleles exceeded stutter, resulting in the correct predictions at all loci. The 15-cycle dataset was skewed due to the low level of coverage, and thus the 30-cycle dataset produced more reliable genotypes. Additionally, we showcased the accurate identification of STR alleles and flanking regions SNPs. These results suggest that this portable, scalable, and rapid sequencing approach could prove extremely valuable in future applications. A maximum of four samples were pooled and sequenced on a single MinION flow cell. Given the high level of coverage achieved at 30 PCR cycles, it may be possible to increase the number of samples sequenced on a single MinION flow cell (without exceeding 75ng total) to reduce overall cost. STRspy leverages ONT sequencing data to profile STRs with unprecedented accuracy regardless of repeat motif, complexity, or length. All relevant studies to date have reported incorrect genotypes at vWA, FGA, and D21S11 due to repeat pattern complexity [34–37]. We demonstrated that the novel method developed and tested in the current paper was able to produce the correct length- and sequence-based allele designations for vWA, FGA, and D21S11 across all samples even at the low-level coverage obtained from 15 PCR cycles.

While the length and continuous sequence information obtained in this study enabled accurate STR identification and phasing using STRspy, current ONT data still suffers from certain biases (homopolymer error rates) that may impact the performance of STRspy. We predict that recent and future developments from ONT (in the base calling algorithm) will improve the

quality of sequencing data, further increasing the accuracy and performance of STRspy-based analyses. Despite current biases in ONT sequencing data, STRspy predicted the correct genotypes for the homopolymer-containing Penta D and Penta E using 30-cycle reads produced on the standard R9 nanopore proteins. Although Penta D was also correctly typed across all samples in the 15-cycle dataset, drop out of the minor Penta E allele was observed in nine samples. Reducing the normalization cutoff value from 0.4 to 0.3 resulted in the correct genotype, suggesting that the establishment of locus-specific normalization thresholds in future studies may be beneficial when analyzing low-coverage samples. Given the improvements in STR data quality previously reported [36], the use of the R10 nanopore proteins may further mitigate this issue by increasing the number of usable reads produced from lower cycle numbers.

In addition to producing robust and reliable STR profiles, STRspy possesses numerous features that support implementation in forensic genetics without the need for extensive bioinformatic training in third-generation sequencing data processing and analysis. Our easy-to-install method can be used on computational infrastructures ranging from personal laptops to high-performance clusters, closely mirroring the scalability of nanopore sequencing platforms. Furthermore, STRspy executes all steps required to go from basecalled reads to STR profiles based on user-defined parameters and input files. The minimal computational requirements and streamlined nature of STRspy not only increase the overall accessibility of ONT sequencing in forensic genetics but also supports field applications. Although beyond the scope of the current study, samples processed with the ONT Field Sequencing Kit should be analyzed with STRspy to establish protocols using the limited laboratory and computational equipment that would be available at crime scenes.

The ability of STRspy to achieve correct genotype predictions depends on the alleles being present in the STR database provided by the user. Because the custom reference database

we generated contains the most common STR alleles observed among the four major U.S. populations [28], it can be used to profile many unknown samples. This list however is not exhaustive, and the database constructed for this study only includes autosomal STRs amplified by the Promega PowerSeq 46GY System. STRspy relies upon a best-fit alignment model, meaning that if the true sequence is not present in the database, reads are mapped to the entry that is the closest match and short indels are inferred. Subsequent iterations of our software will include flags for poor alignment scores so users can manually confirm novel alleles. Moreover, the alleles in our database will be continuously updated in accordance with forthcoming STRSeq BioProject publications. Future efforts will be geared towards adding the 23 PowerSeq Y-STRs to our database and assessing profiles generated using STRspy. Users can also expand upon or create their own database containing STRs of interest if sequence-based allele information is available and formatted in the same manner across all loci.

The data analyzed herein was produced in a conventional laboratory setting using high quality DNA extracts. Each sample was amplified and sequenced in triplicate, providing novel insight into the reproducibility of STR profiling on the MinION device. Although we sequenced more amplicon libraries than all relevant publications [34–37] our study included a small number of unique, single-source samples. Additional experiments involving more reference and probative samples will be conducted in future studies. These data will allow us to evaluate STR profiling from biological material of similar quality to that collected from suspects and crime scenes, respectively. The current release of STRspy selects the top two alleles (at most) with normalized read counts above the user-defined cutoff, and thus can only type single-source samples. We will use future mixture analysis studies to determine optimal thresholds and evaluate read balance ratios. If these results indicate that STRspy is capable of mixture interpretation, we will implement the necessary changes to our bioinformatic pipeline. Further

assessment of amplification at various cycle numbers and input DNA concentrations will also provide important information about the sensitivity of nanopore sequencing devices and expand upon our understanding of PCR-induced artifacts. Through these studies, we will identify the minimum level of coverage at which we can accurately type all STRs to improve SNP calls. These experiments will ultimately form the foundation for establishing ONT-specific protocols and interpretation guidelines for STR profiling.

A key limitation of sequence-based STR typing in routine forensic casework is cost. Despite the relatively low startup fee of the ONT MinION device, the price per sample exceeds that of mainstay short-read sequencing platforms [49]. The rapid evolution of nanopore sequencing since the 2014 release of the MinION device has led to a significant decrease in error rate and increase in throughput that have, in turn, reduced overall cost [33,50]. Continued technological improvements and developments (Flongle adapter and flow cells) in coming years will likely reduce the cost and increase the accessibility of ONT sequencing. Despite the cost of sequencing, the ONT MinION device provides unique advantages including higher resolution over current typing techniques and faster turnaround time with potential on-site analyses. These features would be particularly beneficial in forensic investigations.

CONCLUSION

This chapter assessed the ability to profile forensic STRs using nanopore sequencing data produced on the ONT MinION device. With our forensic-specific analysis method, STRspy, we were able to achieve accurate STR profiles for all autosomal loci amplified at 30 cycles with the Promega PowerSeq 46GY System. The results presented herein demonstrate that nanopore sequencing platforms can produce length-based allele designations consistent with standard forensic nomenclature while revealing an additional level of variation in and around STR loci.

The novel pipeline we developed overcomes the issues reported in previous publications to profile the entire panel rather than a subset of STRs amplified by a commercially available kit. We anticipate that continued improvements in nanopore sequencing technologies, along with further development of STRspy, will increase the feasibility of forensic STR profiling on ONT devices in not only a traditional laboratory setting but also at crime scenes and police stations.

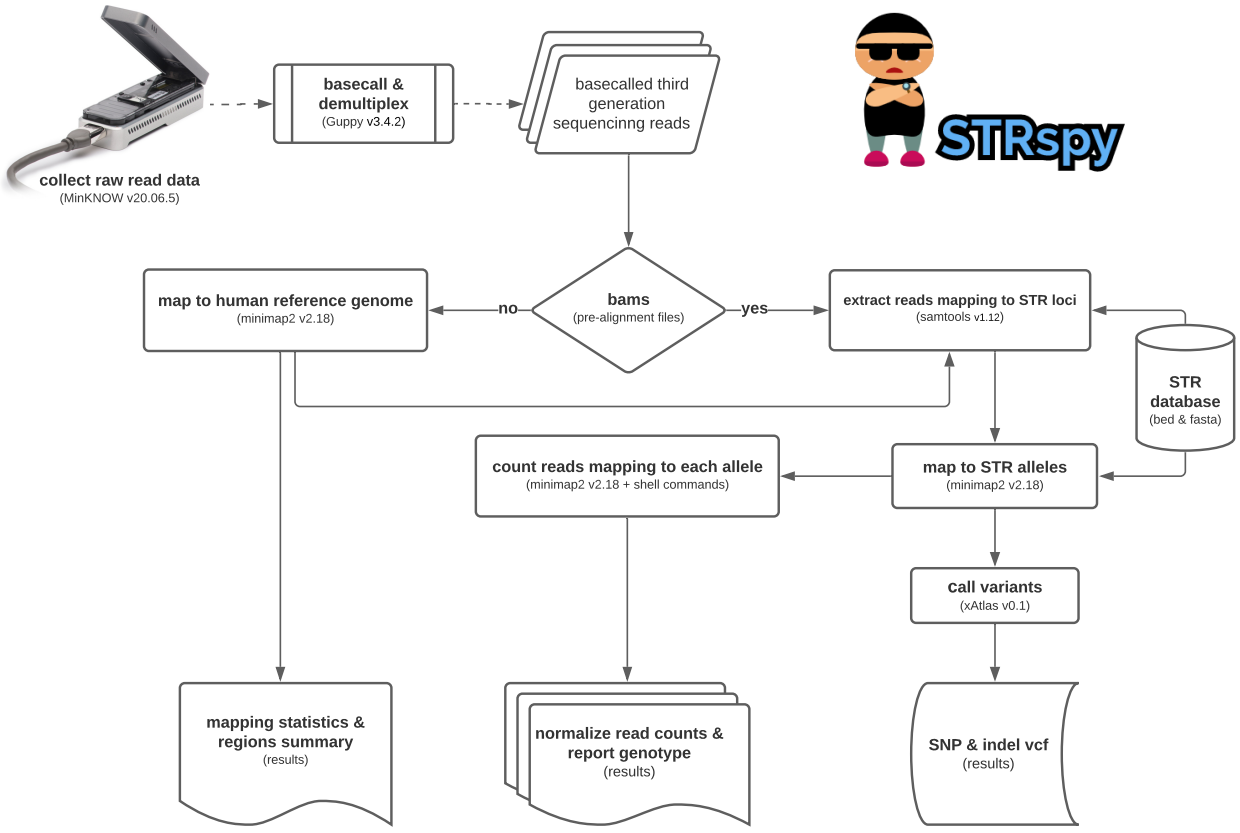
ACKNOWLEDGMENTS

We would like to thank Jonathan King and Promega for providing the validated sequence-based allele designations for the 2800M reference sample.

SUPPLEMENTAL MATERIAL

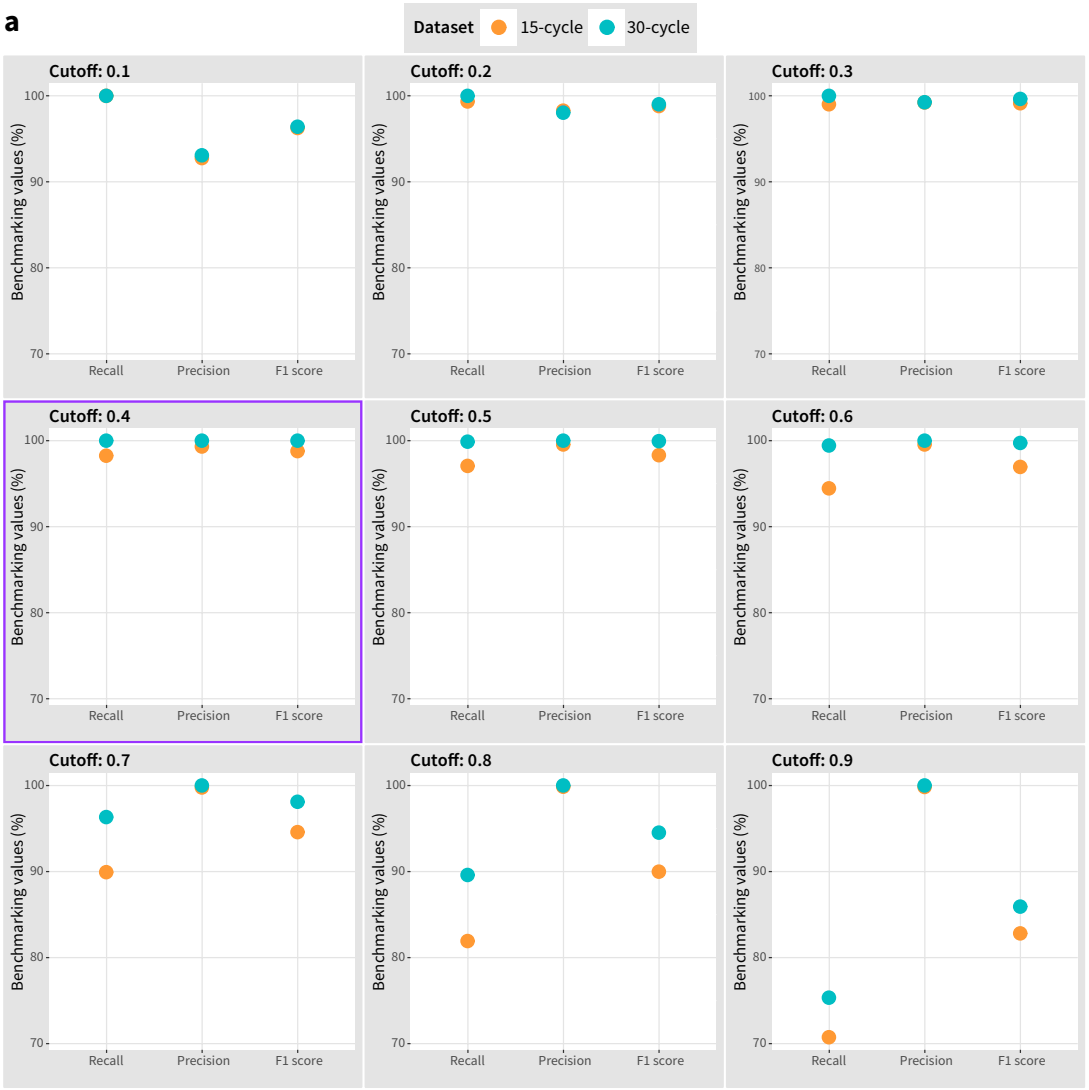
FIGURES

Supplemental Fig. S1 STR allele database structure. The STR allele database required by STRspy consists of the text string of nucleotides corresponding to each sequence-based allele (yellow) along with 500bp flanks (orange). Each entry is labeled with the locus name (pink), bracketed repeat motif (blue), and length-based allele designation used in forensics (green). The isoalleles with a repeat length of 10 at D2S441 are depicted.



