Ludwick, Whitney N. SNP Genotyping of Native DNA using Oxford Nanopore MinION Sequencing. Master of Science (Forensic Genetics), May, 2018, 61 pp., 8 tables, 12 figures, bibliography, 35 titles.

Short tandem repeats (STRs) are the primary system of genetic variation used for human identity testing in forensics; however, STR typing relies on the use of time-consuming polymerase chain reaction and expensive laboratory equipment. The use of single nucleotide polymorphisms (SNPs) in forensics have several advantages over STRs. In this study, a panel of Identity SNPs were interrogated and typed from native genomic DNA sequencing libraries using the Oxford Nanopore Technologies (ONT) MinION sequencer. We determined that SNPs could be effectively captured using existing software. Four different methods of alignment were investigated, and we found that aligning sequence data to the human genomic sequence (hg19) provided partial profiles, while aligning data to a merged reference profile resulted in more complete profiles. As ONT's platform continues to improve, SNP genotyping using the MinION may be used to generate complete SNP profiles with the sufficient depth of coverage for reliable genotype determination.

SNP GENOTYPING OF NATIVE DNA USING OXFORD NANOPORE MINION SEQUENCING

INTERNSHIP PRACTICUM REPORT

Presented to the Graduate Council of the

Graduate School of Biomedical Sciences

University of North Texas

Health Science Center at Fort Worth

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Whitney N Ludwick B.S. B.A.

Fort Worth, Texas

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Short tandem repeats (STRs) have been the primary means for human identity testing in the forensic community over the past 28 years [1]. However, with advances in technology, the utilization of single nucleotide polymorphisms (SNPs) in forensics and paternity testing proves to have several advantages [2-3]. Compared to the currently used combined DNA index system (CODIS) STR loci, SNPs have genotyping methods that are less costly and time-consuming, have lower error rates, and can be accomplished with shorter DNA fragments, similar to those found in forensic evidentiary items [4]. SNPs also have much lower mutation rates than STRs, with the mutation rate being an estimated 10^{-8} for SNPs versus 10^{-3} for STRs [5]. Mutation rates are important in paternity testing. For these reasons, SNPs are thought to be a candidate for new forensic markers [6-7].

The common genotyping method for STRs relies on lengthy PCR-based steps and the use of expensive and large instruments for capillary electrophoresis (CE) [8]. However, there are several SNP genotyping methods, with the most frequently used one being SNP chips. A SNP chip, or SNP array, detects polymorphisms in the human genome [9]. A single SNP chip is composed of multiple arrays that contain beads covered with fragments of single stranded DNA probes that are complementary to the sequence adjacent to a particular SNP. The beads are incubated with genomic DNA strands; the strands of DNA will anneal to beads that have a complementary sequence. DNA polymerase then incorporates a fluorescently labeled nucleotide

to the 3' end, corresponding to the variable site, or SNP [10-12]. Recently, the use of SNP chips has plateaued due to an increased interest in other technologies, such as Next Generation Sequencing [13].

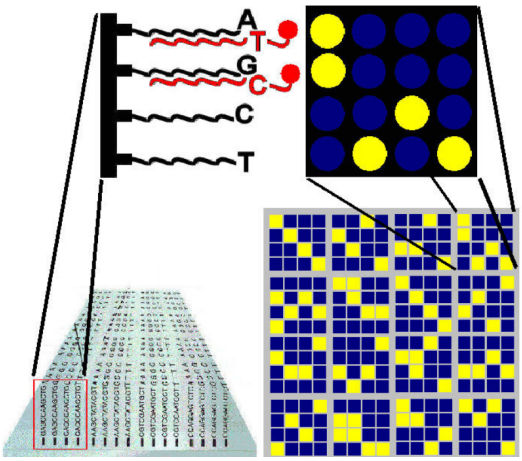


FIGURE 1. SNP CHIP. A SNP chip detects SNPs in the human genome. It is composed of multiple arrays that contain beads covered with fragments of single stranded DNA probes that are complementary to the sequence adjacent to a particular SNP. The beads are incubated with genomic DNA strands; the strands of DNA will anneal to beads that have a complementary sequence. DNA polymerase then incorporates a fluorescently labeled nucleotide to the 3' end, corresponding to the variable site, or SNP (*Figure 1 adapted from* [12]).

Another method used for SNP genotyping is the SNaPshot® assay from Applied Biosystems™ [14]. The SNaPshot® assay uses a single base extension (SBE) technique that is based on the ability of an oligonucleotide primer to bind to a complementary template strand; DNA polymerase then extends the primer by adding a single fluorescently labeled dideoxynucleotide (ddNTP) to the 3' end [15]. Limitations include the need for multiple transfer steps, that increase the likelihood for contamination, the reduced ability to multiplex more than 30 to 40 SNPs at a time, and the need for costly equipment involved in PCR and CE [16-17]. Alternatively, massively parallel sequencing (MPS), also termed next-generation sequencing

(NGS), is a newer genotyping method that can sequence multiple samples at the same time using small amounts of DNA. An example of this current method is MiSeq™ by Illumina®, that can produce as much as 15 GB of sequencing data on a single flow cell and can be applied to interrogating forensic markers [18]. However, Illumina® sequencing also requires PCR and library preparation steps that can take several days to complete [19].

An even newer method, introduced in 2014, is the MinION sequencer from from Oxford Nanopore Technologies (ONT). The MinION device is not only inexpensive and portable, weighing less than 100 grams, but also allows for real-time sequencing. The MinION can eliminate the need for PCR and other time-consuming laboratory preparation in many applications, as well as reduces the cost associated with DNA analysis [20]. The aim of this study was to determine if a panel of forensically related SNPs could be effectively interrogated and typed from native genomic DNA sequencing libraries, saving time and money by eliminating PCR steps.

BACKGROUND

Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when alleles differ by a single nitrogenous base at specific loci [21-22]. They are not only the most common type of DNA variation, but also the most widespread across the entire human genome, appearing approximately one in every five hundred bases [23-25]. SNPs can occur in both noncoding and coding regions; some SNPs can play a role in affecting gene expression, thus making them good biological markers for studying diseases [26-28]. Projects such as the International HapMap Project and the 1000 Genomes Project founded in 2002 and 2008, respectively, created detailed catalogues of genetic variation across the human genome that has since been used to study genetic associations with disease [29-32].

The type and location of a SNP often is what determines the level of phenotypic effect it yields [33]. Many SNPs can be associated with phenotypic traits, whereas only a few STRs are able to give this type of information. SNPs are also suitable markers for ancestral studies and investigating lineage-familial relationships, providing information that STRs are not capable of [34-36]. Unlike SNPs used for the identification of individuals, ideal ancestry informative single nucleotide polymorphisms (AISNPs) will have large allele frequency differences across a set of populations, allowing for the distinction between populations. Much consideration goes into the s into the selection of AISNPs; a small enough number of SNPs should be chosen for cost and time efficiency, and the SNPs must also be highly selective for individual populations in order to accurately determine ancestry [37-38].

In the field of forensics, advances ensuing from the Human Genome and International HapMap Projects has caused an increased interest in investigating the potential use of SNPs as forensic markers. SNP genotyping in the forensic community was first done with HLA-DQA1 and AmpliType® PM kits, that prove to be largely uninformative currently due to the lack of loci examined [39]. Since this first attempt, related research has focused mainly on the number of SNPs required to obtain similar powers of discrimination provided by STRs and the selection of a panel of forensically relevant SNPs [40]. In 2006, Kidd K *et al*. focused on a five-step screening process for SNPs that displayed high heterozygosity and had similar allele frequencies across all populations, that would allow match probabilities to remain constant regardless of the population being used, thus making a good global panel of markers for identity testing. Identification of probable candidates originated from a database of SNPs from Applied Biosystems™; this database was used for the initial selection of markers solely because TaqMan® assays already existed, eliminating the need for new assays to be tested. Once allele frequencies for these SNPs were obtained, SNPs were ranked based on average heterozygosity and minimal allele frequency variation among four major populations. SNPs that had an average heterozygosity of greater than 0.45 and an Fst value less than 0.01 were chosen to be further evaluated against 7 populations. A second screening on an additional 33 populations was completed with SNPs having a maximum Fst value of 0.06. Hardy-Weinberg equilibrium was tested using a chi-squared test and linkage disequilibrium values were assessed before match probabilities for each marker were calculated. As a result, 19 SNPs were selected as final candidates for a global forensic panel with average match probabilities being between 10^{-7} and 10^{-8} and a probability of exclusion being greater than 0.999. Kidd K *et al*. concluded that extending this panel to 45 SNPs or more would result in a match probability value of 10^{-15},

similar to the match probability of the original 13 core STR loci produce. Several other forensic panels of SNPs have been proposed however, these labs did not have the goal of developing a universal panel of SNPs that utilized one allele frequency database like Kidd K *et al*. did [41]. Since this study, Kidd K *et al*. has developed panels which have increased to 45 SNPs with match probabilities averaging 10^{-15} and 92 SNPs with match probabilities as small as 10-*35* [42].

Although Kidd K *et al*. used TaqMan® assays for simplicity, other research groups chose to work towards developing highly multiplexed assays. In 2003, the SNPforID consortium was established; this project consisted of several groups with a similar interest in forensic DNA typing, whose main goal was to develop SNP assays for forensic DNA analysis [43]. Several studies were published from this project, many of which had found methods to detect 20 to 30 SNPs at a time [44-53]. One study focused on a PCR based assay that could detect up to 52 SNPs simultaneously. This assay utilized single-base extension (SBE) methodology and the SNaPshot® reaction mix to identify the chosen markers, and only required 0.5ng of DNA prior to PCR amplification. This particular study resulted in match probabilities spanning from 10^{-19} to 10^{-21} [54]. As technology in the field has progressed, studies have continued to focus on new ways to effectively use SNPs for genotyping, with the recent applications utilizing nanopore sequencing technologies.

NANOPORE SEQUENCING

Nanopore sequencing originated in 1989 and does not require the use of PCR amplification or any type of chemical labeling, such as fluorescent tags, of the sample of interest [55-56]. Oxford Nanopore Technologies utilizes a nanopore that is comprised of a protein whose core is a hollow tube measuring a few nanometers in diameter. The nanopore is embedded in a

synthetic polymer membrane with high electrical resistance. The membrane is bathed in an electrophysiological solution so, when a potential is applied across the membrane an ionic current is generated through the nanopore. Molecules such as DNA, RNA, or proteins that pass through the nanopore or close to its surface will generate characteristic disruptions in the current, referred to as the nanopore signal. The nanopore signal is then measured to identify the molecule [57]. Intact DNA strands are sequenced by mixing the DNA strands of interest with a processive enzyme; the DNA-enzyme complex moves towards the nanopore and the single stranded DNA (ssDNA) is pulled through the aperture one base at a time. Here, the nanopore signal can be used to determine the order of bases on the DNA strand [58]. There are multiple research groups that have used ONT's nanopore technology for SNP genotyping [59-60].

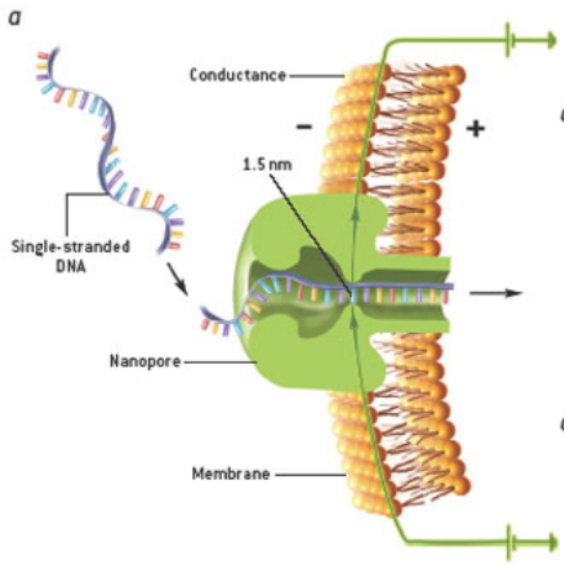


FIGURE 2. NANOPORE STRUCTURE. A nanopore is an enzyme-protein complex embedded in a membrane with high electrical resistance. A voltage is applied to the membrane and the negatively charged DNA is pulled towards the positive charge. The single stranded DNA is pulled through the opening of the pore one base at a time. Each base will produce a characteristic disruption in the current, that can be measured to identify the molecule. (*Figure 2 adapted from* [61]).

Recently, Zaaijer S *et al*. developed an approach to query SNPs using ONT's MinION sequencer, termed MinION sketching. Their MinION sketching strategy does not call for enrichment through PCR but integrates real-time strand sequencing data using a Bayesian search algorithm. The algorithm created, calculates a posterior probability that the sketch either matches or does not match to an item in a reference database, with regards to each SNP's allele frequency. As samples were sequenced by the nanopore, the algorithm compared SNPs from the sequencing data to SNPs from samples in a database in order to find a match. Zaaijer S *et al*. created a database that consisted of 31,000 individuals and results of the study showed that any sample contained in the database could be detected in a sequencing run in as little as 5 minutes of sketching and using approximately 98 to one 134 SNPs. The authors proposed that this strategy could eventually be used at crime scenes for on-site DNA analysis since the protocol requires little preparation and the instrumentation is portable [59].