Supplemental Fig. S2 Steps implemented in STRspy. Additional documentation is provided on the GitHub page (https://github.com/unique379r/strspy).

a



b

	30-cycle			15-cycle		
Cutoff	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**
0.1	860	64	0	857	67	0
0.2	906	18	0	902	16	6
0.3	917	7	0	908	7	9
0.4	924	0	0	902	6	16
0.5	923	0	1	893	4	27
0.6	919	0	5	869	4	51
0.7	890	0	34	829	2	93
0.8	828	0	96	756	1	167
0.9	696	0	228	653	1	270

Supplemental Fig. S3 (a) Benchmarking plots across a range of normalization thresholds for the 30- and 15-cycle datasets. (b) Table showing the number of true positive (TP), false positive (FP), and false negative (FN) predictions produced by STRspy at each cutoff tested. The optimal cutoff (0.4) and default normalization threshold for STRspy is denoted by purple boxes.

FILE

Supplemental File S1 Per allele read counts generated by STRspy for autosomal STRs.

TABLES

Supplemental Table S1. Cycle number triplicates per MinION flow cell.

run #	1	2	3	4	5	6	7	8	9
PCR cycles	15	15	15/10*	30	30	30	30	30	30
sample.triplicate	NISTAc.1	NISTAc.2	NISTCd.3	NISTAd.1	NISTAd.2	NISTAd.3	NISTAc.1	NISTAc.2	NISTAc.3
	NISTBc.1	NISTBc.2	2800M.3	NISTBd.1	NISTBd.2	NISTBd.3	NISTBc.1	NISTBc.2	NISTBc.3
	NISTCc.1	NISTCc.2	NISTAc.1*	NISTCd.1	NISTCd.2	NISTCd.3	NISTCc.1	NISTCc.2	NISTCc.3
	NISTAd.1	NISTAd.2	NISTBc.1*	2800M.1	2800M.2	2800M.3			
	NISTBd.1	NISTBd.2	NISTCc.1*						
	NISTCd.1	NISTCd.2	NISTAc.2*						
	2800M.1	2800M.2	NISTBc.2*						
		NISTAc.3	NISTCc.2*						
		NISTBc.3	NISTAc.3*						
		NISTCc.3	NISTBc.3*						
		NISTAd.3	NISTCc.3*						
		NISTBd.3							

*sequencing data not included in the current study

Supplemental Table S2 Sample coverage for the cycle number triplicates. Table reports the number of total and STR aligned reads as well as the percentage of total reads mapping to STRs.

		30-cycle			15-cycle		
sample	triplicate	total	STR	total:STR	total	STR	total:STR
2800M	1	8211019	7325461	89.22%	10185	6622	65.02%
	2	10485404	9514161	90.74%	23738	16014	67.46%
	3	7811448	7101644	90.91%	20073	13139	65.46%
NISTAc	1	6852921	6103122	89.06%	5300	3904	73.66%
	2	6969767	6116495	87.76%	12605	8806	69.86%
	3	7558006	6682430	88.42%	15800	10550	66.77%
NISTBc	1	13656701	12447212	91.14%	10107	6005	59.41%
	2	11933551	10987665	92.07%	27290	16096	58.98%
	3	15724912	14333631	91.15%	23344	13608	58.29%
NISTCc	1	11726444	10890382	92.87%	7531	5149	68.37%
	2	11895951	10930803	91.89%	19136	12908	67.45%
	3	16201452	14916992	92.07%	21173	14164	66.90%
NISTAd	1	3981187	3518920	88.39%	7141	3639	50.96%
	2	9281500	8414983	90.66%	20586	12047	58.52%
	3	5489336	4941964	90.03%	17547	9858	56.18%
NISTBd	1	7993749	7245492	90.64%	9559	6340	66.32%
	2	13698486	12612967	92.08%	25091	17363	69.20%
	3	7991785	7279976	91.09%	19127	12487	65.28%
NISTCd	1	7670183	7088214	92.41%	10707	7642	71.37%
	2	16719680	15425589	92.26%	26608	19071	71.67%
	3	7701310	7021321	91.17%	17753	12365	69.65%
	average	**9978800**	**9090449**	**90.76%**	**16686**	**10847**	**65.09%**

Supplemental Table S3 Per locus coverage at autosomal loci for the cycle number triplicates. Number of STR aligned reads are colored by cycle number and shaded by relative abundance.

	2800M						NISTAc					
	30-cycle			15-cycle			30-cycle			15-cycle		
locus	1	2	3	1	2	3	1	2	3	1	2	3
CSF1PO	169767	216912	170000	281	562	258	263987	188244	250155	229	459	591
D10S1248	95953	183556	114634	158	313	192	311980	307800	302964	170	291	352
D12S391	21702	44736	28570	20	96	65	60840	74293	80803	15	48	56
D13S317	488914	492982	407104	588	1190	924	345109	322820	368942	193	543	513
D16S539	282577	317466	260642	394	854	535	485513	458703	459694	329	740	870
D18S51	116014	194866	106289	115	284	441	161537	188980	228169	91	176	223
D19S443	245493	396392	261474	129	466	541	227703	244519	241712	110	227	269
D1S1656	37035	68307	51325	64	174	70	101896	132295	120351	117	237	290
D21S11	265224	396627	306047	280	739	676	431928	407419	464096	251	569	511
D22S1045	15878	29305	18080	10	46	24	41063	46420	53816	12	20	34
D2S1338	376779	402458	335286	216	566	416	320787	397136	434729	179	453	503
D2S441	138141	211803	155877	199	449	199	190622	206014	199786	145	393	374
D3S1358	160745	209272	148585	103	311	348	288595	295155	297238	92	140	300
D5S818	111005	116507	101906	94	243	110	131223	143085	131532	93	241	250
D7S820	165077	300183	165107	118	397	302	341196	258409	287685	149	262	467
D8S1179	302155	351651	256893	189	463	288	345556	335082	428901	159	369	447
FGA	128690	176202	129371	180	503	508	229549	223692	271279	181	351	462
Penta D	219969	396228	263543	271	783	640	306652	312606	306243	254	425	559
Penta E	309908	353834	278340	163	405	266	448552	437098	458131	169	438	443
TH01	44309	47579	44024	47	106	47	51285	42818	70111	38	71	99
TPOX	142544	142943	133945	164	333	91	164372	172086	205864	182	326	438
vWA	212730	227728	180048	121	399	204	280494	267956	304777	112	302	386

	NISTBc						NISTCc					
	30-cycle			15-cycle			30-cycle			15-cycle		
locus	1	2	3	1	2	3	1	2	3	1	2	3
CSF1PO	296398	220588	309651	205	587	500	266059	197391	296703	179	459	543
D10S1248	270480	184799	243734	107	301	184	268991	158664	358958	125	254	226
D12S391	90457	76685	94731	27	76	35	83739	58422	106584	28	51	43
D13S317	325552	268551	402410	250	578	498	352039	269528	476273	157	565	567
D16S539	705160	590208	721051	346	1031	894	474035	562772	731277	269	752	947
D18S51	194255	194832	296560	111	230	210	146100	183637	259604	85	241	201
D19S443	272265	240930	385266	119	306	265	213249	259586	326646	97	241	259
D1S1656	103403	84521	96964	94	148	164	92889	76805	119394	71	168	177
D21S11	319888	406941	526133	239	576	521	441315	389342	560341	205	588	524
D22S1045	66636	51582	86476	13	37	32	49062	42650	88512	13	18	26
D2S1338	321069	460762	607548	155	429	390	349702	393974	550938	137	411	415
D2S441	272826	217433	326606	178	517	379	250374	188510	251555	118	332	344
D3S1358	351641	320593	439379	131	263	191	286862	243033	401911	109	146	130
D5S818	154500	124033	200354	75	211	236	113559	131055	209179	50	190	192
D7S820	369689	254079	427833	118	355	216	308724	312060	423097	146	283	243
D8S1179	388063	384737	536389	134	447	430	345588	406055	657146	191	384	464
FGA	202612	163863	227322	169	371	402	208922	239671	293230	147	421	365
Penta D	293763	338933	416388	237	567	607	375616	372316	444385	211	368	509
Penta E	584430	435251	568703	122	358	273	496317	467644	499213	58	192	107
TH01	96272	96940	162633	38	131	124	52830	103280	129611	63	84	122
TPOX	284390	224686	257383	157	513	418	170807	239469	394572	166	383	506
vWA	264348	293201	417577	137	366	366	240996	313745	339879	100	283	328

Supplemental Table S3 (continued) Per locus coverage at autosomal loci for the cycle number triplicates. Number of STR aligned reads colored by cycle number and shaded by relative abundance.

	NISTAd						NISTBd					
	30-cycle			15-cycle			30-cycle			15-cycle		
locus	1	2	3	1	2	3	1	2	3	1	2	3
CSF1PO	159545	379589	165794	224	595	484	163066	315352	142828	226	589	456
D10S1248	94375	291112	204096	105	408	330	96639	219339	122561	107	329	344
D12S391	22947	94625	43846	16	82	53	18274	54483	31544	14	83	51
D13S317	239385	442218	297919	266	593	549	253945	390726	213007	307	756	413
D16S539	343736	572865	412608	333	1136	729	258683	518126	296108	408	809	653
D18S51	79384	216390	125219	63	248	220	102023	213105	109893	105	331	196
D19S443	155988	299279	191178	99	348	262	232579	346445	202582	117	294	233
D1S1656	44017	142208	84760	104	331	305	50908	97216	57390	89	266	189
D21S11	229765	619634	343534	218	625	782	328908	565571	272991	207	767	569
D22S1045	9249	45982	17708	7	38	40	7682	30413	16075	10	26	32
D2S1338	236455	588644	325490	169	751	461	319740	435890	254477	183	559	365
D2S441	119107	279178	159401	178	498	401	150314	239256	145523	188	394	318
D3S1358	92727	310602	166276	66	292	261	120282	238227	169284	67	361	236
D5S818	88319	164974	111669	69	306	208	114668	187782	98921	98	239	134
D7S820	155330	390528	211380	120	378	371	194621	443934	230690	134	522	334
D8S1179	241627	518087	318946	130	444	400	269615	466043	238510	172	495	394
FGA	98070	312697	151859	170	482	505	116332	267750	124667	215	556	416
Penta D	161142	413033	278337	156	609	504	193519	373701	192432	222	675	547
Penta E	292591	785303	424708	224	533	226	280474	463161	296972	280	723	429
TH01	31052	74020	27489	22	135	77	26180	53471	31586	38	101	93
TPOX	86064	228198	144664	106	303	400	92518	134284	93222	143	392	276
vWA	129942	295265	208196	121	560	228	199911	269842	166845	139	436	284

	NISTCd					
	30-cycle			15-cycle		
locus	1	2	3	1	2	3
CSF1PO	181918	347085	139435	268	687	205
D10S1248	118632	289057	119369	185	537	167
D12S391	19526	55658	21820	13	78	43
D13S317	234685	430222	224969	271	696	481
D16S539	272614	665564	278439	369	876	518
D18S51	101880	237211	110931	118	358	318
D19S443	217894	417491	167517	165	415	344
D1S1656	62643	109086	57524	133	241	137
D21S11	300888	717769	259865	408	737	711
D22S1045	12519	37239	16984	7	48	14
D2S1338	325476	684039	295453	232	541	315
D2S441	132398	249292	139692	276	607	152
D3S1358	139117	319652	125776	105	381	245
D5S818	94876	246221	84978	119	295	101
D7S820	201743	484694	209136	146	543	321
D8S1179	310130	639179	284904	211	506	206
FGA	105665	307855	144603	219	533	522
Penta D	198479	448801	222545	278	635	500
Penta E	262278	663597	301355	92	248	197
TH01	34316	77396	33685	46	83	40
TPOX	124382	202296	101818	195	405	81
vWA	171787	422688	183910	162	466	184

CHAPTER 3

Expanding the capabilities of STRspy to profile markers on the Y

C.L. Hall
R.K. Kesharwani
N.R. Phillips
J.V. Planz
F.J. Sedlazeck
R.R. Zascavage

HIGHLIGHTS

- The ONT MinION device supports simultaneous sequencing of autosomal and Y-STRs.
- STRspy produced accurate Y-STR profiles for all samples amplified at 30 PCR cycles.
- At least 24 amplicon libraries can be sequenced on a single MinION flow cell and correctly typed by STRspy across the largest combined panel of autosomal and Y-STRs.