Another study whose focal point is SNP genotyping using nanopore technology aimed to determine the applicability of using the MinION for forensic purposes. Cornelis S *et al*. utilized both Illumina® and MinION sequencing with the SNPforID consortium's 52 SNP-plex assay. The profiles generated from the different sequencing methods were then compared to test the efficiency of the MinION. Cornelis S *et al*. reported that 51 of the 52 SNPs interrogated were correctly genotyped using the Oxford Nanopore sequencer. It is thought that as improvements continue to be made to MinION technology, that it could become suitable for use in forensic laboratories [60].

PREVIOUS RESEARCH

Prior research performed in the laboratory focused on developing a method for mitochondrial DNA (mtDNA) genome sequencing that does not require the use of PCR using ONT's MinION [62]. MtDNA analysis is paramount in forensics when STR profiles are unable to be obtained due to low sample quality or quantity of nuclear DNA (nuDNA) [63]. Thorson developed a pipeline to evaluate data produced by the MinION. A pipeline is a data processing element that is essentially a string of commands, where the output of one command is the input for the following one [64]. Bioinformatics processing approaches of whole genome mtDNA were evaluated by comparing results obtained from runs using 2-directional (2D) library preparation of amplified mtDNA products to 1-directional (1D) rapid library preparation from genomic DNA. 2D is the term used when information from both strands of DNA are being utilized due to the use of a hairpin adapter, while 1D is the term used when forward or reverse strands will pass arbitrarily through the nanopore [65]. Native DNA is DNA that has not undergone amplification from PCR. As previously mentioned, current technology used for genotyping relies on PCR, which is time consuming and costly. If identity profiles could be effectively generated via a method that eliminated PCR, time and costs would be saved; this would be a huge benefit to the field of forensics. From the pilot study, Thorson concluded that comparable results can be achieved with the PCR enriched libraries and the native libraries when using the MinION [62].

INSTRUMENTS

ONT's MinION device is a pocket-size DNA sequencer that allows for real time analysis of data being generated. Each flow cell can produce up to 20 GB of data and is recyclable after use. Read lengths are customizable, with maximum read lengths being hundreds of kilobases long. Library preparation is also relatively simple, with an estimated prep time of ten minutes for the most recently released kit. Since the MinION is transportable, weighing less than one 100 grams, it can be used in the field or on site. This technology is also much cheaper than any other sequencing technology commercially available [66]. Since 2014 when the MinION was released for commercial use, the accuracy of the platform has improved; current error rates for the MinION are approximately one to three percent. Additionally, errors are in non-repetitive regions of the genome are extremely rare. All of the MinION sequencer's benefits make it a favorable tool when compared to other MPS approaches currently available.

SPECIFIC AIMS

This study was an internship project with the goal of using existing genomic data generated with the use of the ONT Minion to determine the efficacy of approach to capture 40 forensically relevant SNPs. The first objective was to investigate and establish a strategy to interrogate a panel of SNPs from the data by utilizing conventional SNP software approaches. The second objective was to determine if heterozygosity and homozygosity could be determined for specific alleles. To conclude if reliable heterozygote determination could be reached, it had to be determined that sufficient depth of coverage of the autosomal markers of interest would be

detected against the genome background. The third objective was to compare multiple methods of alignment to determine the best procedure for generating complete and accurate identity profiles from native DNA samples. One of the samples and the positive control had been previously sequenced using Taqman® assays and mass spectrometry for a previous study. The generated profiles from this study could be compared to the reference profiles for the previous study to determine which method of alignment produced the most accurate results.

CHAPTER 2

MATERIALS AND METHODS

In this project, genome sequence data was utilized that was generated from a previous study (Thorson, 2017). The following text provides a synopsis of the samples and methodology from the previous study from which we received the raw sequence run data.

SAMPLE PREPARATION

Genomic DNA (gDNA) was extracted from eight 200uL liquid blood samples (101, 102, 103, 433, 441, 442, 449, and 459) using the QIAmp® Mini Blood Kit (QIAGEN, Hilden, Germany) according to the manufacturer's specifications. This project was approved by the International Review Board (IRB) at the University of North Texas Health and Science Center (UNTHSC) (Protocol #2010-106: *Assembly of Databank for Development and Validation of Genomic Assays*). Extracted DNA from cell line HL-60 was purchased from ATCC (CCL-240D), which served as the National Institute of Standards and Technology (NIST) control to determine sequencing accuracy. To determine the quantity and quality of DNA from the eight blood samples, extracted DNA was quantified using the gDNA kit and Agilent 4200 Bioanalyzer (Agilent Technologies, Santa Clara, California, USA) according to the manufacturer's specifications.

LIBRARY PREPARATON AND SEQUENCING

Libraries for samples (101, 102, 103, and 433) were prepared using protocol SQK-RAD001 kit V9, and libraries for samples (425, 441, 442, 449, 459, and HL60) were prepared using protocol RAD002 kit V9, both according to the manufacturer's specifications. The only difference between the two protocols is that RAD002 kit V9 utilizes library loading beads (LLB), which brings the DNA closer to the pores prior to sequencing. The suggested target amount of gDNA is 200ng / 7.5 μL; for all samples the input amount varied (see Table 1). Here, library preparation consisted of DNA tagmentation and adapter attachment steps. gDNA and FRM buffer were mixed gently, spun down, and incubated at 30ºC for one minute, followed by another one minute incubation at 75ºC. RAD buffer and 0.4μL of Blunt/TA ligase were added to the tagmented DNA, gently mixed, spun down, and incubated at room temperature. Incubation time was extended to ten minutes however, it should be noted that sample 425 had an incubation time of five minutes. 12μL of library were added to a pre-sequencing mixture consisting of RBF and nuclease-free water; the pre-sequencing mixture for samples 425, 441, 442, 449, 459, and HL60 also contained LLB.

Prior to library loading and sequencing, flow cells were primed with 480μL of RBF and 520μL of nuclease-free water. Libraries were loaded drop-wise on the SpotON port of the flow-cell. Flow cells were then positioned in the MinION device and sequenced for 48 hours without base calling. It should be noted that library preparation and sequencing for samples 101, 102, 103, 433, 441, 442, 449, and 459, as well as the positive control HL-60 were done in a previous

study; library preparation and sequencing for sample 425 was completed during this study. A detailed description of the manufacturer's protocol can be found in Appendix A.

TABLE 1. DNA INPUT FOR LIBRARY PREPARATION. The manufacturer's recommendation for input DNA is 200ng. The table shows the amount of input DNA for the samples and the positive control. Samples 101, 102, 103, 433, 441, 442, 449, 459, and the positive control HL-60 were run in a previous study [61]. Sample 425 was run in this study.

Sample	Input DNA (ng)
HL-60	212.0
101	144.8
102	229.5
103	207.0
425	234.0
433	176.3
441	259.5
442	237.0
449	204.0
459	234.0

DATA ANALYSIS

Data Conversion

The customized data pipeline from the previous study was adapted to mine autosomal SNPs. Base calling was 1-directional and performed by one of three software programs: Albacore v0.8.4 [67], a C++ program, Metrichor v2.45.3 [68], an Amazon cloud-based system, and the ONT local base caller [69]. The MinION device outputs one FAST5 file per read. After being processed by one of the three aforementioned software programs, FAST5 files store metadata and events, such as aggregated bulk current measurements, that were pre-processed by the sequencer, as well as various log files and the raw signal dataset [70]. Conventional software used to interrogate SNPs requires FASTA or FASTQ files. FASTA files are in a text-based format and represent nucleotide or peptide sequences; base pairs or amino acids are represented

by a single letter code [71]. Multiple bioinformatics tools have been developed to convert the files from FAST5 files to FASTA files, like Nanopolish [72] and poretools [73]. Poretools, developed by Aaron Quinlan and Wick Loman, is written in python script and has the ability to filter sequences in a multitude of ways. Nanopolish, developed by Jared Simpson, is a C++ program with a python utility script. The Nanopolish software package analyzes the nanopore signal and calculates a consensus sequence for genome assembly. This software is also able to detect any base modifications, as well as calls SNPs with respect to the reference genome being used [74]. In this study, Nanopolish v0.6.0 was used.

The samples were first aligned to human genome 19 (hg19) [75]; BWA-MEM [76] was then used to generate a SAM file of the alignments. BWA-MEM is an alignment algorithm that can align sequencing reads or assembly contigs against large reference genomes, such as hg19. Using the algorithm, an alignment is seeded with supermaximal exact matches (SMEMs); the algorithm is essentially finding the largest exact match for positions of interest [77]. Samtools[78] converted the SAM files to BAM files, which is the file type needed for data analysis. Samtools is a set of utilities that can convert files into the BAM format. Samtools imports from and exports to the SAM (Sequence Alignment / Map) format. Samtools also performs sorting, merging, indexing, and retrieval of reads [79]. In this study, Samtools was also used to perform an index command and a stats command to generate BAI files and summary statistics, respectively. A BAI file is an index file for the bam file; it can be described as a table of contents for alignments for the corresponding file. The BAI file allows the bioinformatics program being used to sort through the BAM file and go to a specific region without having to read every base of the sequence The stats command collects statistics from the BAM files and outputs them in text format [80].

ALIGNMENT STRATEGIES

Four different types of alignments were performed, each to a different reference file, to gain an understanding of the most appropriate bioinformatics approach to take prior to interrogating the SNPs of interest. Alignments were done on the positive control, HL-60, and analyzed for concordance with NIST's standard HL-60 reference.

ALIGNMENT TO HUMAN GENOME 19

The 9 samples (101, 102, 103, 425, 433, 441, 442, 449, and 459) and the positive control (HL-60) were aligned to hg19. The data from the alignments to hg19 was manually viewed using Integrative Genomics Viewer (IGV) [81]. The 40 forensic relevant SNPs were looked for in the sequencing data. The panel of 40 SNPs interrogated in this study are those initially reported by [82] (Figure 2).

Chromosome	Cyto-genetic band position	†	Locus symbol ‡	ABI catalog #	dbSNP rs#	Nucleotide position UCSC May 2004	ALFRED site UID	F_{st} 40 p	F_{st} 7 p	Average heterozygosity 40 p	Average heterozygosity 7 p
11	q23.2		IGSF4	C___2450075_10	rs10488710	114,712,386	SI001899B	0.025	0.010	0.441	0.460
4	p12	√	GABRA2	C___8263011_10	rs279844	46,170,583	SI001391O	0.030	0.011	0.485	0.495
4	q32.3	√	PALLD	C__11245682_10	rs6811238	170,038,345	SI001910L	0.031	0.014	0.485	0.492
13	q32.3	√	PHGDHL1	C___1619935_1_	rs1058083	98,836,234	SI001402H	0.032	0.014	0.464	0.484
5	q31	√	SPOCK	C___2556113_10	rs13182883	136,661,237	SI001390N	0.033	0.019	0.471	0.489
1	q23.3	√	LY9	C___1006721_1_	rs560681	157,599,743	SI001392P	0.035	0.018	0.434	0.439
8	p21	√	FZD3	C___2049946_10	rs10092491	28,466,991	SI001900K	0.039	0.009	0.456	0.458
10	q26	√	HSPA12A	C___3254784_10	rs740598	118,496,889	SI001393Q	0.040	0.011	0.463	0.477
20	p12.1	√	C20orf133	C___2997607_10	rs445251	15,072,933	SI001912N	0.041	0.013	0.463	0.473
6	q22		TRDN	C___2140539_10	rs1358856	123,936,677	SI001427O	0.042	0.018	0.473	0.486
15	q13	√	Intergenic	C__11673733_10	rs1821380	37,100,694	SI001913O	0.042	0.018	0.464	0.474
20	q13.1	√	Intergenic	C___2508482_10	rs1523537	50,729,569	SI001914P	0.042	0.013	0.472	0.476
18	q11.1		ZNF521	C____105475_10	rs7229946	20,992,999	SI001901L	0.043	0.020	0.464	0.456
20	p11.1		SSTR4	C___3206279_1_	rs2567608	22,965,082	SI001902M	0.044	0.020	0.475	0.490
18	p11.3	√	RAB31	C___1371205_10	rs9951171	9,739,879	SI001395S	0.044	0.020	0.474	0.490
3	q29	√	ATP13A4	C__25749280_10	rs6444724	194,690,082	SI001903N	0.045	0.019	0.468	0.489
6	q16.1	√	Intergenic	C___1817429_10	rs1336071	94,593,976	SI001915Q	0.045	0.007	0.472	0.495
1	p36	√	PRDM2	C____342791_10	rs7520386	13,900,708	SI001394R	0.045	0.018	0.477	0.490
7	p22	√	Intergenic	C___2572254_10	rs1019029	13,667,516	SI001916R	0.045	0.018	0.472	0.485
22	q11.2		loc388882	C__11522503_1_	rs2073383	22,126,725	SI001911M	0.046	0.008	0.452	0.474
6	p24.1	√	HIVEP1	C___9371416_10	rs13218440	12,167,940	SI001397U	0.047	0.013	0.457	0.479
6	q22.31		Intergenic	C___1152009_10	rs1478829	120,602,393	SI001917S	0.047	0.008	0.474	0.491
6	q24.3		SASH1	C___1256256_1_	rs2272998	148,803,149	SI001398V	0.047	0.010	0.468	0.490
22	q12.3	√	loc650568	C__11887110_1_	rs987640	31,884,062	SI001918T	0.048	0.018	0.476	0.488
2	q31.3	√	CERKL	C___1276208_10	rs12997453	182,238,765	SI001396T	0.048	0.019	0.445	0.466
10	p15.1	√	DNMT2	C___2822618_10	rs3780962	17,233,352	SI001904O	0.049	0.020	0.475	0.490
6	q25	√	SYNE1	C___2515223_10	rs214955	152,789,820	SI001403I	0.049	0.017	0.475	0.491
4	q21.1		RCHY1	C___1880371_10	rs13134862	76,783,075	SI001400F	0.054	0.006	0.456	0.467
10	q24.3		SORBS1	C___7538108_10	rs1410059	97,162,585	SI001399W	0.054	0.012	0.471	0.482
16	p13.3	√	a2bp1	C__31419546_10	rs7205345	7,460,255	SI001905P	0.055	0.017	0.469	0.487
7	q33	√	PTN	C___3004178_10	rs321198	136,487,093	SI001906Q	0.056	0.004	0.457	0.489
5	qter	√	ADAMTS2	C___3153696_10	rs338882	178,623,331	SI001401G	0.056	0.019	0.467	0.490
4	q32.1		Intergenic	C___7428940_10	rs1554472	157,847,511	SI001919U	0.057	0.012	0.471	0.494
2	p25.2	√	GRHL1	C___2073009_10	rs1109037	10,036,320	SI001909T	0.058	0.018	0.467	0.482
6	q22.3		RSPO3	C____411273_10	rs2503107	127,505,069	SI001426N	0.058	0.013	0.454	0.463
6	q24		EPM2A	C___2223883_10	rs447818	145,910,689	SI001907R	0.058	0.015	0.471	0.479
5	q33.3		TTC1	C___1995608_10	rs7704770	159,420,531	SI001908S	0.058	0.016	0.450	0.456
5	q35		LCP2	C___3032822_1_	rs315791	169,668,498	SI001404J	0.058	0.018	0.471	0.485
11	q23	√	KBTBD3	C___1636106_10	rs6591147	105,418,194	SI001409O	0.059	0.019	0.449	0.481
18	q11.2		B4GALT6	C___7459903_10	rs985492	27,565,032	SI001413J	0.059	0.015	0.468	0.487
							Averages:	0.047	0.015	0.465	0.480