CHAPTER OVERVIEW

Forensic DNA examinations harness STRs on autosomal and Y chromosomes for human identification. Length-based Y-STR profiles are generated using the same methods as autosomal STRs but require separate sample normalization as well as PCR and CE reactions. This consumes often limited DNA evidence and creates backlog by prolonging the period in which a case is being processed. The high sample throughput and enhanced multiplexing capabilities of NGS platforms enable more powerful STR profiles to be produced in less time compared to conventional typing techniques. Despite the advantages of NGS in forensic DNA examinations, the initial investment of established platforms for STR sequencing is too steep for most small to mid-sized laboratories. Nanopore sequencing on the affordable and portable MinION device could provide an effective alternative for uncovering hidden variation in current STR profiles. To harness these unique features in forensic applications, we developed and presented STRspy in Chapter 2. STRspy is the first bioinformatic method that can predict accurate sequence- and length-based allele designations across an entire panel of autosomal STRs using error-prone ONT reads [51]. In this chapter, we expand upon the STRspy framework to enable simultaneous profiling of autosomal and Y-STRs in male samples The updates implemented into the allele database and STRspy script were first evaluated using the 15- and 30-cycle datasets produced in Chapter 2. We also assessed the effect of sample multiplexing on STRspy profile predictions across the entire panel of autosomal and Y-STRs targeted in the Promega PowerSeq 46GY System. Four control DNAs amplified at 30 PCR cycles were pooled in sets of 12, 18, and 24 barcodes per flow cell and sequenced on the MinION for 72hrs. Basecalled reads were analyzed with the updated version of STRspy and resultant allele designations were compared to the manufacturer-validated profiles. STRspy predicted the correct genotypes across all 22 autosomal and 23 Y-STRs based on both length and sequence. The data presented herein demonstrate that STRspy can produce

accurate profiles for the largest combined autosomal and Y-STR amplification panel and supports multiplexing of at least 24 samples per flow cell. Expanding our method to harness more STRs in larger sample multiplexes decreases cost and increases the forensic potential of the MinION device in future applications.

MATERIALS & METHODS

MULTIPLEXING

Samples. The multiplexing experiment was conducted using three NIST traceable standards and one Promega control (female n = 1; male n = 3) with manufacturer-validated CE and NGS profiles. NIST A, B, and C (SRM 2391d) were quantified on the Qubit 2.0 Fluorometer using the Qubit dsDNA BR Assay (Thermo Fisher Scientific) and diluted to 0.1ng/μL in amplification grade water. The positive control included in the Promega PowerSeq 46GY System (2800M) was prepared and normalized as per manufacturer recommendations. The Qubit 1X dsDNA HS Assay (Thermo Fisher Scientific) was used to confirm the final concentration of all control DNAs before STR amplification.

STR amplification. The 22 autosomal and 23 Y-STRs in the Promega PowerSeq 46GY System (PS4600) were amplified for ONT sequencing using 0.5ng of DNA. Amplification was performed with the recommended thermal profile at 30 cycles on the Eppendorf Mastercycler pro S. Resultant amplicons were then processed with the QIAquick PCR Purification Kit (Qiagen) according to the microcentrifuge protocol. A 10μL aliquot of 3M sodium acetate (pH 5.0) was added to all samples before column binding due to the observed change in color of the pH indicator. DNA was eluted in 50μL of nuclease-free water to produce 48μL of purified amplicons for ONT library preparation.

ONT library prep & sequencing. STR libraries were prepared using the ONT Ligation Sequencing Kit (SQK-LSK109) with Native Barcoding Expansions 1-12 (EXP-NBD104) and 13-24 (EXP-NBD114) as per the modifications described in Chapter 2. Purified amplicons from one PCR reaction (48µL) were used as the input for ONT library preparation. Following end-repair and dA-tailing, unique barcodes were ligated onto both amplicon ends in samples to be sequenced together. The multiplex experiment was performed using stock solutions of barcoded samples to eliminate potential variation in library preparation. The four control DNAs were labeled using all 24 barcodes available for the ligation-based workflow**.** To ensure that sufficient stock solution was available to sequence and resequence different multiplex combinations if needed, six amplicon libraries were prepared and pooled for each barcode. Details regarding sample barcoding and multiplexing per MinION flow cell are provided in **Supplemental Table S4**. Bead-purified amplicon libraries were then quantified on the Agilent TapeStation 4200 with D1000 ScreenTapes and combined according to the concentration of fragments ranging from 175bp to 475bp (**Supplemental Figure S4**). Barcode pools exceeding 65µL were concentrated in an Eppendorf 5301 Vacufuge System. After ligation of ONT sequencing adapters, amplicon libraries were purified using magnetic beads with two washes in short fragment buffer (SFB, ONT). Pooled amplicon barcodes were then quantified and diluted in elution buffer (EB, ONT) to 75ng based on overall concentration before preparing final loading libraries (**Supplemental Figure S4**). Prepared sequencing libraries were loaded via the SpotON port of primed MinION vR9.4.1 flow cells (FLO-MIN106D, ONT) and sequenced on the MinION device with the ONT MinKNOW software. All runs were performed for 72hrs regardless of throughput.

The raw current disruptions recorded on the MinION device (fast5) were converted to nucleotide sequences (fastq) using the GPU-enabled Guppy basecaller (v3.4.2) with the high accuracy model. Guppy was also used to demultiplex and merge reads based on barcode. Merged

fastq files containing reads with a mean q-score value greater than 9 were then processed with the STRspy command line interface.

CYCLE NUMBER

Y-STR cycle number analyses were performed on the nanopore sequencing data generated in Chapter 2. Similar materials and methods were used to prepare samples for the cycle number and multiplexing experiments. The cycle number dataset includes three additional control DNAs (that have since been discontinued) for a total of seven samples amplified in triplicate at 15 and 30 PCR cycles and purified with magnetic beads [38]. Another key difference was the number of barcoded amplicon libraries loaded and sequenced on each MinION flow cell (**Supplemental Fig. S1**). Low amounts of STR amplicons are produced with 15 PCR cycles. These libraries were therefore barcoded and multiplexed in sets of 12 for sequencing on the MinION device. Final prepared libraries for the 15-cycle datasets were under 75ng and run throughput was low with no indication of pore clogging. The twofold increase in PCR cycle number resulted in an exponentially higher concentration of short amplicons that can clog nanopore proteins and cause a rapid decline in flow cell health (**Supplemental Fig. S5**). Given that the depth of coverage needed for accurate STR profiling had not been established, we multiplexed three to four samples for each of the triplicates amplified with 30 PCR cycles.

STRSPY

Implementation. STRspy predicts forensic autosomal STR and Y-STR profiles from third-generation sequencing data. It executes the following steps in a per sample manner:

1. Align unmapped basecalled reads to the entire human reference genome.
2. Extract reads that overlap STR loci based on the user-provided bed file.
3. Map locus-specific reads to the collection of alleles within the custom STR database.

4. Count number of reads aligned to each sequence-based STR allele with mapping quality greater than 1.
5. Calculate locus-specific normalized read counts (number of reads per allele divided by the highest number of reads mapping to a single allele).
6. Rank alleles at each STR locus based on normalized read counts.

A detailed account of each step is provided in Chapter 2. The remainder of this section focuses on the updates that were implemented in the STRspy framework and user-provided files to enable prediction of Y-STR haplotypes in addition to autosomal genotypes.

Allele reporting. STRspy harnesses normalized read counts to rank the sequence-based alleles detected at each STR of interest. The balance of autosomal alleles is used to predict whether the locus is homozygous (reports top allele) or heterozygous (reports top two alleles) according to the user-defined normalization threshold. The default cutoff of STRspy is set to 0.4 based on the benchmarking results presented in Chapter 2 (**Fig. 9**). We have expanded the capabilities of STRspy to predict both autosomal and Y-STR profiles from third-generation sequencing reads. For all Y-STRs except DYS385, STRspy reports the allele with the highest normalized read count. DYS385a and DYS385b represent duplications of DYS385 with identical flanking region sequences [52]. These loci are amplified with the same PCR primer pair in convention amplification reactions [52]. After genome-wide mapping and extraction of locus-specific reads, STRspy merges DYS385a and DYS385b aligned reads and reports the top two alleles.

Database. STRspy reports bracketed repeat motifs as well as length-based allele designations consistent with conventional CE profiles. In addition to basecalled reads in fastq or bam format, users must provide the chromosomal location (bed file) and sequence-based alleles (fasta file) at STR loci of interest. STRspy relies on the STR database to identify sequence-based alleles in

ONT or PacBio reads while also reporting length-based CE designations. The STR allele database developed in Chapter 2 was limited to the 22 autosomal loci amplified by the Promega PowerSeq 46GY System [28]. Here we update both the database and bed file to include the 23 Y-STR loci also targeted in this panel. The collection of sequence-based alleles that were used to construct the autosomal STR database does not contain information about Y-STRs. We therefore harnessed the STR Fact Sheets available on the NIST STRBase website for preliminary testing. Sequence-based alleles in the Y-STR database were generated using the repeat structure and number provided in the STR Fact Sheets. These bracketed repeat motifs were transformed into sequence strings with a stepwise permutation model (**Supplemental Fig. S6**). As with the autosomal STR database, 500bp flanks were retrieved from the human reference genome and appended onto each sequence-based allele to facilitate alignment of shorter nanopore reads. Database entries were labeled with the locus, bracket repeat motif, and length-based designation.

DATA ANALYSIS

STRspy outputs allele designations consistent with the established forensic naming system as well as the raw and normalized read counts supporting the prediction (**Supplemental Fig. S2**). We modified the utilities developed in Chapter 2 to assess concordance between STRspy-predicted and manufacturer-validated allele designations across both the autosomal and Y-STRs. Each allele that STRspy reported in the final profile was categorized as a true positive (correct allele), false positive (incorrect allele), or false negative (missing allele). These counts were used to calculate the precision, recall, and F1 score of our method. Separate benchmarking was performed for the autosomal and Y panels to assess the updates implemented in this chapter. Precision and recall were determined by dividing the number of true positives by the total alleles in the STRspy (true positive + false positive) or ground truth (true positive + false negative)

profiles, respectively. The overall performance of STRspy for autosomal and Y-STRs was evaluated based on F1 score (harmonic mean of precision and recall).

RESULTS

SIMULTANEOUS PROFILING OF AUTOSOMAL & Y-STRS

Few studies have assessed nanopore sequencing in the context of forensic STRs, none of which achieved correct profiles across the entire panel of autosomal and Y loci [35,36,53]. In addition to 22 autosomal STRs, the PowerSeq 46GY System amplifies 23 STRs on the Y chromosome. We were therefore able to reanalyze the five male DNA controls amplified in triplicate at 15 and 30 cycles from Chapter 2. These data were used to determine whether Y-STRs can be amplified and sequenced alongside autosomal STRs on the MinION device. We first compared the number of Y-STR mapped reads in the 15- and 30-cycle datasets. As with the autosomal targets, per locus coverage across STRs on the Y chromosome varied based on PCR cycle number (**Supplemental Table S5**). The number of reads that mapped to alleles in the Y-STR database for samples in the 30-cycle dataset ranged from 13207 to 63272 with an average of 322998. A lower number of supporting reads was observed across Y-STRs amplified with 15 PCR cycles which ranged from 5 to 531 and averaged 268. The observed difference in Y-STR aligned reads is consistent with the exponential increase in STR amplicons produced with each additional cycle of PCR. Nonetheless, these data suggest that short Y-STR fragments generated with commercial NGS amplification kits can be sequenced alongside autosomal STRs on the MinION device to obtain sufficient coverage across the 23 Y-STR targets for subsequent profiling with STRspy.