FIGURE 3. 40 FORENSICALLY RELEVANT SNPs. These 40 SNPs have been previously determined to be ideal forensic markers for identity testing, producing a match probability similar to that of the CODIS STRs. (*Figure 3 taken from* [82])

In order to look at the SNP positions and determine what bases were called, the SNP locations on chromosomes were first found by using UCSC's Table Browser [83]. The Table Browser controls were changed from the default settings in order to only generate SNP positions of interest. The Assembly was changed to "GRCh37/hg19", the Group was changed to "Variation", the track was changed to "Common SNPs(150)", the Table was changed to

"snp150Common", and the 40 SNPs were typed into the Identifiers Paste List. This automated process allowed for the quick retrieval of chromosome locations that correlated with the 40 SNPs from the panel. Details of the settings used for the Table Browser as well as an example of the output file used to gather locations of interest can be found in Appendix B. The BAM files generated from Samtools, as described previously, were input into IGV with a genome setting of human hg19 and evaluated. Total reads mapped, total bases mapped, and the number of SNPs mapped were recorded for later comparisons (see Appendix C for total reads mapped and total bases mapped). A read is a sequence of base pairs that correspond to all or part of a DNA fragment [84]. Total reads mapped refers to all of the reads that were generated from the alignment. Total bases mapped is the total number of base pairs that were mapped to the targeted region.

ALIGNMENT TO A CHROMOSOME

To compare various methods for optimal data alignments and analysis, additional alignments to different references were completed using the same FAST5 file to BAM file workflow as described previously. The first of these alternate alignments was one of the samples aligned to a specific chromosome rather than to the entire human genome. The chromosome used was chromosome 22 and the sample used was sample 449. Of the 40 forensic relevant SNPs being utilized in this study, only two of them reside on chromosome 22. Rs2073383 and rs987640 (the only two out of the 40 SNPs located on chromosome 22), were the two positions viewed via IGV for this alignment.

ALIGNMENT TO A MERGED REFERENCE FILE

An alignment to regions on the chromosomes was also performed by using a reference file (SNPs_100kb_merged_even) consisting of flanking data surrounding all 40 SNPs from the proposed panel. This reference file had to be created in a step-wise manner. Individual FASTA files for the 40 SNPs were generated using UCSC's Genome Browser [85]. Each SNP was searched and the option to choose one of multiple links was given. The dbSNP 150 link was chosen for its corresponding SNP. UCSC's Genome Browser has a default setting of 250bp flanking the SNP both upstream and downstream of the variant. An additional 50,000bp were added to either side of the variant upon retrieval. Detailed instructions on how the Genome Browser was used can be found in Appendix D. Each SNP and its flanking region was saved as FASTA files. FASTA files have headers, and since the files had to be merged in order to make one reference the headers had to be removed. This was done to all 40 files simultaneously by using a bash command from Samtools. Samples 103 and HL-60 were aligned to the new merged reference file using the same commands as the previous alignments. Commands used included an index command, which breaks apart the files into sections, a pipeline command, which is a series of commands; in this case the pipeline was used to output BAM files. Using Samtools, an index command and a stats command was performed to generate BAI files and summary statistics, respectively.

The BAM files from the alignments were input into IGV with a genome setting of SNPs_100kb_merged_even and evaluated. FASTA sequences of the 40 SNPs from NCBI's GenBank database [86] were found and used in order to determine the specific location of each

of the variants for this particular alignment. The resulting SNP genotypes for sample 103 and HL-60 were compared to reference genotypes to see how well using the MinION device for identity testing works.

ALIGNMENT TO INDIVIDUAL FASTA FILES

Additionally, alignments of the positive control HL-60 to select individual FASTA files, previously generated via UCSC's Genome Browser, were made in order to compare the results from the merged reference to smaller flanking regions around a single variant. Since the merged reference file (SNPs_100kb_merged_even) is comprised of the individual FASTA files, the results should be in concordance with one another.

CHAPTER 3

RESULTS AND DISCUSSION

The purpose of this study was to determine the applicability of ONT's MinION device to accurately and efficiently generate DNA profiles from native DNA samples. For the purpose of this study, a panel of SNPs described in [82] was used. This particular panel of 40 SNPs can give match probabilities averaging 10^{-15}, which is a comparable statistical value of what the original 13 core STR loci produce. In order to obtain a match probability similar to that of the currently used STRs, a panel of approximately 80 SNPs would have to be utilized [42, 87]. Four different types of alignments were performed to gain an understanding of the most appropriate bioinformatics approach to take prior to interrogating SNPs.

RESULTS FOR ALIGNMENT TO HUMAN GENOME 19

Alignments of the nine samples and the positive control to the whole genome resulted in varied partial profiles across all of the samples. SNP locations across hg19 and the partial profile for HL-60 can be seen in Table 2 and Table 3, respectively.

TABLE 2. SNP LOCATIONS ON CHROMOSOMES. Individual SNP positions on their corresponding chromosomes were determined using UCSC's Table Browser. The table below shows the chromosome position and the exact location on that chromosome where the SNP of interest is located.

SNPs	Chromosome	Position	SNPs	Chromosome	Position
rs10092491	chr8	28411072	rs2567608	chr20	23017082
rs1019029	chr7	13894276	rs279844	chr4	46329655
rs10488710	chr11	115207176	rs315791	chr5	169735920
rs1058083	chr13	100038233	rs321198	chr7	137029838
rs1109037	chr2	10085722	rs338882	chr5	178690725
rs12997453	chr2	182413259	rs3780962	chr10	17193346
rs13134862	chr4	76425896	rs445251	chr20	15124933
rs13182883	chr5	136633338	rs447818	chr6	145868996
rs13218440	chr6	12059954	rs560681	chr1	160786670
rs1336071	chr6	94537255	rs6444724	chr3	193207380
rs1358856	chr6	123894978	rs6591147	chr11	105912984
rs1410059	chr10	97172595	rs6811238	chr4	169663615
rs1478829	chr6	120560694	rs7205345	chr16	7520254
rs1523537	chr20	51296162	rs7229946	chr18	22739001
rs1554472	chr4	157489906	rs740598	chr10	118506899
rs1821380	chr15	39313402	rs7520386	chr1	14155402
rs2073383	chr22	23802171	rs7704770	chr5	159487953
rs214955	chr6	152697706	rs985492	chr18	29311034
rs2272998	chr6	148761456	rs987640	chr22	33559508
rs2503107	chr6	127463376	rs9951171	chr18	9749879

TABLE 3. SNP PROFILE FOR POSITIVE CONTROL HL-60. Alignments were done on the positive control, HL-60, and analyzed for concordance with NIST's standard HL-60 reference. A partial profile (19 called SNPs of the 40 SNPs of interest) was generated from the data produced by the alignment of HL-60 to hg19. Blue indicates the SNPs that were in concordance with

SNPs	Called Alleles	SNPs	Called Alleles
rs10092491	No Data	rs2567608	No Data
rs1019029	No Data	rs279844	A
rs10488710	C	rs315791	No Data
rs1058083	No Data	rs321198	No Data
rs1109037	No Data	rs338882	No Data
rs12997453	No Data	rs3780962	No Data
rs13134862	A	rs445251	No Data
rs13182883	G	rs447818	No Data
rs13218440	No Data	rs560681	A
rs1336071	T	rs6444724	C
rs1358856	No Data	rs6591147	C
rs1410059	T	rs6811238	No Data
rs1478829	No Data	rs7205345	G,C
rs1523537	No Data	rs7229946	G
rs1554472	G	rs740598	G
rs1821380	C	rs7520386	No Data
rs2073383	No Data	rs7704770	G,A
rs214955	C	rs985492	G
rs2272998	C	rs987640	No Data
rs2503107	No Data	rs9951171	A

Five to 19 SNPs were detected from each sample (Table 3). Data from the other samples (can be found in Appendix E). A reference profile for sample 103 existed; the reference profile

was generated in a previous study by TaqMan® assays and mass spectrometry. Profiles for sample 103 and positive control HL-60 were compared to reference profiles to determine accuracy. These comparisons can be seen in Table 4.

TABLE 4. PARTIAL PROFILES OF SAMPLES 103 AND HL-60 COMPARED TO REFERENCE PROFILES. The partial profiles generated from the data produced by the alignment of sample 103 and HL-60 to hg19 were compared to reference profiles generated from TaqMan® assays and mass spectrometry. Asterisks denote samples that were not in concordance with the corresponding reference.

SNPs	HL-60	HL-60 Reference	103	103 Reference
rs10092491	No Data	C,T	No Data	C
rs1019029	No Data	C,T	No Data	T
rs10488710	C	C,G	No Data	C,G
rs1058083	No Data	G	No Data	A
rs1109037	No Data	A	No Data	G
rs12997453	No Data	G	No Data	A,G
rs13134862	A	A	No Data	G
rs13182883	G	G	No Data	G
rs13218440	No Data	G	G	G
rs1336071	T *	A	No Data	A
rs1358856	No Data	A	No Data	A
rs1410059	T	C,T	No Data	C
rs1478829	No Data	A,T	No Data	T
rs1523537	No Data	C	No Data	C,T
rs1554472	G *	C,T	No Data	C
rs1821380	C	G	No Data	C,G
rs2073383	No Data	C	No Data	C,T
rs214955	C *	A,G	No Data	A,G
rs2272998	C	C	G	G
rs2503107	No Data	A,C	No Data	C
rs2567608	No Data	A,G	No Data	A
rs279844	A	A	No Data	T
rs315791	No Data	A	No Data	C
rs321198	No Data	C	T *	C
rs338882	No Data	C,T	No Data	C,T
rs3780962	No Data	C	No Data	C,T
rs445251	No Data	G	No Data	C
rs447818	No Data	G	C,T *	A,G
rs560681	A	A	No Data	A,G
rs6444724	C	C,T	No Data	T
rs6591147	C	C	No Data	C
rs6811238	No Data	G	No Data	G
rs7205345	G,C	C,G	C	C,G
rs7229946	G	G	No Data	A,G
rs740598	G	A,G	No Data	A
rs7520386	No Data	A,G	No Data	G
rs7704770	G,A	A	No Data	C
rs985492	G *	C	No Data	A
rs987640	No Data	T	No Data	A
rs9951171	A	A	No Data	A

For alignment to hg19, HL-60 had nine SNP calls that were in concordance with the reference profile, five calls that were partially correct, and four calls that were not in concordance with the reference profile. The five calls that were partially correct were true heterozygotes, however only one SNP was called; this is thought to be a result of low coverage. Sample 103 had three SNP calls which were in concordance with the reference profile and two calls that were not. The data generated at the locations of interest had extremely low coverage; the maximum coverage across all samples at the 40 locations was three reads. For this reason, heterozygosity and homozygosity designations could not be definitively determined. However, it should be noted that some locations across various samples did result in heterozygous calls, and 50 percent were accurate. During the analysis of this particular set of data large numbers of reads with high coverage were mapped to unmapped scaffolds in the genome. Unmapped scaffolds are places in the genome that we know exist, however we do not know their precise location [88].