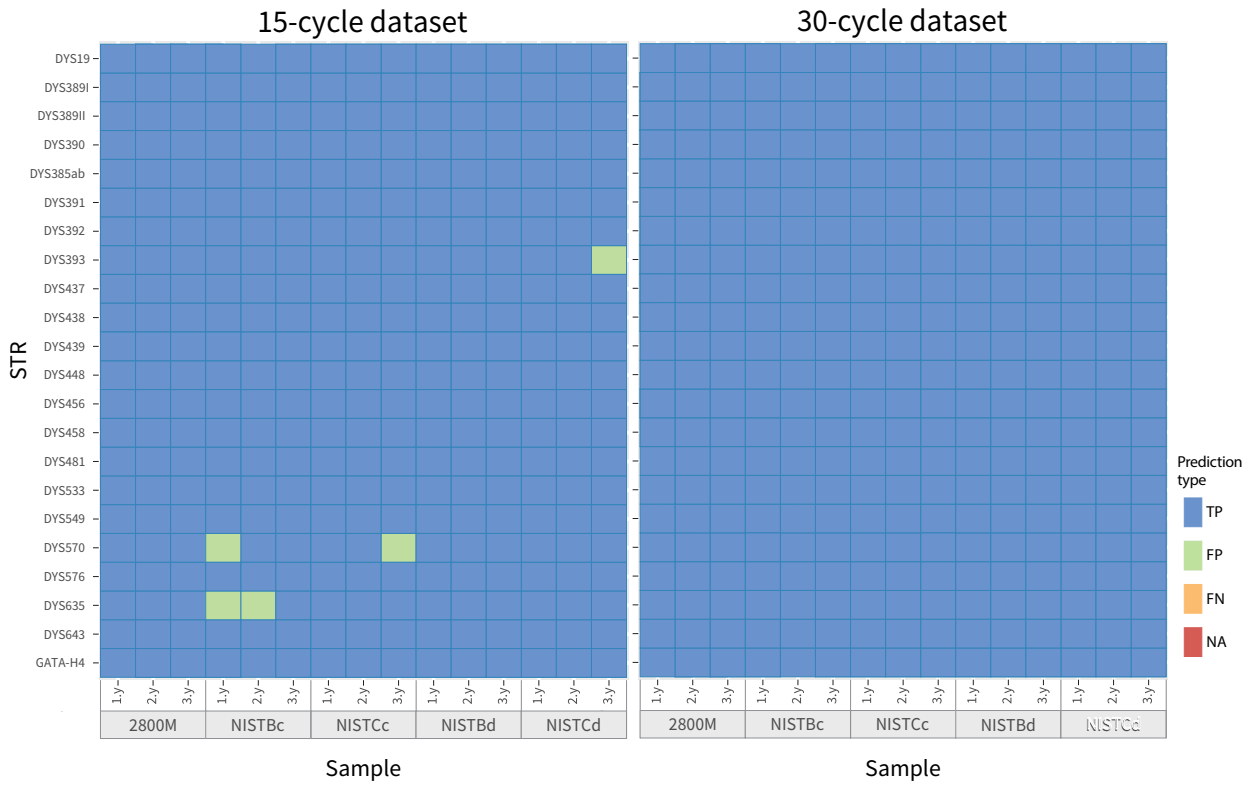


Fig. 12 Y-STR profile predictions. Heatmap comparison of STRspy predictions to manufacturer-validated length- and sequence-based genotypes across the 15-cycle and 30-cycle datasets. True positive (TP) predictions are depicted in blue, false positives (FP) in green, and false negatives (FN) in orange. Reference samples (grey boxes) are labeled by triplicate (1, 2, 3).

To harness all the genetic information contained within commercial NGS amplification kits, we expanded the STRspy framework to support simultaneous profiling of autosomal and Y-STRs. These updates were first assessed in the 15- and 30-cycle datasets to evaluate the impact of cycle number on the accuracy of Y-STR profiles generated with STRspy. Our method predicted the correct allele designations based on both length and sequence for the 23 Y-STRs amplified at 30 PCR cycles (**Fig. 12**). STRspy achieved complete concordance with manufacturer-validated profiles for the five male triplicates in the 30-cycle dataset, resulting in a recall, precision, and F1 score of 100%. Moreover, our method was able to resolve Y-STRs of the same length with different underlying sequences. STRspy reported the correct bracketed repeat motifs for

DYS391II isoalleles with repeat length of 31 for component B from NIST SRM 2391c (NISTBc: [TCTG]6 [TCTA]12 N48 [TCTG]3 [TCTA]10), component C from NIST SRM 2391d (NISTCd: [TAGA]9 [CAGA]3 N48 [TAGA]13 [CAGA]6), and the Promega control (2800M: [TCTG]4 [TCTA]13 N48 [TCTG]3 [TCTA]11) triplicates (**Supplemental File S2**). STRspy can therefore reveal nucleotide-level variation in ONT reads to produce more powerful Y-STR profiles than conventional CE approaches.

We also evaluated haplotype predictions for the same five male DNA controls amplified with 15 PCR cycles. STRspy reported the correct calls for 310 of the 315 alleles in the 15-cycle dataset, achieving a precision, recall, and F1 score of 98.41% (**Fig. 12**). Details about the three loci and five incorrect allele designations are described in **Supplemental File S2**. Reduced amplification resulted in a low number of reads to support Y-STR allele designations. All false positive predictions were one repeat unit less than the true allele and thus overlap with the most common type of PCR-induced stutter artifact (-1 stutter) [54]. These genetic markers also feature homopolymeric stretches either within or around the repeat region known to accumulate partial deletions in nanopore sequencing reads [34,55]. These ONT-specific errors could hinder accurate alignment to the correct allele in the STR database and result in false positive predictions. Nonetheless, STRspy reported the correct alleles for most Y-STRs even with low coverage. Although our method was able to resolve all isoalleles in the 15-cycle dataset, these results indicate that more than 15 cycles of PCR are required to consistently produce accurate STR profiles from ONT sequencing data. The multiplexing experiment was therefore conducted using samples amplified with 30 PCR cycles.

MAXIMUM SAMPLES PER MINION FLOW CELL

NGS platforms provide higher sample throughput and enhanced locus multiplex capabilities over conventional CE typing techniques. Three to four samples were pooled and sequenced on each MinION flow cell for the triplicates amplified with 30 PCR cycles. The high level of coverage and accurate profiles achieved across all autosomal and Y-STRs in the 30-cycle dataset suggest that more samples can be multiplexed per run. We therefore aimed to determine the maximum number of amplicon libraries that can be loaded onto a single MinION flow cell and correctly typed with STRspy. Stock solutions of amplicon barcodes were pooled and sequenced in sets of 12, 18, and 24 samples on the MinION device for 72hrs (**Supplemental Table S4**). The number of STR-aligned reads decreased with increasing multiplex size. This observation is consistent with the notion that DNA fragments compete for pore access during the sequencing run [3]. The highest level of sample coverage was obtained in the 12-sample multiplex (**Fig. 13**). Although about half the number of mapped reads were produced for amplicon libraries in the 24-sample multiplex, STRspy predicted the correct allele designations across all 22 autosomal and 23 Y-STRs.

These data also provided novel insight into variation between library preparations for the same sample**. Fig. 13** depicts the raw number of reads mapping to D2S441 isoalleles in the six barcodes for component B from NIST SRM 2391d (NISTBd). D2S441 isoalleles were balanced within the NISTBd barcodes across the three multiplexes. The number of reads supporting each sequenced-based allele for this length-based 11 homozygote ranged from 2055 to 2957 with an average of 2571. The lowest coverage was observed for barcode 18 in all three multiplexes, suggesting inefficient ligation or reduced basecalling and demultiplexing of this barcode. The example depicted in **Fig. 13** also demonstrates that STRspy can unambiguously differentiate between isoalleles in multiplexes of at least 24 samples. These results suggest that accurate and

reproducible profiles can be generated for the largest multiplex tested in forensic STR sequencing applications to date.

These data also provided novel insight into variation between library preparations for the same sample**. Fig. 13** depicts the raw number of reads mapping to D2S441 isoalleles in the six barcodes for component B from NIST SRM 2391d (NISTBd). D2S441 isoalleles were balanced within the NISTBd barcodes across the three multiplexes. The number of reads supporting each sequenced-based allele for this length-based 11 homozygote ranged from 2055 to 2957 with an average of 2571. The lowest coverage was observed for barcode 18 in all three multiplexes, suggesting inefficient ligation or reduced basecalling and demultiplexing of this barcode. The example depicted in **Fig. 13** also demonstrates that STRspy can unambiguously differentiate between isoalleles in multiplexes of at least 24 samples. These results suggest that accurate and reproducible profiles can be generated for the largest multiplex tested in forensic STR sequencing applications to date.

DISCUSSION

In the previous chapter, we presented the first forensic-specific bioinformatic method for generating CODIS-compatible autosomal STR profiles from third-generation sequencing data [51]. Here we expand upon the STRspy framework as well as the user-provided allele database to enable simultaneous profiling of autosomal and Y-STRs. We demonstrate that the updates implemented in our method produce accurate allele designations for the five male control DNAs amplified at 30 PCR cycles. With these updates, users can now harness third-generation sequencing reads in larger loci panels to achieve more powerful STR profiles than CE and even the first version of STRspy.

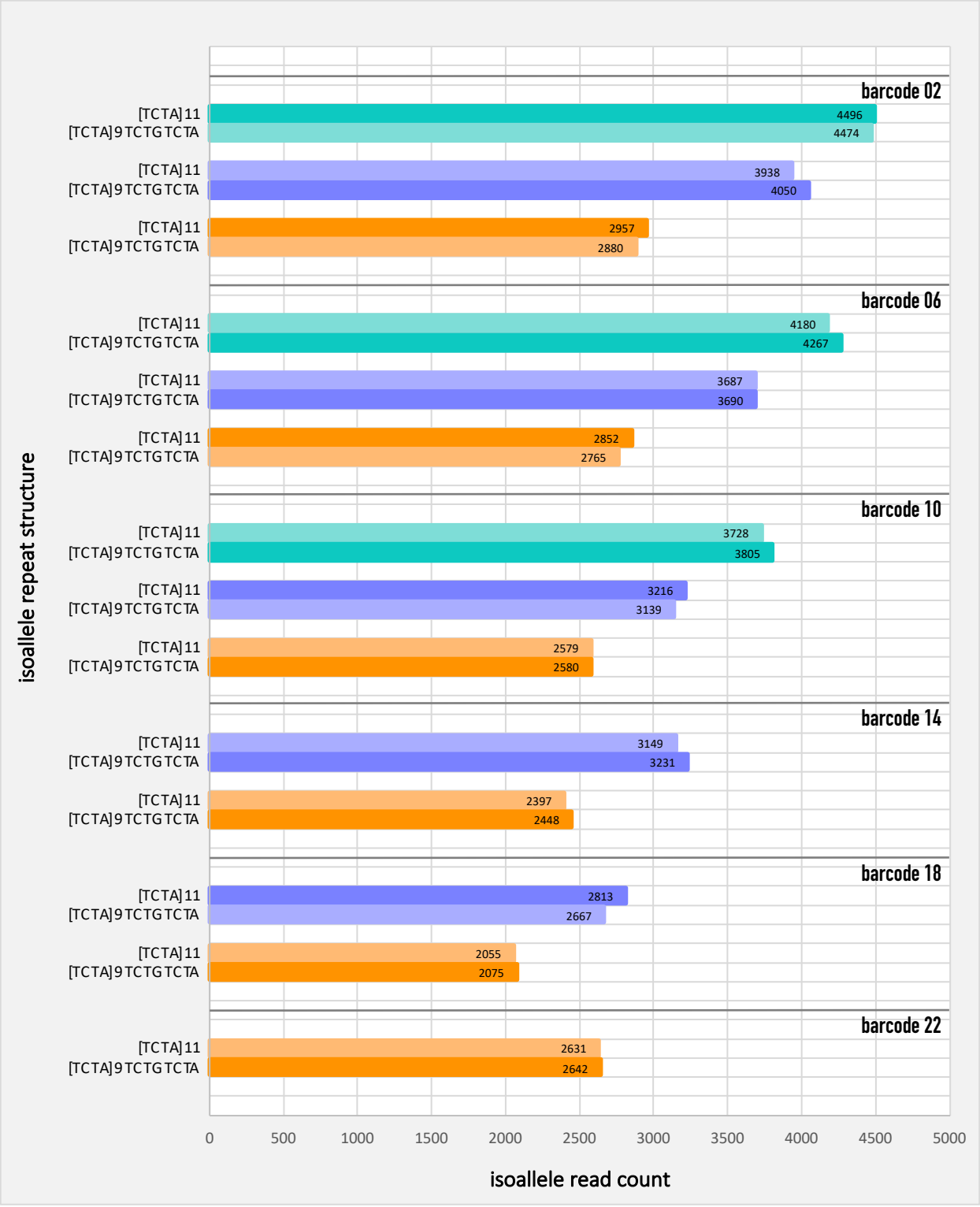


Fig. 13 STRspy can resolve isoalleles in 24-sample multiplexes. The number of raw reads supporting sequence-based isoalleles with repeat length of 11 at D2S441 in component B from NIST SRM 2391d (NISTBd). Bars are grouped by barcode and colored based on multiplex (12 samples = teal, 18 samples = purple, 24 samples = orange). Darker shading represents the isoallele with the highest supporting read count.