RESULTS FOR ALIGNMENT TO A CHROMOSOME

Only two positions, rs2073383 and rs987640, were viewed on the alignment of sample 449 to chromosome 22. There were additional reads aligned to chromosome 22, however the reads did not span the two locations of the SNPs. The additional reads had an increased depth of coverage when comparing coverage to hg19 and they were close to the locations of our SNPs; the increased coverage and proximity to our SNPs led us to believe that aligning the samples to a more specific reference (chromosome compared to the entire genome) could result in an overall increase in depth of coverage as well as an increase in reads mapped to locations of interest. An example of the additional reads produced can be found in Appendix F; additional reads are

shown by a comparison of the alignment to the whole genome and the alignment to chromosome 22.

RESULTS FOR ALIGNMENT TO A MERGED REFERENCE FILE

Since alignment to a chromosome indicated increased coverage when compared to alignment to hg19, we proposed that performing an alignment to an even more specific region (SNP and flanking sequence compared to an entire chromosome) would result in a further increase in depth of coverage. We utilized a merged reference file rather than aligning samples to 40 independent FASTA files for time efficiency. Locations of the SNPs had to be found in the merged reference file. The first potential SNP location was found to be at 50,251bp; this is a result of the 250bp default setting combined with the extra 50,000bp put on either side of the variant during the creation of the FASTA files. Hereafter, the approximate locations of the SNPs were found by adding 100,500bp to the previous SNP; reasoning behind this addition method is as follows: 50,250bp downstream of the current SNP and 50,250bp upstream of the adjacent SNP must be combined to find the subsequent SNP. FASTA sequences of the 40 SNPs from NCBI's GenBank database [89] were found and used in order to determine the specific location of each of the variants for this particular alignment. Table 5 displays the variant locations. The resulting SNP genotypes for sample 103 and positive control HL-60 were compared to reference genotypes to see how well using the MinION device for identity testing works.

TABLE 5. SNP LOCATIONS ON THE MERGED REFERENCE. Individual SNP positions on the reference, SNPs_100kb_merged_even, were determined by adding the number of base pairs on either side of adjacent variants. The first potential SNP location was found to be at 50,251bp; this is a result of the 250bp default setting combined with the extra 50,000bp put on either side of the variant during the creation of the FASTA files. Hereafter, the approximate locations of the SNPs were found by adding 100,500bp to the previous SNP; reasoning behind this addition method is as follows: 50,250bp downstream of the current SNP and 50,250bp upstream of the adjacent SNP must be combined to find the subsequent SNP.

SNPs	Position (bp)	SNPs	Position (bp)
rs10092491	50251	rs2567608	2060251
rs1019029	150751	rs279844	2160751
rs10488710	251251	rs315791	2261251
rs1058083	351751	rs321198	2361251
rs1109037	452251	rs338882	2462251
rs12997453	552751	rs3780962	2562751
rs13134862	653251	rs445251	2663251
rs13182883	753751	rs447818	2763751
rs13218440	854251	rs560681	2864251
rs1336071	954751	rs6444724	2964751
rs1358856	1055251	rs6591147	3065251
rs1410059	1155751	rs6811238	3165751
rs1478829	1256251	rs7205345	3266251
rs1523537	1356751	rs7229946	3366751
rs1554472	1457251	rs740598	3467251
rs1821380	1557751	rs7520386	3567751
rs2073383	1658251	rs7704770	3668251
rs214955	1758751	rs985492	3768751
rs2272998	1859251	rs987640	3869251
rs2503107	1959751	rs9951171	3969751

Alignments of samples 103 and HL-60 to the merged reference file also resulted in varied partial profiles across the two samples, similar to the profiles generated via alignment to hg19. However, data produced here ranged from 16 to 32 SNPs per sample being called. The profiles of samples 103 and HL-60 were compared to reference profiles to determine accuracy. These comparisons can be seen in Table 6.

TABLE 6. PARTIAL PROFILES OF SAMPLES 103 AND HL-60 COMPARED TO REFERENCE PROFILES. The partial profiles generated from the data produced by the alignment of samples 103 and HL-60 to the merged reference were compared to reference profiles generated from TaqMan® assays and mass spectrometry. Asterisks denote samples that were not in concordance with the corresponding reference. Ambiguity codes are shown in parenthesis.

SNPs	Reads Mapped	HL-60	HL-60 Reference	Reads Mapped2	103	103 Reference
rs10092491	No Data	No Data	C,T	No Data	No Data	C
rs1019029	45	(X/N) 16%A, 4%C, 71%G, 4%T	C,T	5	(R) 20%A, 80%G *	T
rs10488710	7	(D) 43%A, 43%G, 9%T	C,G	1	100%G	C,G
rs1058083	No Data	No Data	G	No Data	No Data	A
rs1109037	1	100%C *	A	No Data	No Data	G
rs12997453	1	100%C *	G	3	100%C *	A,G
rs13134862	148	(X/N) 18%A, 60%C, 4G, 18%T	A	19	(B) 79%C, 11%G, 11%T	G
rs13182883	1	100%A *	G	No Data	No Data	G
rs13218440	No Data	No Data	G	1	100%T *	G
rs1336071	56	(X/N) 13%A, 7%C, 14%G, 37%T	A	5	100%T *	A
rs1358856	10	(H) 40%A, 50%C, 10%T	A	No Data	No Data	A
rs1410059	1	100%A *	C,T	No Data	No Data	C
rs1478829	6	100%T	A,T	No Data	No Data	T
rs1523537	No Data	No Data	C	No Data	No Data	C,T
rs1554472	2	100%C	C,T	11	100%T *	C
rs1821380	No Data	No Data	G	No Data	No Data	C,G
rs2073383	1	100%C	C	11	100%C	C,T
rs214955	No Data	No Data	A,G	No Data	No Data	A,G
rs2272998	22	100%A *	C	11	100%A *	G
rs2503107	26	(X/N) 15%A, 4%C, 4%G, 77%T	A,C	33	100%T *	C
rs2567608	1	100%C *	A,G	No Data	No Data	A
rs279844	5	100%C *	A	No Data	No Data	T
rs315791	3	100%A	A	No Data	No Data	C
rs321198	No Data	No Data	C	11	100%T *	C
rs338882	No Data	No Data	C,T	No Data	No Data	C,T
rs3780962	4	(R) 50%A, 50%G *	C	No Data	No Data	C,T
rs445251	41	(X/N) 65%A, 5%C, 10%G, 22%T	G	4	(V) 25%A, 25%C, 50%G	C
rs447818	1	100%A *	G	2	100%C *	A,G
rs560681	3	100%A	A	No Data	No Data	A,G
rs6444724	1	100%A *	C,T	No Data	No Data	T
rs6591147	1	100%G *	C	No Data	No Data	C
rs6811238	1	100%C *	G	No Data	No Data	G
rs7205345	14	(B) 57%C, 21%G, 21%T	C,G	6	(B) 67%C, 17%G, 17%T	C,G
rs7229946	212	(X/N) 22%A, 5%C, 58%G, 16%T	G	44	(X/N) 14%A, 7%C, 66%G, 1%T	A,G
rs740598	2	100%T *	A,G	No Data	No Data	A
rs7520386	No Data	No Data	A,G	No Data	No Data	G
rs7704770	2	100%T *	A	1	100%T *	C
rs985492	1	100%C	C	No Data	No Data	A
rs987640	7	(K) 14%G, 86%T	T	No Data	No Data	A
rs9951171	1	100%G *	A	No Data	No Data	A

HL-60 had five SNP calls that were in concordance with the merged reference profile and 15 calls that were not; sample 103 had two SNP calls which were in concordance with the reference profile and ten calls that were not. While some of the data generated at the locations of interest still had extremely low coverage (similar to that generated by alignment to hg19), some areas had substantially more coverage; the maximum coverage across all samples at the 40 locations was 212 reads. While this method allowed for an increase in depth of coverage, there were more non-concordant allele calls than in the method where samples were aligned to hg19.

RESULTS FOR ALIGNMENT TO INDIVIDUAL FASTA FILES

Since the merged reference file was comprised of the 40 individual FASTA files, we would expect to see identical results when comparing depth of coverage and allele calls for alignments to the merged reference file and alignments to the individual FASTA files. However, alignments to the individual FASTA files had an increased depth of coverage when compared to all other methods. Also, the majority of reads for this alignment method did not span the areas of interest. A limited number of alignments were performed using this method due to time efficiency. One alignment of HL60 to an individual FASTA file took approximately 54 minutes to complete. Multiple alignments can be performed at the same time, but when this was done the time needed to complete an alignment increased.

SUMMARY OF COMPARISONS

When comparing the four methods of alignment used in this study, it was found that aligning samples to a smaller portion of the genome resulted in increased depth of coverage and an increase in total reads mapped. Alignments of samples to individual FASTA files and the

merged reference file had more reads at locations of interest when compared to alignments of samples to the entire human genome. However, even when aligning to a smaller region of the genome, there were multiple areas of interest that had either no coverage or coverage as low as one read.

Sample 103 and positive control HL-60 were compared to their respective reference profiles. The sample and the positive control displayed the most concordance when aligned to hg19 as opposed to being aligned to the merged reference file. Alignments of sample 103 and HL-60 to both an individual chromosome and individual FASTA files were not compared to reference profiles due to the lack of reads at locations of interest. Additionally, alignments of the samples to every chromosome and all 40 FASTA files would have to be completed in order to compare the accuracy of these methods to that of those that used hg19 and the merged reference file.

Ultimately, the customized data pipeline for the four different alignment methods used allowed us to interrogate a panel of SNPs. Reliable heterozygote determination could not be reached, due to the lack of sufficient depth of coverage of the autosomal markers of interest. The method of alignment of samples to hg19 produced the most accurate results; the method of alignment to the merged reference file produced the most reads spanning the SNPs of interest. Performing alignments to individual FASTA files is not time efficient, especially when 40 alignments would have to be done per sample.

Very little prior research has been done on SNP genotyping using the Oxford Nanopore sequencer. Most research utilizing ONT's MinION involves using protocols that rely on PCR. Zaaijer S *et al*. [59] is the only known published research to date that has interrogated SNPs to generated identity profiles without the use of some sort of amplification. While our study

interrogated a set number of SNPs to generate a genotype, Zaaijer S *et al*. sequenced a varied number of SNPs until a match out of a database was made. Their method involved the continuation of SNP interrogation until a definitive match was reached, much different than our method.

CHAPTER 4

LIMITATIONS

There are a number of existing limitations that should be addressed when considering future research. A limitation specific to the forensic field includes the number of SNPs that are needed in order to obtain the same statistical value that the FBI approved CODIS STRs produce. A panel of approximately 80 SNPs would be needed in order to obtain similar match probabilities to those produced by the 20 STRs used currently. Manually interrogating 80 SNPs is not time efficient for a crime laboratory's workflow. The process of interrogating SNPs could be automated using software such as Galaxy [89]; however, in order to use software for the interrogation of SNPs a certain depth of coverage has to obtained.

Another limitation of the applicability of the MinION in forensics is the amount of suggested input DNA needed in order to run the MinION sequencer according to the manufacturer's standards. Evidentiary samples obtained from forensic cases vary greatly in the amount of sample available to be consumed, and often times 200ng of DNA may not be available, that is why the use of PCR is extremely prevalent in the forensic field.

Low coverage when performing nanopore sequencing can be a result of the existing sampling bias from trying to interrogate 40 nucleotides against a background of 3.2 billion nucleotides; this leads to coverage of random parts of the whole genome and can lead to a depth of coverage as low as one read. When coverage is this low, it is impossible to detect

heterozygous alleles and determine if alleles at specific loci are truly homozygous, rendering an incomplete DNA profile.

Another challenge of nanopore sequencing is the bioinformatics background that is needed in order to be well versed with generating profiles from large amounts of data and being able to then interrogate variants. The workflow described in Chapter 2 Materials and Methods for converting FAST5 files to BAM files proves to be a complicated process that involves several computerized conversion steps. Currently, coursework and experience dealing with bioinformatics is not a requirement that DNA analysts possess in order to gain employment at a crime lab, resulting in a lack of the knowledge that would be favorable if implementing this process into standard operating procedures. as technologies continue to develop the need to have advanced computer skills has become apparent in many career fields to the point where many universities require some type of advanced computer or coding class be taken prior to graduation. The forensics field also requires DNA analysts to take a set number of hours towards continued education each year, presenting an ample opportunity for nanopore sequencing to be introduced to already practicing forensic scientists.

Even though considerable limitations currently exist for the use of ONT's MinION for forensic purposes does not mean that this technology is not useful within the forensics field. Overall, any existing limitations could be eventually overcome.

CHAPTER 5

FUTURE RESEARCH

Although this study focused on ONT's MinION applicability for SNP analysis, the majority of research associated with nanopore sequencing technology has a much broader scope. Studies utilizing MinION sequencing have resulted in publications centering around areas such as environmental biology, clinical studies of specific and infectious diseases, cancer research, the microbiome, and animal related research, to name a few [90]. As an example of the wide range of topics covered, recent publications include topics such as identifying novel genetic variations in a strain of tuberculosis, finding RNA splicing profiles of a calcium channel gene in the human brain, detection of multiple virus species involved in bovine respiratory diseases, and determining that specific liquid substrates are capable of repowering sewage microbiomes [91-94].

Future directions for research using MinION sequencing for forensics could involve running samples and HL-60 on new chemistries and comparing generated profiles to reference profiles in order to compare and contrast the improvements that the new chemistries have. Nanopore sequencing protocols that utilize PCR could also be performed and compared to sequencing runs of native DNA in order to measure applicability if amplification of DNA is implemented. Adjusting the penalties for insertions and deletions during alignments could also be attempted in order to force data to align to the positions of interest by down weighting InDels.