Despite the low coverage obtained across Y-STR loci in the 15-cycle dataset, only five incorrect allele designations were reported by STRspy. All false positive predictions were one repeat unit shorter than the true allele with few supporting reads. These incorrect calls are consistent with -1 stutter which is the prevalent artifact of STR amplification [54]. The combination of low coverage and stutter artifacts across these loci is one potential explanation for the false positives reported by STRspy. The three loci with incorrect allele designations also feature homopolymers in the repeat and flanking regions. Homopolymer-containing STRs are known to accumulate partial deletions that can complicate alignment and preclude detection of the true allele. These results suggest that more than 15 PCR cycles are required to produce accurate allele designations across all Y-STRs. Additional studies are needed to determine the level of coverage required to overcome PCR-induced stutter and other nanopore sequencing errors throughout forensic STR panels. This information would provide the foundation for generating consistent and accurate STR profiles across all loci. It can also guide future efforts that aim to harness the adaptive sampling capabilities of nanopore sequencing platforms with ReadUntil to further maximize flow cell usage (see Chapter 4) [56].

With a startup fee of $1k, the MinION device represents an affordable alternative to established STR sequencing platforms. This low initial investment is countered by the current high price of ONT consumables. To reduce cost, samples can be barcoded for simultaneous sequencing on one MinION flow cell. Although multiplexing amplicon libraries results in higher throughput and lower cost, it also increases competition for pore access and decreases the number of reads produced for each sample. One of the main aims of this chapter was to determine the maximum number of samples that can be sequenced and accurately profiled with STRspy. These results confirmed that the laboratory and bioinformatic methods we developed support multiplexing of at least 24 samples (which is the maximum number of ligation barcodes

available). ONT also offers a PCR-based barcoding kit that can accommodate up to 96 samples [32]. Additional amplification of low complexity repeats such as STRs is not ideal. It may be possible to scale back on the 30 PowerSeq cycles used herein to account for the amplification performed during PCR barcoding reactions. Future studies should assess the potential to load more STR amplicon libraries per MinION flow cell to further increase sample throughput while reducing cost.

STRspy predictions are based on the user-provided STR allele database. It is therefore critical that the allele sequences and associated length-based designations in the database are both correct and comprehensive. STRspy was able to achieve complete concordance for all samples amplified with 30 PCR cycles. However, the Y-STR database used herein was constructed by permutating each repeat unit reported in the STRBase Fact Sheets. Although this well-established collection also contains rare variants, it is based on CE profiling data and thus may lack a subset of sequence-based Y-STR alleles. Our current efforts are focused on updating both the autosomal and Y-STR databases to contain all loci and alleles reported in the common autosomal and Y-STR subdivisions of the STRSeq BioProject [57]. The updated database will represent the most comprehensive collection of validated autosomal and Y-STR alleles based on NGS data for over four thousand individuals [57]. Users can also expand upon our database or construct their own, but this process requires conversion of sequence-based allele information to an STRspy-compatible format. The same STR database can be used to profile all unknown samples of interest but the process of generating and updating this collection of alleles can be time-consuming for users with limited computational skills. We are therefore developing and testing a new script that can construct custom STR allele databases from GenBank records in the STRSeq BioProject. We intend to release this script with the next iteration of STRspy to make our method more accessible to forensic DNA analysts.

Y-STR amplicons harbor additional variation in the flanking regions that were not assessed in the current study [58,59]. The autosomal analyses presented in the previous chapter suggest that our approach to SNP calling requires further improvements, especially for samples amplified at 30 PCR cycles. One potential solution is integration of SNPs into the Y-STR database. Attempts to implement this approach for autosomal STRs were unsuccessful due to the large number of sequence-based alleles in the STRSeq BioProject. The Y-STR subdivision is much smaller with less flanking variation and could thus enable accurate profiles to be generated using a SNP-aware STR allele database.

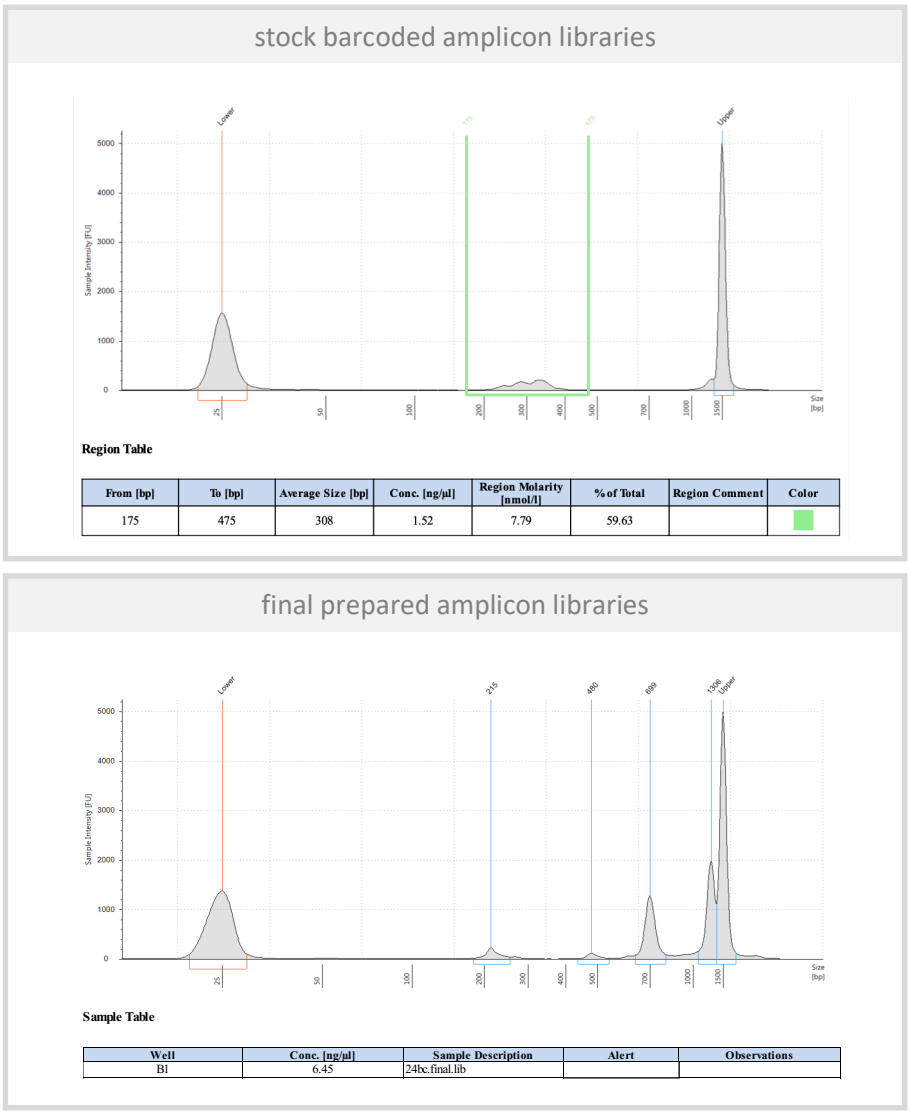
All samples analyzed up to this point are high quality reference materials. Ongoing research from our group aims to determine whether STRspy can generate CODIS-compatible length-based genotypes while revealing additional nucleotide-level variation in biological materials often encountered in forensic investigations (buccal swab, blood, and bone). Sequencing metrics suggest that data were successfully generated for these sample types in recent experiments. We are currently working to improve our database and SNP detection approach before analyzing the casework-relevant samples.

CONCLUDING REMARKS

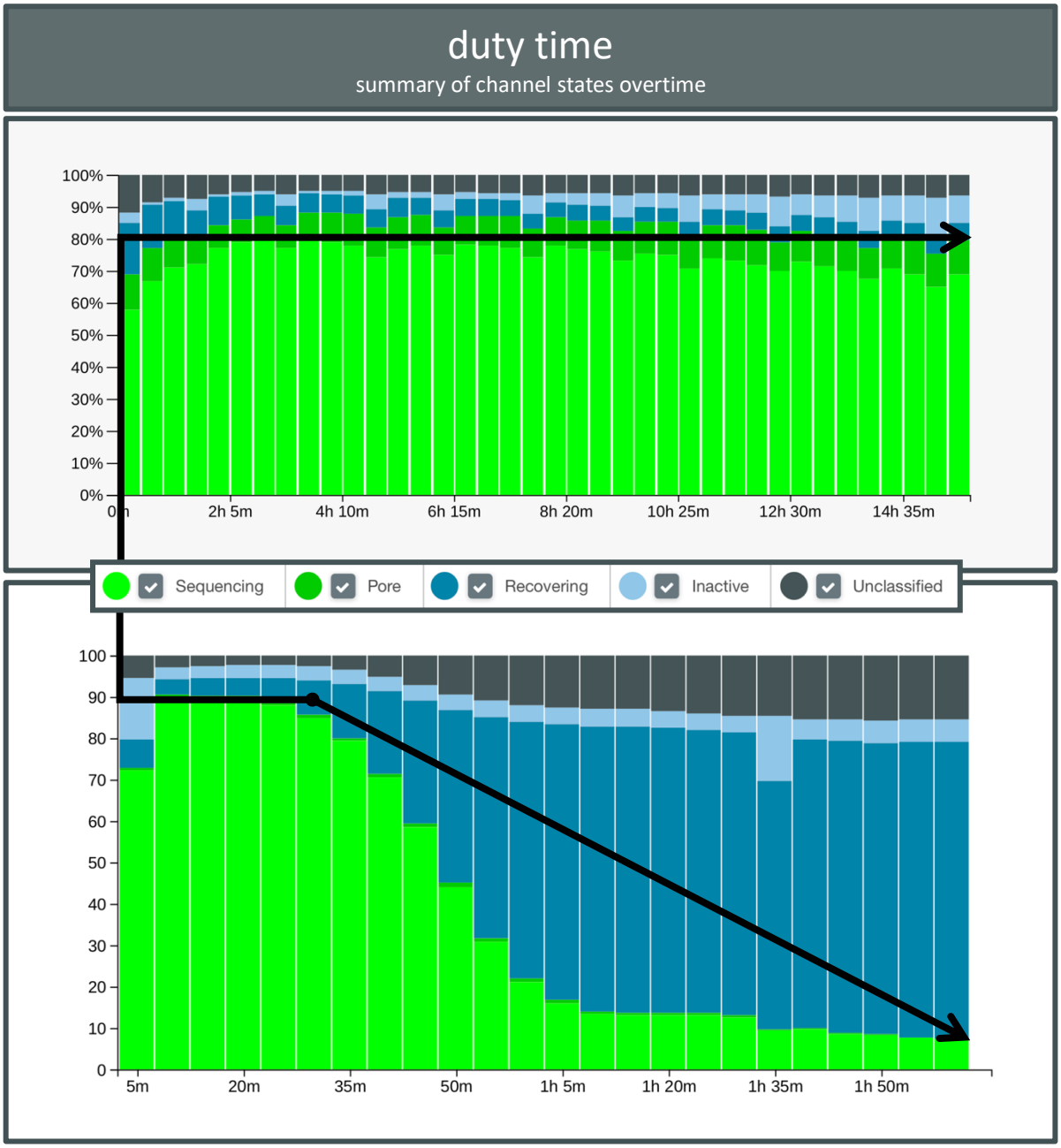
This chapter aimed to profile Y-STRs using nanopore sequencing data produced on the ONT MinION device. The updates implemented in our method achieved 100% concordance across all 23 Y-STRs amplified with the Promega PowerSeq 46GY System at 30 PCR cycles. We expanded upon our cycle number study by conducting a comprehensive multiplexing experiment to assess how sample number impacts autosomal and Y-STR profiles. These results demonstrate that at least 24 samples can be sequenced on MinION flow cells and profiled with STRspy. Our method generated accurate profiles across the entire PowerSeq panel rather than a subset of autosomal

and Y-STRs. Continued improvements in nanopore sequencing technologies along with further development of STRspy could make forensic STR sequencing more accessible, feasible, and affordable on the ONT MinION device.

FIGURES



Supplemental Fig. S4 Barcoded amplicon stock and multiplex tapes. PowerSeq amplicons were barcoded and pooled based on the concentration of fragments ranging from 175bp–475bp (green) produced on the Agilent TapeStation 4200 with D1000 screentapes. Prepared multiplexes were quantified before loading and diluted to 75ng if needed.



Supplemental Fig. S5 Summary of nanopore channel states during amplicon sequencing runs on the MinION device. Final prepared libraries were loaded onto MinION flow cells at 75ng (top) and 132ng (bottom). The rapid decline in sequencing and active pores (dark and light green) and increase in recovering pores (dark blue) is indicative of pore clogging.

Supplemental Fig. S6 Permutation approach used to construct the Y-STR allele database. Bracketed repeat motifs from STRBase Fact Sheets (blue) were expanded and contracted (orange) to produce the text string of nucleotides for each allele in our Y-STR database (teal). Flanking sequencing of 500bp were retrieved from the human reference genome and appended to the repeat sequences (purple).

FILE

Supplemental File S2 Per allele read counts generated by STRspy for Y-STRs.

TABLES

Supplemental Table S4. Sample and barcode combinations per MinION flow cell for the multiplexing experiment.

	samples per MinION flow cell		
barcode	12	18	24
01	NISTAd	NISTAd	NISTAd
02	NISTBd	NISTBd	NISTBd
03	NISTCd	NISTCd	NISTCd
04	2800M	2800M	2800M
05	NISTAd	NISTAd	NISTAd
06	NISTBd	NISTBd	NISTBd
07	NISTCd	NISTCd	NISTCd
08	2800M	2800M	2800M
09	NISTAd	NISTAd	NISTAd
10	NISTBd	NISTBd	NISTBd
11	NISTCd	NISTCd	NISTCd
12	2800M	2800M	2800M
13		NISTAd	NISTAd
14		NISTBd	NISTBd
15		NISTCd	NISTCd
16		2800M	2800M
17		NISTAd	NISTAd
18		NISTBd	NISTBd
19			NISTCd
20			2800M
21			NISTAd
22			NISTBd
23			NISTCd
24			2800M

Supplemental Table S5 Reads mapping to Y-STR alleles for the male cycle number triplicates. Number of STR aligned reads are colored by cycle number and shaded by relative abundance.

	2800M						NISTBc					
	30-cycle			15-cycle			30-cycle			15-cycle		
locus	1	2	3	1	2	3	1	2	3	1	2	3
DYS19	132038	148068	112377	145	309	152	268321	205567	192488	97	339	316
DYS385ab	303706	396890	288567	150	365	335	564361	361020	573598	160	471	228
DYS389I	309057	402965	293591	152	368	327	597435	387083	612477	160	505	241
DYS389II	223637	255008	218146	120	273	340	350376	404148	376503	156	370	227
DYS390	36161	83976	46628	74	138	98	103905	89929	107043	55	148	143
DYS391	18320	28422	33259	62	103	39	73343	40889	54987	37	90	95
DYS392	41111	62708	56968	19	64	11	33778	33218	36870	15	70	35
DYS393	150288	126239	124126	142	295	216	292314	239305	265153	170	335	232
DYS437	14175	13270	19505	29	49	27	50651	39474	35533	28	91	55
DYS438	19845	29897	21527	19	49	37	43456	29649	30866	16	64	56
DYS439	165117	160362	190017	76	207	359	357109	254054	473978	100	308	172
DYS448	70158	129340	71438	76	220	98	211913	125393	115274	59	163	107
DYS456	85042	87901	93774	109	238	250	223917	224111	229367	95	250	203
DYS458	43464	80197	47494	52	164	111	151637	108506	141348	66	153	181
DYS533	202039	242991	167771	118	326	285	400840	389657	349872	116	343	291
DYS549	177906	161220	149659	133	224	223	166287	238178	296913	134	362	253
DYS570	62974	62015	47822	67	98	129	141455	91455	114311	54	128	151
DYS576	62013	72183	69212	101	201	213	130714	73690	114515	73	211	224
DYS635	21944	25161	20219	37	85	30	56322	32141	52087	29	71	63
DYS643	142968	123169	104909	183	387	451	237421	153608	222026	119	381	326
GATA-H4	270547	540435	232458	212	425	287	437920	448254	500517	130	440	357

	NISTCc						NISTBd					
	30-cycle			15-cycle			30-cycle			15-cycle		
locus	1	2	3	1	2	3	1	2	3	1	2	3
DYS19	195949	164474	216702	113	262	258	183724	265573	140480	125	450	242
DYS385ab	480491	512656	479075	110	402	305	326125	452479	291641	155	381	298
DYS389I	514067	548550	512560	111	442	330	346613	484684	311958	158	411	312
DYS389II	357289	373756	441639	85	233	264	197164	410884	238974	133	328	304
DYS390	118372	87526	160116	47	140	207	47759	98667	58518	68	203	171
DYS391	60802	61852	59800	39	73	54	18501	38298	41201	33	125	61
DYS392	40933	33428	47776	16	48	55	23224	36332	20377	29	47	38
DYS393	190648	205335	274237	104	244	265	137620	196440	218240	126	342	173
DYS437	38915	27140	40412	26	57	51	23602	33081	30748	50	89	63
DYS438	37822	26023	60282	20	55	44	22886	46800	17806	42	82	53
DYS439	205938	252135	298004	56	210	224	244810	278768	262857	126	272	133
DYS448	140877	111022	169234	84	138	145	68052	136358	87094	81	179	147
DYS456	162127	131205	233261	74	195	163	105306	156775	88894	128	267	199
DYS458	167806	116030	186850	71	154	156	47797	152511	78306	89	251	192
DYS533	338392	295195	429026	108	269	339	231213	520472	222386	129	361	221
DYS549	228562	182075	288086	103	216	348	158809	264970	147938	129	356	270
DYS570	118804	91290	142541	54	116	155	59935	89189	60678	50	133	122
DYS576	105409	81404	106577	52	144	163	43546	104487	59775	71	215	116
DYS635	35756	21649	51312	25	72	62	21019	41776	22446	37	90	63
DYS643	215211	162897	270312	89	283	304	158360	228472	137794	147	433	346
GATA-H4	412687	536320	598431	206	357	518	268042	566195	301310	214	493	485

Supplemental Table S5 (continued) Reads mapping to Y-STR alleles for the male cycle number triplicates. Number of STR aligned reads are colored by cycle number and shaded by relative abundance.

	NISTCd					
	30-cycle			15-cycle		
locus	1	2	3	1	2	3
DYS19	139482	320757	120739	163	376	135
DYS385ab	269461	521816	246419	129	504	338
DYS389I	278248	535405	254470	133	531	346
DYS389II	213038	441807	237404	152	311	389
DYS390	57667	103742	47458	94	257	66
DYS391	16819	43405	28298	77	127	27
DYS392	23604	50632	18939	26	82	5
DYS393	110611	216051	153187	129	397	213
DYS437	10828	26333	29488	50	66	69
DYS438	17329	38802	18024	31	55	21
DYS439	106697	273754	171030	119	336	406
DYS448	81439	142278	64640	88	213	47
DYS456	101158	236099	112393	138	305	208
DYS458	80123	176654	96055	138	303	165
DYS533	287962	632727	240910	190	379	292
DYS549	150549	308639	142588	139	352	174
DYS570	45167	101265	50026	65	177	139
DYS576	62191	146636	57192	83	245	206
DYS635	24786	49826	24270	39	75	36
DYS643	126908	302917	158609	190	402	498
GATA-H4	304496	612235	319338	233	487	356

CHAPTER 4

The MinION device: a small sequencer with big potential in forensic DNA examinations

CHAPTER OVERVIEW

The research presented herein forms the foundation for future efforts that aim to harness nanopore sequencing technologies in forensic DNA examinations. The limitations of each experiment are discussed in the respective chapter. This chapter highlights potential solutions to current challenges of nanopore sequencing in forensic investigations. Implementation of novel ONT features and devices, further development of streamlined workflows, and expansion of our methods to other biological materials and genetic marker systems could facilitate forensic adoption of the MinION device in the future.

MORE ACCURATE READS

High error rate is the most cited limitation of nanopore sensing systems [27,50,60]. Significant efforts have therefore been geared towards development of new sequencing chemistries and analytical methods capable of producing more accurate raw read data. Available flow cells are composed of nanopores from either the R9 or R10 series [31]. Structural differences between R9 and R10 proteins can impact the quality of resultant reads. Nanopore sequencing relies on current disruptions as unique combinations of nucleotides (or k-mers) pass through individual pores. The reader head is the position at which nucleotides within the DNA strand have the most influence on the current [61]. The R9.4 version used throughout the current project features a single reader head that spans k-mers of three to five nucleotides [31]. Although sufficient for complex genomic sequences, this reader head is unable to resolve longer stretches of identical k-mers in homopolymeric and repeat regions [50,60,61]. R10 nanopore proteins have an elongated neck with an additional reader head that has been shown to improve basecalling through these low complexity sequences [62].

We and others have highlighted the potential benefits of R10 nanopore proteins in forensic STR sequencing [3,34,36,51,53]. The first R10 flow cell released was shown to achieve higher resolution through homopolymer-containing STR amplicons [36]. Despite producing more accurate reads, genotype predictions did not improve, resulting in profiles that were comparable to R9.4 data [36]. A subsequent publication from this group reported correct profiles across all autosomal loci using a combination of supported and developmental algorithms to base call current disruptions from R10.3 flow cells [53]. Interestingly, these researchers attributed the inability to produce accurate profiles with a single basecaller to the R10 nanopore protein itself [53].

One potential limitation of the data generated in both studies was the low throughput and coverage [36,53]. Short fragments can clog nanopore proteins and cause a rapid decline in flow cell health. ONT therefore recommends loading approximately 20ng of prepared STR amplicon libraries [36,53]. Experiments that we conducted before ONT-supported amplicon sequencing suggest that much higher amounts of STRs (up to 75ng) can be loaded onto R9.4 MinION flow cells. The throughput achieved for our 30 PCR cycle dataset is orders of magnitude higher than those reported for both R9 and R10 flow cells in previous studies [36,53]. We therefore maintain that the newest R10 flow cells and library preparation kit (for which ONT has reported the highest raw read accuracy to date [63]) have the potential to further improve STRspy allele predictions. Future studies will aim to determine the optimal amount of final prepared library to load and the maximum number of samples that can be multiplexed on R10.4.1 MinION flow cells. STRspy allele predictions will be used to determine which pore series can achieve accurate genotypes with the lowest coverage. This information could enable us to maximum flow cell usage by reducing PCR cycle number and implementing adaptive sampling

with ReadUntil [56]. The ability to multiplex more forensic markers and samples would in turn decrease the cost of STR nanopore sequencing.

MORE INFORMATION

Both the cycle number (Chapter 2) and multiplex (Chapter 3) studies were performed using high quality DNA controls. A primary objective of ongoing research is to demonstrate that accurate STR profiles can also be achieved from DNA evidence collected in routine forensic casework such as buccal swab, blood, and bone. Future efforts will also be geared towards expanding the current capabilities of our novel bioinformatic method for the simultaneous detection of SNPs and variation within mtDNA. In addition to traditional PCR amplification, we will assess the use of probe-based capture methods to minimize stutter and improve profiling of severely degraded samples [38,64]. These studies could enable us to achieve the most comprehensive representation of forensic genetic variation to date with the pocket-sized MinION device.

MORE STREAMLINED METHODS

VOLTRAX

The ligation-based kits used to prepare STR sequencing libraries in this project are relatively labor intensive and time consuming. Native amplicon barcoding and sequencing requires multiple reactions and magnetic bead-based cleanups amounting to an overall preparation time of more than 60min (depending on the number of samples). ONT also offers transposase-based library preparation protocols that can be completed in approximately 10min [32]. These rapid workflows are optimized for samples containing at least 400ng of high molecular weight DNA and thus are not suitable for the short STR amplicons used in forensic investigations [3]. ONT has therefore designed a fully automated solution for library preparation known as the VolTRAX.

After the samples of interest and appropriate reagents are loaded onto a cartridge, this small, USB-powered device transports liquids in one of the predefined paths selected via the control software [3,32]. The VolTRAX performs all required reactions and allows users to achieve consistent library preparations regardless of kit or skill level. Reducing potential for contamination and human errors would be particularly beneficial when processing the often fragile and limited DNA evidence collected in forensic investigations. The VolTRAX has not been assessed in this context due in part to the fact that researchers were unable to extract accurate STR profiles from nanopore sequencing reads in previous studies. STRspy makes it possible to compare allele designations obtained from manual and automated library preparation workflows. Our method could therefore be a critical component of future development and implementation of streamlined nanopore sequencing workflows in forensic DNA examinations.