Multiple validations studies would also have to be performed in accordance with Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines. Under these guidelines sensitivity studies, reproducibility studies, and mixture studies would have to be completed using both reference and mock evidentiary samples to determine the real value such technology would have in forensic casework [95].

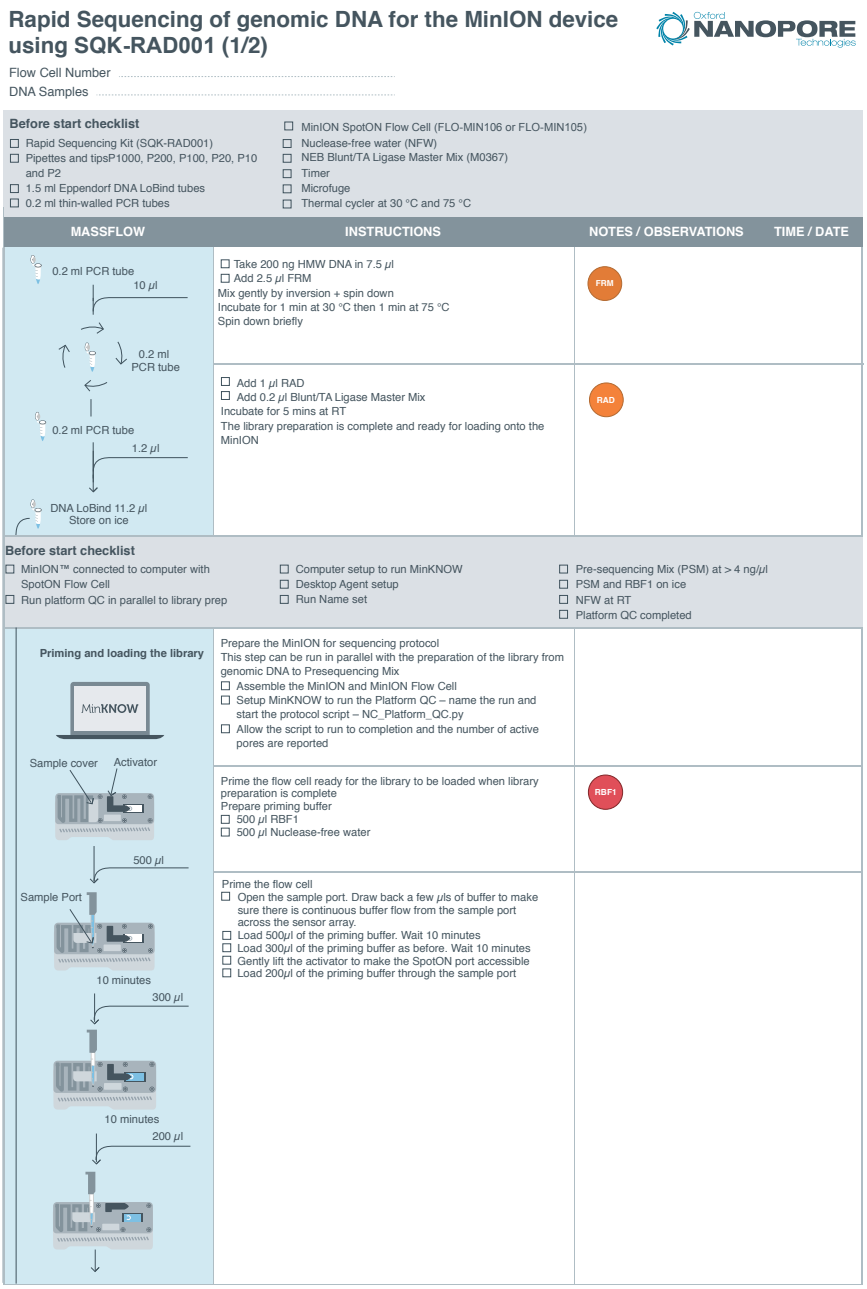
CHAPTER 6

CONCLUSIONS

This study demonstrated that ONT's MinION could be used to generate DNA profiles from samples that did not previously undergo PCR. Profiles determined were partial and varied depending on the alignment method used. The best method used to accurately determine alleles was the method using the alignment to hg19. This SNP genotyping method employed was able to accurately determine about half of the alleles present in the positive control. The best method used to attain the greatest depth of coverage was the method using the merged reference alignment. Using this method, the positive control had a total of thirty-two out of forty alleles called. At this point heterozygosity and homozygosity could not be determined from the data generated because of the lack of sufficient depth of coverage. Increased coverage and validation studies setting thresholds for detection would have to be performed in order to be able to confidently make that decision.

As nanopore technology continues to progress, the limiting obstacles that currently exist will be able to be overcome. Until such restraints are overcome, studies such as this one will have to continue in order to conclusively determine the role that nanopore sequencing and SNP genotyping as a whole can have in the ever-evolving forensic crime laboratory.

APPENDIX A



Rapid Sequencing of genomic DNA for the MinION device using SQK-RAD001 (1/2)

Oxford NANOPORE Technologies

Flow Cell Number .....

DNA Samples .....

Before start checklist

- ☐ Rapid Sequencing Kit (SQK-RAD001)
- ☐ Pipettes and tipsP1000, P200, P100, P20, P10 and P2
- ☐ 1.5 ml Eppendorf DNA LoBind tubes
- ☐ 0.2 ml thin-walled PCR tubes
- ☐ MinION SpotON Flow Cell (FLO-MIN106 or FLO-MIN105)
- ☐ Nuclease-free water (NFW)
- ☐ NEB Blunt/TA Ligase Master Mix (M0367)
- ☐ Timer
- ☐ Microfuge
- ☐ Thermal cycler at 30 °C and 75 °C

MASSFLOW	INSTRUCTIONS	NOTES / OBSERVATIONS	TIME / DATE
0.2 ml PCR tube → 10 µl → 0.2 ml PCR tube → 0.2 ml PCR tube → 1.2 µl → DNA LoBind 11.2 µl Store on ice	☐ Take 200 ng HMW DNA in 7.5 µl ☐ Add 2.5 µl FRM Mix gently by inversion + spin down Incubate for 1 min at 30 °C then 1 min at 75 °C Spin down briefly	FRM	
	☐ Add 1 µl RAD ☐ Add 0.2 µl Blunt/TA Ligase Master Mix Incubate for 5 mins at RT The library preparation is complete and ready for loading onto the MinION	RAD	

Before start checklist

- ☐ MinION™ connected to computer with SpotON Flow Cell
- ☐ Run platform QC in parallel to library prep
- ☐ Computer setup to run MinKNOW
- ☐ Desktop Agent setup
- ☐ Run Name set
- ☐ Pre-sequencing Mix (PSM) at > 4 ng/µl
- ☐ PSM and RBF1 on ice
- ☐ NFW at RT
- ☐ Platform QC completed

Priming and loading the library MinKNOW Sample cover, Activator 500 µl Sample Port 10 minutes 300 µl 10 minutes 200 µl	Prepare the MinION for sequencing protocol This step can be run in parallel with the preparation of the library from genomic DNA to Presequencing Mix ☐ Assemble the MinION and MinION Flow Cell ☐ Setup MinKNOW to run the Platform QC – name the run and start the protocol script – NC_Platform_QC.py ☐ Allow the script to run to completion and the number of active pores are reported		
	Prime the flow cell ready for the library to be loaded when library preparation is complete Prepare priming buffer ☐ 500 µl RBF1 ☐ 500 µl Nuclease-free water	RBF1	
	Prime the flow cell ☐ Open the sample port. Draw back a few µls of buffer to make sure there is continuous buffer flow from the sample port across the sensor array. ☐ Load 500µl of the priming buffer. Wait 10 minutes ☐ Load 300µl of the priming buffer as before. Wait 10 minutes ☐ Gently lift the activator to make the SpotON port accessible ☐ Load 200µl of the priming buffer through the sample port		

Figure 4. ONT's RAPID SEQUENCING OF GENOMIC DNA FOR THE MinION DEVICE PROTOCOL (SQK-RAD001). The manufacturer's protocol for generating 1D Rapid libraries. This instruction set was used for samples 101, 102, 103, and 433. Image was taken from ONT [96].

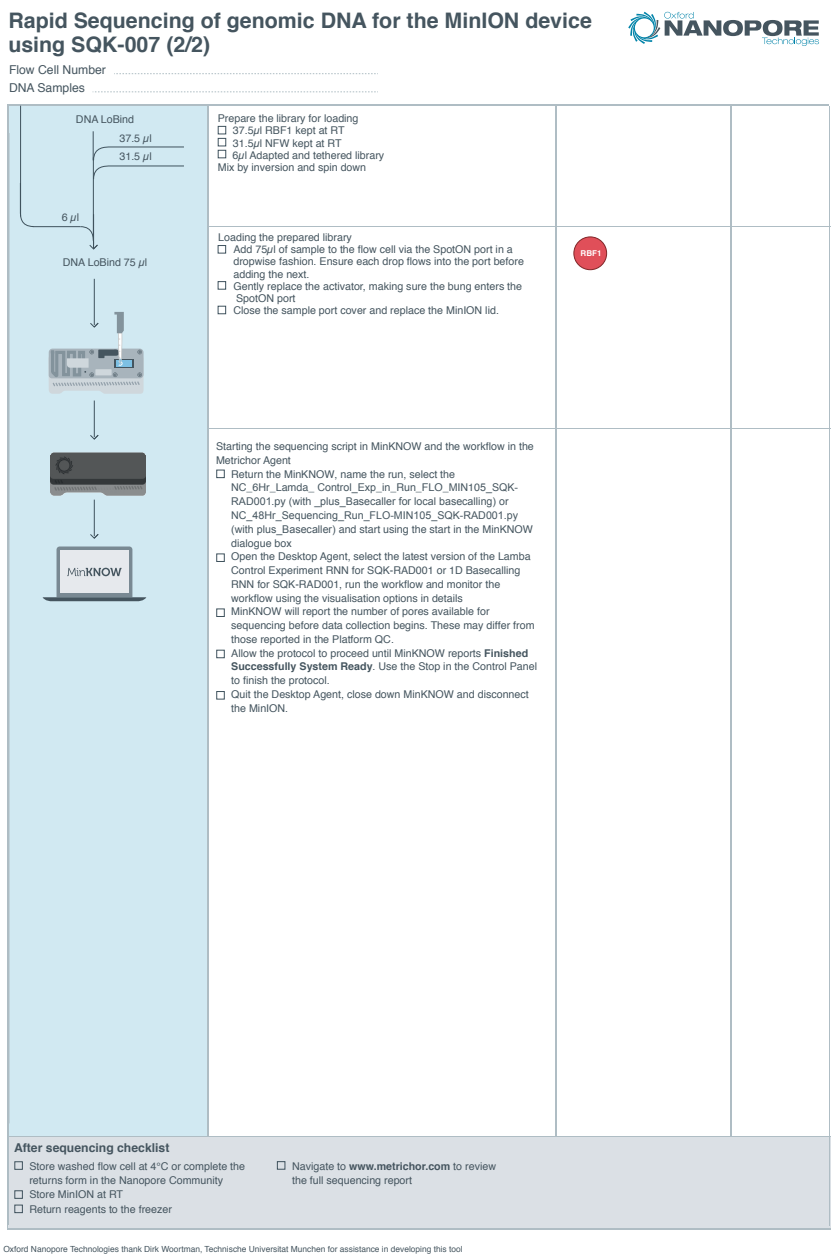
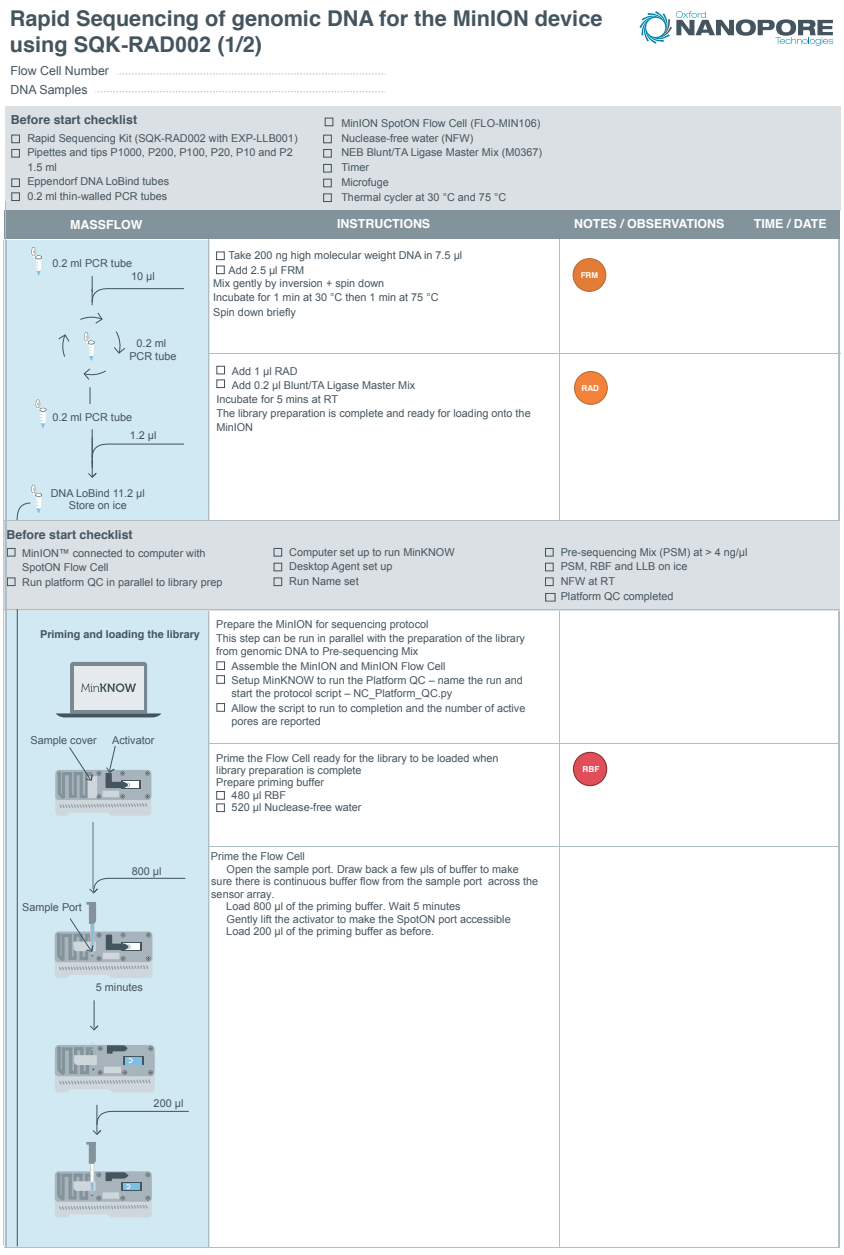


Figure 4. ONT's RAPID SEQUENCING OF GENOMIC DNA FOR THE MinION DEVICE PROTOCOL (SQK-RAD001). The manufacturer's protocol for generating 1D Rapid libraries. This instruction set was used for samples 101, 102, 103, and 433. Image was taken from ONT [96]



Rapid Sequencing of genomic DNA for the MinION device using SQK-RAD002 (1/2)

Oxford NANOPORE Technologies

Flow Cell Number .....

DNA Samples .....

Before start checklist

- ☐ Rapid Sequencing Kit (SQK-RAD002 with EXP-LLB001)
- ☐ Pipettes and tips P1000, P200, P100, P20, P10 and P2
- ☐ 1.5 ml
- ☐ Eppendorf DNA LoBind tubes
- ☐ 0.2 ml thin-walled PCR tubes
- ☐ MinION SpotON Flow Cell (FLO-MIN106)
- ☐ Nuclease-free water (NFW)
- ☐ NEB Blunt/TA Ligase Master Mix (M0367)
- ☐ Timer
- ☐ Microfuge
- ☐ Thermal cycler at 30 °C and 75 °C

MASSFLOW	INSTRUCTIONS	NOTES / OBSERVATIONS	TIME / DATE
0.2 ml PCR tube 10 µl 0.2 ml PCR tube 0.2 ml PCR tube 1.2 µl DNA LoBind 11.2 µl Store on ice	☐ Take 200 ng high molecular weight DNA in 7.5 µl ☐ Add 2.5 µl FRM Mix gently by inversion + spin down Incubate for 1 min at 30 °C then 1 min at 75 °C Spin down briefly	FRM	
	☐ Add 1 µl RAD ☐ Add 0.2 µl Blunt/TA Ligase Master Mix Incubate for 5 mins at RT The library preparation is complete and ready for loading onto the MinION	RAD	

Before start checklist

- ☐ MinION™ connected to computer with SpotON Flow Cell
- ☐ Run platform QC in parallel to library prep
- ☐ Computer set up to run MinKNOW
- ☐ Desktop Agent set up
- ☐ Run Name set
- ☐ Pre-sequencing Mix (PSM) at > 4 ng/µl
- ☐ PSM, RBF and LLB on ice
- ☐ NFW at RT
- ☐ Platform QC completed

Priming and loading the library MinKNOW Sample cover Activator 800 µl Sample Port 5 minutes 200 µl	Prepare the MinION for sequencing protocol This step can be run in parallel with the preparation of the library from genomic DNA to Pre-sequencing Mix ☐ Assemble the MinION and MinION Flow Cell ☐ Setup MinKNOW to run the Platform QC – name the run and start the protocol script – NC_Platform_QC.py ☐ Allow the script to run to completion and the number of active pores are reported		
	Prime the Flow Cell ready for the library to be loaded when library preparation is complete Prepare priming buffer ☐ 480 µl RBF ☐ 520 µl Nuclease-free water	RBF	
	Prime the Flow Cell Open the sample port. Draw back a few µls of buffer to make sure there is continuous buffer flow from the sample port across the sensor array. Load 800 µl of the priming buffer. Wait 5 minutes Gently lift the activator to make the SpotON port accessible Load 200 µl of the priming buffer as before.		

Figure 5. ONT's RAPID SEQUENCING OF GENOMIC DNA FOR THE MinION DEVICE PROTOCOL (SQK-RAD002). The manufacturer's protocol for generating 1D Rapid libraries. This instruction set was used for samples 425, 441, 442, 449, 459, and HL60. This method utilized library loading beads. Image was taken from ONT [97].

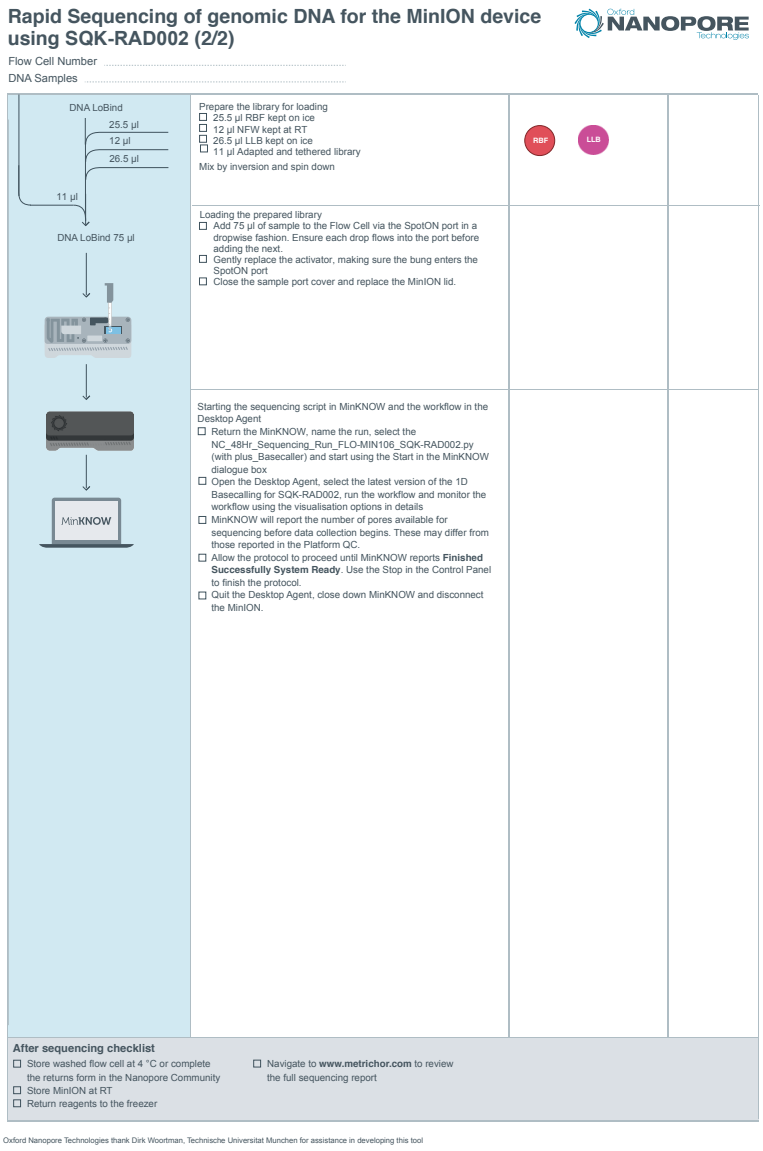


Figure 5. ONT's RAPID SEQUENCING OF GENOMIC DNA FOR THE MinION DEVICE PROTOCOL (SQK-RAD002). The manufacturer's protocol for generating 1D Rapid libraries. This instruction set was used for samples 425, 441, 442, 449, 459, and HL60. This method utilized library loading beads. Image was taken from ONT [97].

APPENDIX B

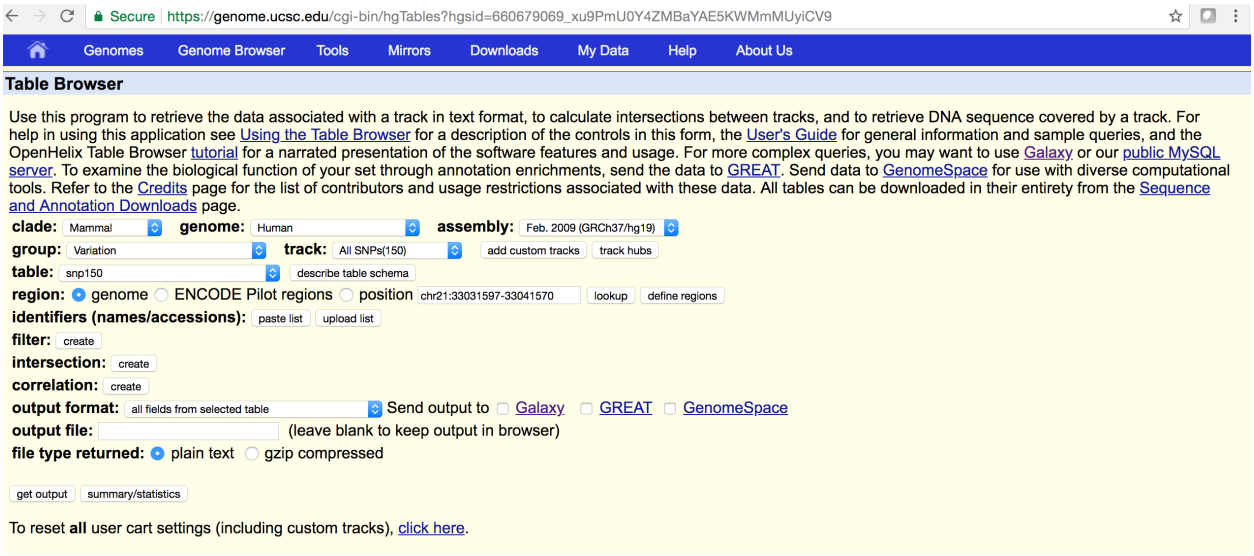


Figure 6. UCSC Table Browser. Settings were altered on the Table Browser tool in order to determine the specific locations on the chromosomes of each of the 40 SNPs used in this study. Assembly was changed to hg19, group was set to Variation, track was set to All SNPs(150), and the 40 SNPs were input into identifiers paste list [83].

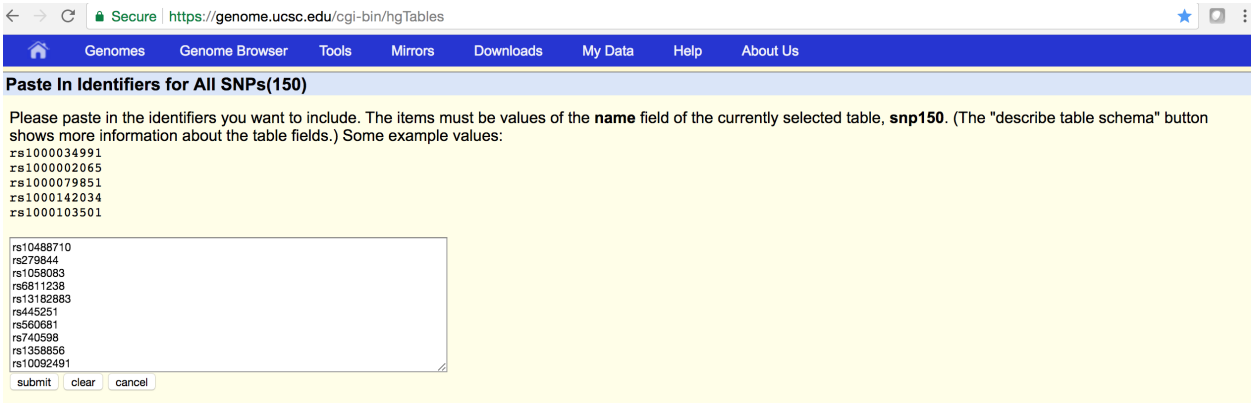


Figure 7. USCS Table Browser Identifiers (Names/Accessories) Window. The 40 SNPs used were typed into the window in no particular order; this allowed for the generation of specific information [83].

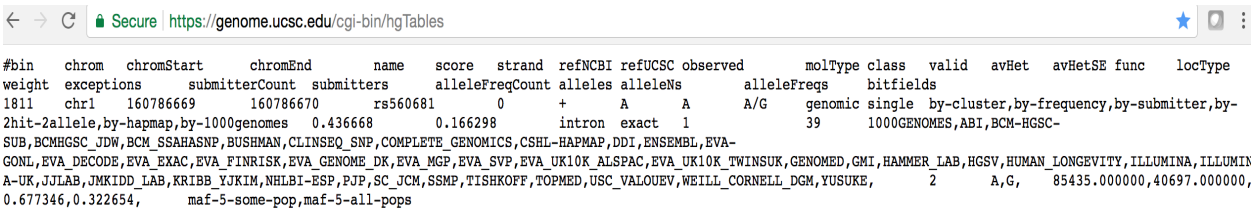


Figure 8. UCSC's Table Browser Output File. Below is an example of the output file generated using the Table Browser function. The example below is information regarding rs560681, which is on chromosome 1 specifically at 160786670bp.