USER INTERFACE

Several specialized methods have been developed to align and detect tandem repeats in nanopore sequencing reads. Despite the significance of these advancements in biomedical applications, researchers have demonstrated that available tandem repeat tools are unsuitable for forensic STRs [35]. Recent efforts have focused on developing forensic-specific pipelines that accommodate the structural diversity and length-based nomenclature of established STR panels [35,53]. These workflows harness different bioinformatic tools for processing and filtering reads. Extraction and interpretation of the desired information requires computational skills not common among most forensic DNA analysts. We therefore developed a streamlined method that converts third-generation sequencing data into the STR language used in forensic laboratories across the world. STRspy automatically analyzes user-provided read data and reports genotype predictions as the bracketed repeat region and length-based designation. Despite the relative ease of use, STRspy must be installed and executed at the command line. Forensic DNA analysts

are accustomed to the electropherogram peaks in conventional CE profiles. To address potential interpretation challenges and minimize training requirements, user interfaces have been developed and adopted for Illumina SBS data [65]. An interactive STRspy user interface would be beneficial for analysts that are not familiar with the command line. Further streamlining the reads-to-profile process would make nanopore sequencing more accessible and appealing to forensic laboratories.

CHAPTER 5

Discussion & conclusion

Forensic DNA examinations harness the high repeat length variation observed at STRs throughout the genome for human identification. Conventional typing approaches involve PCR amplification followed by length-based separation and fluorescent detection via CE. These well-established techniques are used in forensic laboratories across the nation to produce genetic information that can be uploaded and searched against FBI databases. Although CE profiles are both powerful and reliable, nucleotide-level variation within and around STRs is hidden in the length-based allele designations. Researchers have demonstrated that the additional resolution achieved with STR sequencing data can facilitate interpretation in challenging casework scenarios such as the deconvolution of mixed DNA profiles. This project assessed the potential to sequence forensic STRs on the handheld MinION device. As the newest and smallest DNA sequencer available, the MinION has undergone limited testing but offers unique features that could be beneficial in forensic applications.

We first aimed to determine whether STR amplicons generated using a commercial NGS kit can be sequenced on the MinION. Seven DNA controls were amplified in triplicate at 15 and 30 PCR cycles with the Promega PowerSeq 46GY System. High coverage was observed across samples and autosomal loci in the 30-cycle dataset compared to the 15-cycle. This observation is consistent with the exponential increase in STR-containing DNA fragments during the process of PCR. Despite differences in coverage, these data indicate that STR amplicon libraries can be barcoded and multiplexed for sequencing on the pocket-sized MinION.

Researchers were unable to extract accurate STR allele designations from error-prone ONT reads in previous studies. We therefore developed and assessed STRspy in the second aim of **Chapter 2**. Our novel bioinformatic method can reveal nucleotide-level variation in and around STRs while also reporting length-based profiles consistent with CE. The 15- and 30-cycle datasets from above were processed with STRspy and resultant allele designations were

compared to the manufacturer-validated genotypes. STRspy achieved 100% concordance across the 22 autosomal STRs profiled in the 30-cycle dataset based on both sequence and length. STRspy also detected variation in the flanking regions with a high level of accuracy.

A primary advantage of NGS in forensic DNA examinations is that more STR loci can be profiled in each run. In Chapter 2, we focused on autosomal loci because validated sequence-based information for STRs on the Y chromosome had not been published when the STR allele database was constructed. **Chapter 3** builds upon our novel approach to enable simultaneous sequencing and profiling of autosomal and Y-STRs targeted in common NGS amplification kits. We first reanalyzed the five male samples in the 15- and 30-cycle triplicates sequenced in Chapter 2. As with autosomal STRs, per locus read coverage was significantly lower across Y-STRs in the 15-cycle dataset compared to samples amplified with 30 PCR cycles. The five false positive predictions in the 15-cycle dataset were overcome with the higher coverage 30-cycle dataset, resulting in complete concordance with manufacturer-validated genotypes across all Y-STRs.

Three to four 30-cycle libraries were multiplexed in the previous experiments. The high coverage and accurate profiles obtained suggest that more samples can be sequenced at once. We therefore assessed the impact of multiplexing on MinION throughput and STRspy predictions in the second aim of **Chapter 3**. To determine the number of samples that can be sequenced and profiled on a single MinION flow cell, we prepared stock solutions of barcoded amplicon libraries from four DNA controls. Samples were pooled to 75ng in sets of 12, 18, and 24 barcodes per flow cell and sequenced on the MinION device for 72hrs. Per locus coverage for each sample decreased with increasing multiplex size. Nevertheless, STRspy predicted the correct allele designations across all STR loci. These results demonstrate that our methods can support

multiplexes of at least 24 samples to produce accurate profiles, which equates to less than 4.5ng per barcoded amplicon library. This in turn reduces the cost per sample of ONT sequencing.

The data presentation throughout Chapters 2 and 3 demonstrate that forensic autosomal and Y-STRs are amenable to sequencing on the MinION device when analyzed with STRspy. STRspy is the first and only method shown to predict the correct allele designations based on both sequence and length from error-prone ONT reads. We demonstrate the capabilities of our method by profiling the largest combination of autosomal and Y-STRs amplified by a single commercial kit. The resultant profiles achieve higher resolution by revealing nucleotide-level variation within and around STRs while also maintaining compatibility with current CODIS databases.

Despite the significance of the results presented herein, there are key limitations in both our laboratory and data analysis approaches. As with current CE and NGS techniques, this project relied on PCR to generate enough STR-containing DNA fragments for nanopore sequencing. Amplification increased depth of coverage and enabled the correct alleles to outcompete both noise and stutter in subsequent analyses. However, higher cycle number reduced the accuracy of SNP calls in the flanking regions. These observations suggest that it would be beneficial to determine the optimal number of cycles for simultaneous detection of both genetic marker systems (somewhere between 15 and 30 cycles). Further optimization of the PCR reaction could produce more balanced profiles with fewer stutter artifacts and reduced coverage for improved SNP calling compared to the 30-cycle dataset.

In addition to challenges associated with amplification, some inherent features of STRspy limit the scope of samples that it can profile. STRspy can only predict the correct profile if the true alleles are present in the STR database used for alignment. Although we are updating our database to include all validated sequence-based alleles for autosomal and Y loci published in

the STRSeq BioProject, it may still lack rare variants. STRspy is also unable to predict mixed DNA profiles at this time. The high rate of error through low complexity regions such as STRs makes it difficult to differentiate between minor contributors and sequencing artifacts. We designed STRspy to report the top Y and top two autosomal alleles at most, and thus it is not suitable for DNA mixtures. Continued improvements in nanopore sequencing could increase the feasibility of mixture profiling in which case we will determine how to implement the appropriate modifications into the STRspy framework.

Nanopore sequencing could make it possible to generate nucleotide-level data for STRs and other genetic markers used for human identification on a single platform that costs a mere $1k. The flexibility and affordability of the MinION have captured the attention of forensic researchers across the world. Here we developed a novel approach to a problem that has perplexed forensic application of nanopore sequencing since 2018. With STRspy, we are one step closers to harnessing the big potential of this small device in forensic DNA examinations.

REFERENCES

[1] M.A. Jobling, P. Gill, Correction: Encoded evidence: DNA in forensic analysis, Nat. Rev. Genet. 5 (2004) 739–751. https://doi.org/10.1038/nrg1455.

[2] J.M. Butler, The future of forensic DNA analysis, Philos. Trans. R. Soc. Lond. B Biol. Sci. 370 (2015) 20140252. https://doi.org/10.1098/rstb.2014.0252.

[3] C.L. Hall, R.R. Zascavage, F.J. Sedlazeck, J.V. Planz, Potential applications of nanopore sequencing for forensic analysis, Forensic Sci. Rev. 32 (2020) 23–54. https://www.ncbi.nlm.nih.gov/pubmed/32007927.

[4] J.M. Butler, Short tandem repeat typing technologies used in human identity testing, Biotechniques. 43 (2007) ii–v. https://doi.org/10.2144/000112582.

[5] J.M. Butler, Genetics and genomics of core short tandem repeat loci used in human identity testing, J. Forensic Sci. 51 (2006) 253–265. https://doi.org/10.1111/j.1556-4029.2006.00046.x.

[6] A. Edwards, A. Civitello, H.A. Hammond, C.T. Caskey, DNA typing and genetic mapping with trimeric and tetrameric tandem repeats, Am. J. Hum. Genet. 49 (1991) 746–756. https://www.ncbi.nlm.nih.gov/pubmed/1897522.

[7] H.A. Hammond, L. Jin, Y. Zhong, C.T. Caskey, R. Chakraborty, Evaluation of 13 short tandem repeat loci for use in personal identification applications, Am. J. Hum. Genet. 55 (1994) 175–189. https://www.ncbi.nlm.nih.gov/pubmed/7912887.

[8] J.M. Butler, DNA Databases, in: Advanced Topics in Forensic DNA Typing, Elsevier, 2012: pp. 213–270. https://doi.org/10.1016/b978-0-12-374513-2.00008-7.

[9] D.R. Hares, Expanding the CODIS core loci in the United States, Forensic Sci. Int. Genet. 6 (2012) e52-4. https://doi.org/10.1016/j.fsigen.2011.04.012.

[10] D.R. Hares, Selection and implementation of expanded CODIS core loci in the United States, Forensic Sci. Int. Genet. 17 (2015) 33–34. https://doi.org/10.1016/j.fsigen.2015.03.006.

[11] H. Fan, J.-Y. Chu, A brief review of short tandem repeat mutation, Genomics Proteomics Bioinformatics. 5 (2007) 7–14. https://doi.org/10.1016/S1672-0229(07)60009-6.

[12] Slipped-strand mispairing: a major mechanism for DNA sequence evolution, Mol. Biol. Evol. (1987). https://doi.org/10.1093/oxfordjournals.molbev.a040442.

[13] D. Pumpernik, B. Oblak, B. Borstnik, Replication slippage versus point mutation rates in short tandem repeats of the human genome, Mol. Genet. Genomics. 279 (2008) 53–61. https://doi.org/10.1007/s00438-007-0294-1.

[14] J.M. Butler, Recent developments in Y-short tandem repeat and Y-single nucleotide polymorphism analysis, Forensic Sci. Rev. 15 (2003) 91–111. https://www.ncbi.nlm.nih.gov/pubmed/26256727.

[15] M.A. Jobling, A. Pandya, C. Tyler-Smith, The Y chromosome in forensic analysis and paternity testing, Int. J. Legal Med. 110 (1997) 118–124. https://doi.org/10.1007/s004140050050.

[16] N.M.M. Novroski, F.R. Wendt, A.E. Woerner, M.M. Bus, M. Coble, B. Budowle, Expanding beyond the current core STR loci: An exploration of 73 STR markers with increased

diversity for enhanced DNA mixture deconvolution, Forensic Sci. Int. Genet. 38 (2019) 121–129. https://doi.org/10.1016/j.fsigen.2018.10.013.
[17] N.M.M. Novroski, A.E. Woerner, B. Budowle, Potential highly polymorphic short tandem repeat markers for enhanced forensic identity testing, Forensic Sci. Int. Genet. 37 (2018) 162–171. https://doi.org/10.1016/j.fsigen.2018.08.011.
[18] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, R.A. Guerrieri, P.M. Vallone, Sequence variation of 22 autosomal STR loci detected by next-generation sequencing, Forensic Sci. Int. Genet. 21 (2016) 15–21. https://doi.org/10.1016/j.fsigen.2015.11.005.
[19] M.C. Kline, C.R. Hill, A.E. Decker, J.M. Butler, STR sequence analysis for characterizing normal, variant, and null alleles, Forensic Sci. Int. Genet. 5 (2011) 329–332. https://doi.org/10.1016/j.fsigen.2010.09.005.
[20] C. Allor, D.D. Einum, M. Scarpetta, Identification and characterization of variant alleles at CODIS STR loci, J. Forensic Sci. 50 (2005) 1128–1133. https://doi.org/10.1520/jfs2005024.
[21] E.-M. Dauber, A. Kratzer, F. Neuhuber, W. Parson, M. Klintschar, W. Bär, W.R. Mayr, Germline mutations of STR-alleles include multi-step mutations as defined by sequencing of repeat and flanking regions, Forensic Sci. Int. Genet. 6 (2012) 381–386. https://doi.org/10.1016/j.fsigen.2011.07.015.
[22] R.L.M. Huel, L. Basić, K. Madacki-Todorović, L. Smajlović, I. Eminović, I. Berbić, A. Milos, T.J. Parsons, Variant alleles, triallelic patterns, and point mutations observed in nuclear short tandem repeat typing of populations in Bosnia and Serbia, Croat. Med. J. 48 (2007) 494–502. https://www.ncbi.nlm.nih.gov/pubmed/17696304.
[23] R.A. Griffiths, M.D. Barber, P.E. Johnson, S.M. Gillbard, M.D. Haywood, C.D. Smith, J. Arnold, T. Burke, A.J. Urquhart, P. Gill, New reference allelic ladders to improve allelic designation in a multiplex STR system, Int. J. Legal Med. 111 (1998) 267–272. https://doi.org/10.1007/s004140050167.
[24] A.M. Lins, K.A. Micka, C.J. Sprecher, J.A. Taylor, J.W. Bacher, D.R. Rabbach, R.A. Bever, S.D. Creacy, J.W. Schumm, Development and population study of an eight-locus short tandem repeat (STR) multiplex system, J. Forensic Sci. 43 (1998) 1168–1180. https://doi.org/10.1520/jfs14381j.
[25] C. Phillips, L. Fernandez-Formoso, M. Garcia-Magariños, L. Porras, T. Tvedebrink, J. Amigo, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, A. Freire-Aradas, A. Gomez-Carballa, A. Mosquera-Miguel, A. Carracedo, M.V. Lareu, Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel, Forensic Sci. Int. Genet. 5 (2011) 155–169. https://doi.org/10.1016/j.fsigen.2010.02.003.
[26] J.V. Planz, T.A. Hall, Hidden variation in microsatellite loci: Utility and implications for forensic DNA analysis, Forensic Sci. Rev. 24 (2012) 27–42. https://www.ncbi.nlm.nih.gov/pubmed/26231359.