APPENDIX C

Sample	Reads Mapped	Bases Mapped
HL60	226317	2673955811
101	281951	973314868
102	96486	376158673
103	150528	631437287
425	87679	367685729
433	414351	1282947569
441	242328	1161125124
442	221478	1159427343
449	170072	1318376493
459	168861	1167841465

Table 7. Total Reads and Total Bases Mapped. The number of total reads and total bases mapped for each of the 10 samples (HL-60, 101, 102, 103, 425, 433, 441, 442, 449, and 459) was obtained from the summary statistics file generated with the use of Samtools.

APPENDIX D

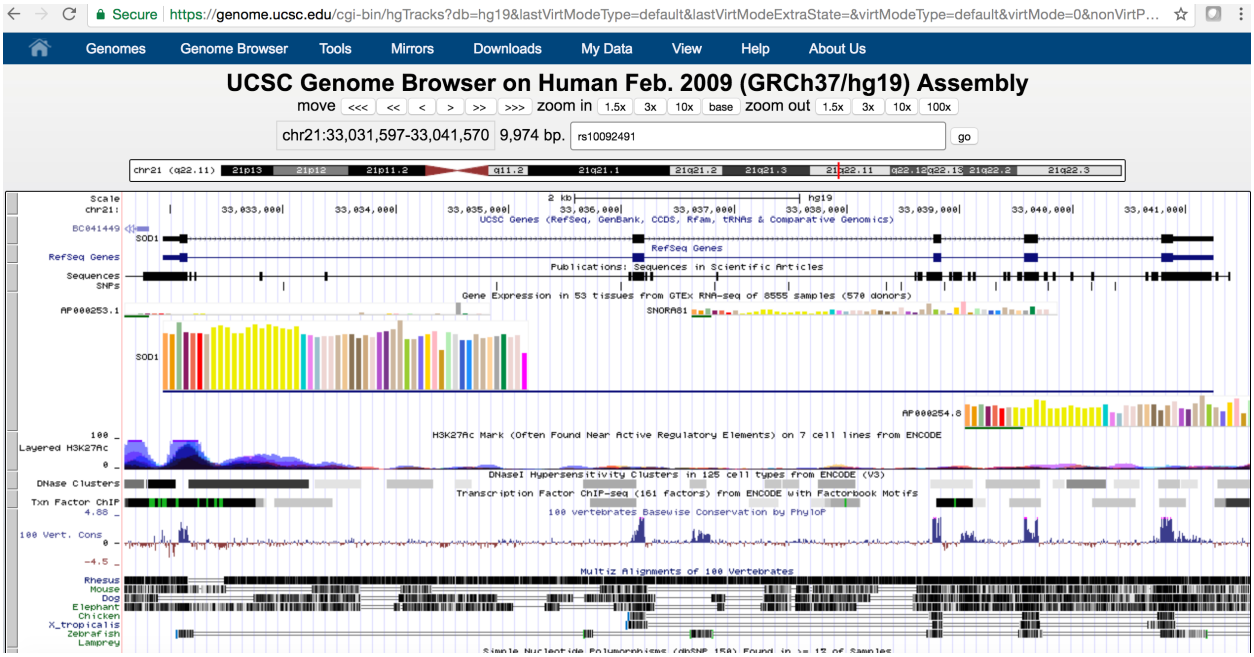


Figure 9. UCSC's Genome Browser. Each individual SNP was searched via the Genome Browser tool. After searching for a specific SNP in the search bar, the option to choose one of multiple links was given. The dbSNP 150 link was chosen. From there, the View tab was opened and the option to Get DNA was chosen [85].

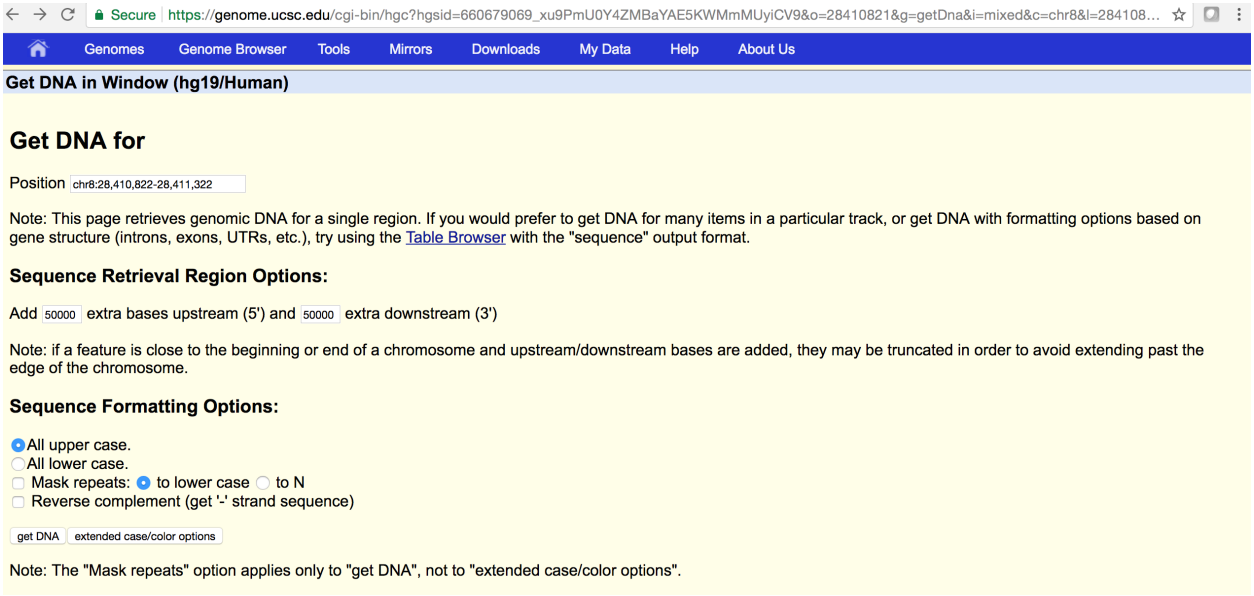


Figure 10. UCSC's Genome Browser Get DNA Window. UCSC's Genome Browser default setting for sequence retrieval around a SNP is 250bp on either side of the variation. An additional 50,000bp were added on either side of the variation, so that the FASTA file generated had 50,250bp both upstream and downstream of the SNP [85].

```
>hg19_dna range=chr8:28360822-28461322 5'pad=50000 3'pad=50000 strand=+ repeatMasking=none
TGCTATTTTAAATTATATCTTATAAATTTAAAAATGTATATGTTTTGCTT
CAGACTCTAATAGATTGAAGAATTCACATTTGGTGAATGTTACTCTTATA
TTTTAGACATTATACTTTTATAAAGAAATTTGAAACATCAGAACATGGGA
GATAGCAGCCATTCAGAAGTTTAAAATTTAAGAATAACTCCCAAGCAGTG
AAAAAGTTTTTGTCAATGCGAATTATCTGTTCTTGATCTCTCAGAATTTT
ATGAAAGTTTGCTGCCTTATTTGAATTTTTCAGTGGTAATTCTTCAGACA
TTTCTTAATTATTTTAGTTTAGAAACTATTTCAAATTTAGAAACTATGGA
CACATTAATTGATATAGTTAATAATAGGTTTAAATAAGGGTATATCTGTC
TATAATGGTTTTAGAATTGAGGACAGGGATATATATACCAGATAAGAACT
```

Figure 11. FASTA File. The depiction below is an example of the first part of 1 of the 40 FASTA files generated using UCSC's genome browser. These FASTA files were then used to create a merged sequence file. Note the header at the top of the example; this line was removed using a BASH command.

APPENDIX E

SNPs	101	102	103	425	433	441	442	449	459
rs10092491	No Data	No Data	No Data	No Data	No Data	C	No Data	No Data	No Data
rs1019029	A	No Data	No Data	G	No Data	No Data	No Data	No Data	No Data
rs10488710	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
rs1058083	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
rs1109037	A	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
rs12997453	No Data	No Data	No Data	No Data	No Data	G	No Data	No Data	No Data
rs13134862	A	No Data	No Data	A	No Data	G	G	No Data	No Data
rs13182883	G	No Data	No Data	No Data	No Data	G	No Data	No Data	No Data
rs13218440	No Data	No Data	G	No Data	No Data	G	No Data	No Data	No Data
rs1336071	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
rs1358856	No Data	No Data	No Data	No Data	C	No Data	C	No Data	A
rs1410059	T	No Data	No Data	No Data	No Data	A	C	No Data	No Data
rs1478829	No Data	A	No Data	No Data	No Data	No Data	A	No Data	A
rs1523537	T	C	No Data	No Data	No Data	No Data	No Data	No Data	No Data
rs1554472	No Data	No Data	No Data	No Data	G	No Data	No Data	G,A	No Data
rs1821380	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
rs2073383	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	C
rs214955	C	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
rs2272998	G	G	G	No Data	No Data	No Data	No Data	No Data	No Data
rs2503107	No Data	C	No Data	No Data	No Data	No Data	C	No Data	No Data
rs2567608	No Data	T	No Data	No Data	C	No Data	No Data	No Data	No Data
rs279844	No Data	No Data	No Data	No Data	No Data	No Data	A	T	No Data
rs315791	No Data	No Data	No Data	C	No Data	No Data	No Data	No Data	No Data
rs321198	No Data	No Data	T	No Data	No Data	No Data	No Data	T	No Data
rs338882	No Data	A	No Data	No Data	No Data	No Data	No Data	No Data	A
rs3780962	No Data	No Data	No Data	No Data	No Data	G	G	No Data	No Data
rs445251	No Data	No Data	No Data	No Data	C	G	C	No Data	No Data
rs447818	C,T	No Data	C,T	No Data	T	No Data	No Data	No Data	T
rs560681	G,A	No Data	No Data	No Data	G	No Data	No Data	No Data	A
rs6444724	T	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
rs6591147	No Data	No Data	No Data	T	No Data	No Data	No Data	No Data	No Data
rs6811238	T	No Data	No Data	No Data	No Data	No Data	No Data	G	No Data
rs7205345	No Data	No Data	No Data	No Data	No Data	C	C	No Data	No Data
rs7229946	No Data	No Data	No Data	No Data	No Data	A	No Data	G	No Data
rs740598	No Data	No Data	No Data	No Data	A	No Data	No Data	No Data	A
rs7520386	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
rs7704770	No Data	No Data	No Data	No Data	G	No Data	G	No Data	No Data
rs985492	No Data	No Data	No Data	G	No Data	No Data	A	No Data	G
rs987640	T	No Data	No Data	No Data	A	No Data	No Data	No Data	No Data
rs9951171	G	A	No Data	No Data	G	No Data	No Data	G	No Data

Table 8. SNP Profiles (Samples 101, 102, 103, 425, 433, 441, 442, 449, and 459). The table shows profiles generated from sample alignments to hg19. All of the profiles are partial.

APPENDIX F

Figure 12. Additional Coverage on Chromosome 22. Only two positions, rs2073383 and rs987640, were viewed on the alignment of sample 449 to chromosome 22. There were additional reads aligned to chromosome 22, however the reads did not span the two locations of the SNPs. The additional reads had an increased depth of coverage when comparing coverage to hg19 and they were close to the locations of our SNPs (see red arrows); the increased coverage and proximity to our SNPs led us to believe that aligning the samples to a more specific reference (chromosome compared to the entire genome) could result in an overall increase in depth of coverage as well as an increase in reads mapped to locations of interest [81].

REFERENCES

1. Jeffrey AJ, Allen MJ, Hagelberg E, *et al*. (1992). Identification of the skeletal remains of Josef Mengele by DNA analysis. *Forensic Science International*, 56: 65-76.

2. Butler JM. (2006). Genetics and genomics of core short tandem repeat loci used in human identity testing. *J Forensic Sci*, 51(2): 253-265.

3. Gill P. (2002). Role of short tandem repeat DNA in forensic casework in the UK – past, present, and future perspectives. *BioTechniques*, 32: 366-372.

4. Kidd KK, Pakstis AJ, Speed WC, *et al*. (2013). Microhaplotype loci are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series*, 4:123-124.

5. Biesecker LG, Bailey-Wilson JE, Ballantyne J, *et al*. (2005). DNA identifications after the 9/11 World Trade Center attack. *Science*, 310: 1122-1123.

6. Brooks AJ. (1999). The essence of SNPs. *Gene*, 234: 177-186.

7. Wakeley J, Nielsen R, Liu-Cordero SN, *et al*. (2001). The discovery of single-nucleotide polymorphisms – and inferences about human demographic history. *Am J Hum Genet*, 69(6): 1332-1347.

8. Hiratsuka M, Tsukamoto N, Konno Y, *et al*. (2005). Forensic assessment of 16 single nucleotide polymorphisms analyzed by hybridization probe assay. *Exp Med*, 207: 255-261.

9. Lam CW, Lau KC, and Tong SF. (2010). Microarrays for personalized genomic medicine. *Advances in Clinical Chemistry*, 52: 1-18

10. John S, Shepard N, Liu G, *et al*. (2004). Whole-genome scan in a complex disease using 11,245 single nucleotide polymorphisms: comparison with microsatellites. *Am J Hunt Genet*. 75: 54-64.