[27] S. Goodwin, J.D. McPherson, W.R. McCombie, Coming of age: ten years of next-generation sequencing technologies, Nat. Rev. Genet. 17 (2016) 333–351. https://doi.org/10.1038/nrg.2016.49.
[28] K.B. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based U.S. population data for 27 autosomal STR loci, Forensic Sci. Int. Genet. 37 (2018) 106–115. https://doi.org/10.1016/j.fsigen.2018.07.013.
[29] How nanopore sequencing works, Oxford Nanopore Technologies. (n.d.). https://nanoporetech.com/how-it-works (accessed October 6, 2022).
[30] Flow cells and nanopore, Oxford Nanopore Technologies. (n.d.). https://nanoporetech.com/how-it-works/flow-cells-and-nanopores (accessed October 6, 2022).
[31] F.J. Rang, W.P. Kloosterman, J. de Ridder, From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy, Genome Biol. 19 (2018). https://doi.org/10.1186/s13059-018-1462-9.
[32] Nanopore product brochure, Oxford Nanopore Technologies. (2022). https://nanoporetech.com/sites/default/files/s3/literature/product-brochure.pdf (accessed October 5, 2022).
[33] M. Jain, I.T. Fiddes, K.H. Miga, H.E. Olsen, B. Paten, M. Akeson, Improved data analysis for the MinION nanopore sequencer, Nat. Methods. 12 (2015) 351–356. https://doi.org/10.1038/nmeth.3290.
[34] S. Cornelis, S. Willems, C. Van Neste, O. Tytgat, J. Weymaere, A.-S.V. Plaetsen, D. Deforce, F. Van Nieuwerburgh, Forensic STR profiling using Oxford Nanopore Technologies' MinION sequencer, BioRxiv. (2018) 433151. https://doi.org/10.1101/433151.
[35] 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, L. Kong, X.-C. Bo, S.-Q. Wang, M. Ni, J.-W. Yan, Forensic nanopore sequencing of STRs and SNPs using Verogen's ForenSeq DNA Signature Prep Kit and MinION, Int. J. Legal Med. (2021). https://doi.org/10.1007/s00414-021-02604-0.
[36] 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 (2020). https://doi.org/10.3390/genes11040381.
[37] 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. 65 (2020) 21–24. https://doi.org/10.1038/s10038-019-0688-z.
[38] R.R. Zascavage, C.L. Hall, K. Thorson, M. Mahmoud, F.J. Sedlazeck, J.V. Planz, Approaches to whole mitochondrial genome sequencing on the Oxford Nanopore MinION, Curr. Protoc. Hum. Genet. 104 (2019) e94. https://doi.org/10.1002/cphg.94.
[39] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmão, D.R. Hares, J.A. Irwin, J.L. King, P. de Knijff, N. Morling, M. Prinz, P.M. Schneider, C. Van Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal

nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63. https://doi.org/10.1016/j.fsigen.2016.01.009.
[40] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics. 34 (2018) 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
[41] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, Bioinformatics. 25 (2009) 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
[42] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics. 26 (2010) 841–842. https://doi.org/10.1093/bioinformatics/btq033.
[43] J. Farek, D. Hughes, A. Mansfield, O. Krasheninina, W. Nasser, F.J. Sedlazeck, Z. Khan, E. Venner, G. Metcalf, E. Boerwinkle, D.M. Muzny, R.A. Gibbs, W. Salerno, xAtlas: Scalable small variant calling across heterogeneous next-generation sequencing experiments, BioRxiv. (2018) 295071. https://doi.org/10.1101/295071.
[44] Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M.J. Ziller, V. Amin, J.W. Whitaker, M.D. Schultz, L.D. Ward, A. Sarkar, G. Quon, R.S. Sandstrom, M.L. Eaton, Y.-C. Wu, A.R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R.A. Harris, N. Shoresh, C.B. Epstein, E. Gjoneska, D. Leung, W. Xie, R.D. Hawkins, R. Lister, C. Hong, P. Gascard, A.J. Mungall, R. Moore, E. Chuah, A. Tam, T.K. Canfield, R.S. Hansen, R. Kaul, P.J. Sabo, M.S. Bansal, A. Carles, J.R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T.R. Mercer, S.J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R.C. Sallari, K.T. Siebenthall, N.A. Sinnott-Armstrong, M. Stevens, R.E. Thurman, J. Wu, B. Zhang, X. Zhou, A.E. Beaudet, L.A. Boyer, P.L. De Jager, P.J. Farnham, S.J. Fisher, D. Haussler, S.J.M. Jones, W. Li, M.A. Marra, M.T. McManus, S. Sunyaev, J.A. Thomson, T.D. Tlsty, L.-H. Tsai, W. Wang, R.A. Waterland, M.Q. Zhang, L.H. Chadwick, B.E. Bernstein, J.F. Costello, J.R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J.A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes, Nature. 518 (2015) 317–330. https://doi.org/10.1038/nature14248.
[45] C.R. Hill, D.L. Duewer, M.C. Kline, C.J. Sprecher, R.S. McLaren, D.R. Rabbach, B.E. Krenke, M.G. Ensenberger, P.M. Fulmer, D.R. Storts, J.M. Butler, Concordance and population studies along with stutter and peak height ratio analysis for the PowerPlex ® ESX 17 and ESI 17 Systems, Forensic Sci. Int. Genet. 5 (2011) 269–275. https://doi.org/10.1016/j.fsigen.2010.03.014.
[46] P. Müller, A. Alonso, P.A. Barrio, B. Berger, M. Bodner, P. Martin, W. Parson, Systematic evaluation of the early access applied biosystems precision ID Globalfiler mixture ID and Globalfiler NGS STR panels for the ion S5 system, Forensic Sci. Int. Genet. 36 (2018) 95–103. https://doi.org/10.1016/j.fsigen.2018.06.016.
[47] M.G. Ensenberger, J. Thompson, B. Hill, K. Homick, V. Kearney, K.A. Mayntz-Press, P. Mazur, A. McGuckian, J. Myers, K. Raley, S.G. Raley, R. Rothove, J. Wilson, D. Wieczorek, P.M.

Fulmer, D.R. Storts, B.E. Krenke, Developmental validation of the PowerPlex 16 HS System: an improved 16-locus fluorescent STR multiplex, Forensic Sci. Int. Genet. 4 (2010) 257–264. https://doi.org/10.1016/j.fsigen.2009.10.007.
[48] P. Hölzl-Müller, M. Bodner, B. Berger, W. Parson, Exploring STR sequencing for forensic DNA intelligence databasing using the Austrian National DNA Database as an example, Int. J. Legal Med. (2021). https://doi.org/10.1007/s00414-021-02685-x.
[49] W. De Coster, M.H. Weissensteiner, F.J. Sedlazeck, Towards population-scale long-read sequencing, Nat. Rev. Genet. (2021). https://doi.org/10.1038/s41576-021-00367-3.
[50] F.J. Sedlazeck, H. Lee, C.A. Darby, M.C. Schatz, Piercing the dark matter: bioinformatics of long-range sequencing and mapping, Nat. Rev. Genet. 19 (2018) 329–346. https://doi.org/10.1038/s41576-018-0003-4.
[51] C.L. Hall, R.K. Kesharwani, N.R. Phillips, J.V. Planz, F.J. Sedlazeck, R.R. Zascavage, Accurate profiling of forensic autosomal STRs using the Oxford Nanopore Technologies MinION device, Forensic Sci. Int. Genet. 56 (2022) 102629. https://doi.org/10.1016/j.fsigen.2021.102629.
[52] L. Gusmão, J.M. Butler, A. Carracedo, P. Gill, M. Kayser, W.R. Mayr, N. Morling, M. Prinz, L. Roewer, C. Tyler-Smith, P.M. Schneider, DNA Commission of the International Society of Forensic Genetics, DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis, Forensic Sci. Int. 157 (2006) 187–197. https://doi.org/10.1016/j.forsciint.2005.04.002.
[53] O. Tytgat, S. Škevin, D. Deforce, F. Van Nieuwerburgh, Nanopore sequencing of a forensic combined STR and SNP multiplex, Forensic Sci. Int. Genet. 56 (2022) 102621. https://doi.org/10.1016/j.fsigen.2021.102621.
[54] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic Acids Res. 24 (1996) 2807–2812. https://doi.org/10.1093/nar/24.14.2807.
[55] S. Cornelis, Y. Gansemans, L. Deleye, D. Deforce, F. Van Nieuwerburgh, Forensic SNP genotyping using nanopore MinION sequencing, Sci. Rep. 7 (2017) 41759. https://doi.org/10.1038/srep41759.
[56] S. Kovaka, Y. Fan, B. Ni, W. Timp, M.C. Schatz, Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED, Nat. Biotechnol. 39 (2021) 431–441. https://doi.org/10.1038/s41587-020-0731-9.
[57] 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. https://doi.org/10.1016/j.fsigen.2017.08.017.
[58] F.R. Wendt, J.L. King, N.M.M. Novroski, J.D. Churchill, J. Ng, R.F. Oldt, K.L. McCulloh, J.A. Weise, D.G. Smith, S. Kanthaswamy, B. Budowle, Flanking region variation of ForenSeq™ DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans, Forensic Sci. Int. Genet. 28 (2017) 146–154. https://doi.org/10.1016/j.fsigen.2017.02.014.

[59] C.R. Steffen, T.I. Huszar, L.A. Borsuk, P.M. Vallone, K.B. Gettings, A multi-dimensional evaluation of the “NIST 1032” sample set across four forensic Y-STR multiplexes, Forensic Sci. Int. Genet. 57 (2022) 102655. https://doi.org/10.1016/j.fsigen.2021.102655.
[60] C. Delahaye, J. Nicolas, Sequencing DNA with nanopores: Troubles and biases, PLoS One. 16 (2021) e0257521. https://doi.org/10.1371/journal.pone.0257521.
[61] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K.F. Au, Nanopore sequencing technology, bioinformatics and applications, Nat. Biotechnol. 39 (2021) 1348–1365. https://doi.org/10.1038/s41587-021-01108-x.
[62] M. Sereika, R.H. Kirkegaard, S.M. Karst, T.Y. Michaelsen, E.A. Sørensen, R.D. Wollenberg, M. Albertsen, Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing, Nat. Methods. 19 (2022) 823–826. https://doi.org/10.1038/s41592-022-01539-7.
[63] Oxford Nanopore Technologies, Nanopore Sequencing Accuracy, Nanoporetech. (n.d.). https://nanoporetech.com/accuracy.
[64] A. Tillmar, K. Sturk-Andreaggi, J. Daniels-Higginbotham, J.T. Thomas, C. Marshall, The FORCE panel: An all-in-one SNP marker set for confirming investigative genetic genealogy leads and for general forensic applications, Genes (Basel). 12 (2021) 1968. https://doi.org/10.3390/genes12121968.
[65] J.L. King, A.E. Woerner, S.N. Mandape, K.B. Kapema, R.S. Moura-Neto, R. Silva, B. Budowle, STRait Razor Online: An enhanced user interface to facilitate interpretation of MPS data, Forensic Sci. Int. Genet. 52 (2021) 102463. https://doi.org/10.1016/j.fsigen.2021.102463.