11. Mei R, Galipeau PC, Prass C, *et al*. (2000). Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Research* 10(8): 1126-1137.

12. Baggerly K and Broom B. (2009). Analysis of microarray data. http://bioinformatics.mdanderson.org/MicroarrayCourse/Lectures09/snp1_bw.pdf (Accessed March 2018).

13. LaFramboise T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational, and technological advances. *Nucleic Acids Research*, 37: 4181-4193.

14. Mehta B, Daniel R, Phillips C, *et al*. (2016). Massively parallel sequencing of customised forensically informative SNP panels on the Miseq. *Electrophoresis*, 37: 2832-2840.

15. Sobrino B, Brion M, and Carracedo A. (2005). SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Science International*, 154(2): 181-194.

16. Mehta B, Daniel R, and McNevin D. (2017). HRM and SNaPshot as alternative forensic SNP genotyping methods. *Forensic Science, Medicine and Pathology*, 13(3): 293-301.

17. Applied Biosystems. SNaPshot Multiplex System for SNP Genotyping. https://assets.thermofisher.com/TFS-Assets/LSG/brochures/cms_101014.pdf (Accessed January 2018).

18. Illumina. Introduction to Next-Generation Sequencing. https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf (Accessed January 2018).

19. Illumina. Run Time Estimates. https://support.illumina.com/bulletins/2017/02/run-time-estimates-for-each-sequencing-step-on-illumina-sequenci.html (Accessed January 2018).

20. Nanopore Technologies. Products: MinION. https://nanoporetech.com/products/minion (Accessed December 2017).

21. Keats BJB and Sherman SL. (2013). *Emery and Rimoin's principles and practice of medical genetics*. Ed. Sixth. Academic Press. 1-12.

22. NIH. Single Nucleotide Polymorphisms (SNPs). https://ghr.nlm.nih.gov/primer/genomicresearch/snp (Accessed December 2017).

23. Labuschagne C, Dalton DL, Grobler JP, *et al*. (2017). SNP discovery and characterisation in white rhino (*Ceratotherium simum*) with application to parentage assignment. *Genet Mol Bio*, 40(1): 84-92.

24. Ryynanen HJ and Primmer CR. (2006). Single nucleotide polymorphisms (SNP) discovery in duplicated genomes: intron-primers exon-crossing (IPEC) as a strategy for avoiding amplification of duplicated loci in Atlantic salmon (*Salmo salar*) and other salmonid fishes. *BMC Genomics*, 7: 192.

25. Pertoldi C, Bijlsma R, and Loeschcke V. (2007). Conservation genetics in a globally changing environment: present problems, paradoxes and future challenges. *Biodivers Conserv*, 16: 4147-4163.

26. Murphy KM, Cooper A, and Tobias ES. (2014). The human genome, gene regulation, and genomic variation. *Handbook of pharmarcogenomics and stratified medicine*. Academic Press. 41-56.

27. Brody T. (2016). Biomarkers. *Clinical Trials*. Ed. Second. Academic Press. 377-419.

28. Pratt CW, Gill KJ, Barrett NM, *et al*. (2014). Symptoms and etiology of serious mental illness. *Psychiatric rehabilitation*. Ed. Third. Elsevier. 22-74.

29. NCBI. HapMap Resource Retiring. https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/ (Accessed December 2017).

30. NIH. International HapMap Project. https://www.genome.gov/10001688/international-hapmap-project/ (Accessed December 2017).

31. Warshawsky I. (2009). Molecular biology basics for the pathologist. *Cell and tissue based molecular pathology*. Churchill Livingstone. 3-9.

32. IGSR. About the 1000 Genomes Project. http://www.internationalgenome.org/about/ (Accessed December 2017).

33. Haraksingh RR and Snyder MP. (2013). Impacts of variation in the human genome on gene regulation. *J Mol Bio*, 425(21): 3970-3977.

34. Kidd KK and Speed WC. (2015). Criteria for selecting microhaplotypes: mixture detection and deconvolution. *Investigative Genetics*, 6:1

35. Shriver MD, Mei R, Parra EJ, *et al*. (2005). Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. *Hum Genomic*, 2: 81-89.

36. Kersbergen P, Van Duijn K, Kloosterman AD, *et al*. (2009). Developing a set of ancestry-informative DNA markers reflecting continental origins of humans. *BMC Genet*, 10: 69.

37. Kidd KK, Speed WC, Pakstis AJ, *et al*. (2014). Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Science International: Genetics*, 10: 23-32.

38. Ding L, Wiener T, Abebe M, *et al*. (2011). Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics*, 12: 622.

39. Butler JM. (2012). Single nucleotide polymorphisms and applications. *Advanced topics in forensic DNA typing: methodology*. Academic Press. 347-369.

40. Butler JM, Coble MD, and Vallone PM. (2007). STRs vs. SNPs: thoughts on the future of forensic DNA testing. *Forensic Sci Med Pathol*, 3: 200-205

41. Kidd KK, Pakstis AJ, Speed WC, *et al*. (2006). Developing a SNP panel for forensic identification of individuals. *Forensic Science International*, 164: 20-30.

42. Pakstis AJ, Speed WC, Fang R, *et al*. (2010). SNPs for a universal individual identification panel. *Hum Genet*, 127(3): 315-324.

43. SNPforID. Browser. http://spsmart.cesga.es/snpforid.php (Accessed January 2018).

44. NIST. Forensic SNP Information. http://strbase.nist.gov/SNP.htm (Accessed January 2018).

45. Sanchez JJ, Borsting C, Hallenberg C, *et al*. (2003). Multiplex PCR and minisequenicng of SNPs – a model with 35 Y chromosome SNPs. *Forensic Science International*, 137: 74-84.

46. Vallone PM and Butler JM. (2004). Y-SNP typing of U.S. African American and Caucasian samples using allele-specific hybridization and primer extension. *J Forensic Science*, 49: 723-732.

47. Vallone PM, Just RS, Coble MD, *et al*. (2004). A multiplex allele-specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome. *Int J Legal Med*, 118: 147-157.

48. Dixon LA, Murray CM, Archer EJ, *et al*. (2005). Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. *Forensic Science International*, 154: 62-77.

49. Brion M, Sanchez JJ, Balogh K, *et al*. (2005). Introduction of a single nucleotide polymorphism-based "major Y-chromosome haplogroup typing kit" suitable for predicting the geographical origin of male lineages. *Electrophoresis*, 23: 4411-4420.

50. Phillips C, Salas A, Sanchez JJ, *et al*. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet*, 1(3-4): 273-280.

51. Nelson TM, Just RS, Loreille O, *et al*. (2007). Development of a multiplex single base extension assay for mitochondrial DNA haplogroup typing. *Croat Med*, 48: 460-472.

52. Fang R, Pakstis AJ, Hyland F, *et al*. (2009). Multiplexed SNP detection panels for human identification. *Forensic Sci Int Genet*, 2(1): 538-539.

53. Walsh S, Liu F, Ballantyne KN, *et al*. (2011). IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci Int Genet*, 5(3): 170-180.

54. Sanchez JJ, Phillips C, Borsting C, *et al*. (2006). Development of a multiplex PCR assay detecting 52 autosomal SNPs, *International Congress Series*. 1288: 67-69.

55. Niedringhaus TP, Millanova D, Kerby MB, *et al*. (2011). Landscape of next-generation sequencing technologies. *Analytical Chemistry*, 83(12): 4327-4341.

56. Deamer D, Akeson M, and Branton D. (2016). Three decades of nanopore sequencing. *Nature Biotchnology*, 34(5): 618-524.

57. Nanopore Technologies. How does nanopore DNA/RNA sequencing work. https://nanoporetech.com/how-it-works. (Accessed March 2018).

58. Nanopore Technologies. Nanopore DNA sequencing. https://nanoporetech.com/resource-centre/videos/nanopore-dna-sequencing. (Accessed March 2018).

59. Zaaijer S, Gordon A, Speyer D, *et al*. (2017). Rapid DNA re-identification for cell line authentication and forensics.

60. Cornelis S, Gansemans Y, Deleye L, *et al*. (2017). Forensic SNP genotyping using nanopore MinION sequencing. *Sci Rep*, 7:41749

61. Church GM. (2006). Nanopore Sequencing. *Scientific American*, 294: 46-54.

62. Thorson K. (2017). Approaches to mitochondrial genome sequencing using the Oxford Nanopore MinION Device.

63. Lodish H, Berk A, Kaiser CA, *et al*. (2013). Molecular cell biology 7th edition. W.H. Freeman and Co. Chapter 12.

64. Quinn MJ. (2003). Parallel programming in C with MPI and OpenMP. McGraw-Hill Education Group.

65. Lu H, Giordano F, and Ning Zemin. (2016). Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics*, 14: 265-279.

66. Nanopore Technologies. Applications: Population Genomics. https://nanoporetech.com/applications/population-genomics (Accessed December 2017).

67. Nanopore Technologies. Protocols: Albacore. https://community.nanoporetech.com/protocols/albacore-offline-basecalli/v/abec_2003_v1_revh_29nov201. (Accessed March 2018).

68. Nanopore Technologies. Protocols: Metrichor. https://community.nanoporetech.com/protocols/epi2me/v/mte_1014_v1_revu_11apr2016. (Accessed March 2018).

69. Nanopore Technologies. Protocols: Local Basecalling. https://community.nanoporetech.com/protocols/local-basecalling-in-minknow/v/lbec_2001_v1_revi_15jul201. (Accessed March 2018).

70. Bioinformatics I/O. Exploring the FAST5 Format. http://bioinformatics.cvr.ac.uk/blog/exploring-the-fast5-format/. (Accessed March 2018).

71. Zhang Lab. What is FASTA Format. https://zhanglab.ccmb.med.umich.edu/FASTA/. (Accessed March 2018).

72. Github. Nanopolish. https://github.com/jts/nanopolish. (Accessed March 2018).

73. Poretools. https://github.com/arq5x/poretools. (Accessed March 2018).

74. Nanopolish. https://github.com/jts/nanopolish. (Accessed March 2018).

75. GRCh37. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/. (Accessed March 2018).

76. BWA. https://github.com/lh3/bwa/blob/master/bwa.1. (Accessed March 2018).

77. Heng, Li. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Broad Institute.

78. SAMtools. http://samtools.sourceforge.net/. (Accessed March 2018).

79. Samtools. http://www.htslib.org/doc/samtools.html. (Accessed March 2018).

80. Stats Output of SAM Tools. https://biostar.usegalaxy.org/p/21856/. (Accessed March 2018).

81. Integrative Genomics Viewer. http://software.broadinstitute.org/software/igv/. (Accessed January 2018).

82. Pakstis AJ, Speed WC, Kidd JR, *et al*. (2007). SNPs for Individual Identification. *Forensic Sci Int Genet*, 1(2008): 479-481.

83. Table Browser. https://genome.ucsc.edu/cgi-bin/hgTables. (Accessed December 2017).

84. Sims D, Sudbery I, Ilott NE, *et al*. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15: 121-132.

85. UCSC Genome Browser. https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr21%3A33031597-33041570&hgsid=664458397_r1CmdrqnHWXy8NGfvW6Jvip1frqY. (Accessed February 2018).

86. GenBank Overview. https://www.ncbi.nlm.nih.gov/genbank/. (Accessed October 2017).

87. Statistics and Population Genetics. https://strbase.nist.gov/pub_pres/NJSP2006_Statistics.pdf. (Accessed February 2018).

88. Faber-Hammond JJ and Brown KH. (2016). Pseudo-de novo assembly and analysis of unmapped genome sequence reads in wild zebrafish reveal novel gene content. *Zebrafish*, 13(2): 95-102.

89. Galaxy. https://usegalaxy.org/. (Accessed March 2018).

90. Nanopore Technologies. Publications. https://nanoporetech.com/publications?field_tags_target_id_1[123]=123&keys=. (Accessed March 2018).

91. Bainomugisa A, Duarte T, Lavu E, *et al*. (2018). A complete nanopore-only assembly of an XDR Mycobacterium tuberculosis Beijing lineage strain identifies novel geneticvariation in repetitive PE/PPE gene regions. BioRxiv.

92. Clark M, Wrzesinski T, Garcia-Bea A, *et al*. (2018). Long-read sequencing reveals the splicing profile of the calcium channel gene CACNA1C in human brain. BioRxiv.

93. McCabe M, Cormican P, Johnston D, *et al*. (2018). Simultaneous detection of DNA and RNA virus species involved in bovine repiratory disease by PCR-free rapid tagmentation-based library preparation and MinION nanopore sequencing. BioRxiv.

94. Hardegen J, Latorre-Perez A, Vilanova C, *et al*. (2018). Liquid co-substrates repower sewage microbiomes. BioRxiv.

95. Scientific Working Group on DNA Analysis Methods. Validation Guidelines for DNA Analysis Methods. https://docs.wixstatic.com/ugd/4344b0_813b241e8944497e99b9c45b163b76bd.pdf. (Accessed March 2018).

96. Nanopore Technologies. Protocols: Rapid Sequencing. https://community.nanoporetech.com/protocols/rapid-sequencing/v/res_9018_v7_revu_04jul2016 (Accessed October 2017).

97. Nanopore Technologies. Protocols: Rapid Sequencing. https://community.nanoporetech.com/protocols/rapid-sequencing-sqk-rad002/v/rse_9018_v2_rev_21nov2016. (Accessed October 2017).