Agency: National Institute of Justice

Award number: 2015-DN-BX-K068

Project Title: Development of PCR-Free DNA Typing Strategies via Real-Time High Throughput Sequencing

PI: John V. Planz, Ph.D.
Associate Professor
john.planz@unthsc.edu
817-735-2397

Submitting official: John V. Planz, Ph.D.
Associate Professor

Submission date: 09/11/2019

DUNS: 110091808

EIN: 1756064033A1

Recipient Organization: UNT Health Science Center
3500 Camp Bowie Blvd., Fort Worth, TX 76107

Award Period: 01/01/2016 to 12/31/2018

Reporting Period End Date: 12/31/2018

Final Summary Overview

ACCOMPLISHMENTS

1) What are the major goals of the project?

The purpose of this study was to empirically evaluate the effectiveness of an alternative approach to targeted library development coupled with a newly available direct read, real-time sequencing approach. A suite of loci including the expanded panels of auSTRs, Y-STRs, and full mtGenome served as the test bed for comparison to contemporary forensic typing methods. The focus of this project was to develop and critically evaluate a new testing process capable of concurrently typing well-characterized and accepted markers with high levels of confidence, increased allelic resolution, portability, and overall decreased costs. A goal of this project was to selectively mine the set of currently utilized forensic markers available in kits worldwide directly from DNA extracted from reference and evidentiary type samples and read the DNA sequence for the targeted loci from the enriched sample. Alternate methods were developed that capitalize on the sensitivity and flexibility of the sequencing platform to permit typing of low-level evidentiary DNA sources as well as reference samples with abundant DNA.

Our overall objective in this project is to develop a renewed approach to forensic DNA typing using both novel and well-characterized methods that have yet to be assembled in a unified process. The quantitative nature of this method allows us to evaluate the efficiency, implementation potential, and utility of the process for forensic DNA analysis. Developing this novel approach will provide the forensic community with the opportunity to explore a high-throughput, quantitative DNA typing method with the added benefit of scalability for field and/or laboratory applications with minimal equipment cost, simplified library preparation, and using direct-read sequencing that utilizes a laptop computer.

2) What was accomplished under these goals?

Over the course of the project two workflows were developed using ONT MinION sequencing to characterize forensically-relevant markers: mtDNA and STRs. The two marker systems, although they are markedly different in their data forms, follow almost identical workflows in the sequencing process. Mitochondrial DNA sequencing, to a large extent, is the more directly relatable to routine DNA sequencing in terms of the data produced.

One of the goals of this study were to determine the efficacy of sequencing full Human mitochondrial genomes utilizing the ONT MinION device and evaluate the accuracy of this DNA sequencing platform by direct comparison to the sequence of a NIST-traceable mtDNA sequencing standard to serve as a benchmark for expected results. Sequencing results obtained from native (non-enriched) genomic DNA samples (a major goal of this project) and sample libraries in which the mtGenomes were enriched through PCR are directly compared to evaluate the efficacy of both methods for clinical or forensic samples. Additionally, 25 blood and buccal swab samples ranging in age from <0.5 years to 12 years and two extracted DNA typing controls

were evaluated. Sequencing data was analyzed to assess depth of coverage necessary to obtain accurate consensus data. With the exception of HL60, which was used to generate an independent library, all of the enriched samples were pooled and barcoded. With six samples represented in a single library, the number of total reads per sample was less than 10,000. However, because the pool was made up of targeted mitochondrial amplicons, the majority of reads mapped to the reference genome, therefore generating a greater depth of coverage than that observed with the native samples. Average depth of coverage for barcoded, PCR enriched samples ranged from 295x-813x. The lowest coverage for a single position in any amplified sample was 54x. Each of these samples was also used to generate an independent, single-sample genome library without amplification to assess the efficacy or interrogating forensic markers directly form a genomic DNA sample. The total number of reads per sample was significantly higher, with the smallest result yielding 175,492 reads. However, only a few hundred reads from each actually mapped to the mtGenome due to the substantially greater concentration of nuclear fragments being interrogated by the nanopores. The average depth of coverage across the mtGenome ranged from 15-118x, with the lowest coverage for a single base in any given native sample being 8x. While this may seem "insufficient", the sequencing results for this native sample, sample 449, were equally informative/accurate as its counterpart generated from an amplicon-enriched library with an average coverage of 334x. The average number of differences between the amplified and the native consensus sequences across the samples tested was 102.66 differences across the ~16,569 base mtGenome. The libraries using the two approaches were 99.38% concordant

The differences observed between the two sets of sequence data were further assessed to determine the types of variations that were occurring during sequencing By and large, the majority of inconsistencies occurred within homopolymeric stretches, an issue common across sequencing platforms. In this study, 79 percent of the observed differences (727 of the 924 total) consisted of discrepancies between number of bases within homopolymeric stretches, defined here by a single base being repeated three or more times in a row. Unlike other platforms, however, there are no downstream errors in alignment induced beyond the homopolymeric region. As expected, there does not appear to be a high degree of error with dinucleotide repeats, e.g. when two different adjacent bases are repeated at least twice in tandem, (compromising only 3.2%, or 30, of all observed differences). This finding indicates that the incorporation of any variable signal, e.g. the presence a hetero-nucleotide polymer string within the k-mer passing through the nanopore, is sufficient to prevent homopolymeric errors to be induced through signal capture during strand translocation.

A single sample exhibited 51 of the 82 observed base disagreements, indicating that there was a combined total of only 31 base disagreements across the other samples sequenced. DNA

from this sample, along with several others, were extracted from frozen blood samples up to ten years old. DNA degradation is common in older stored samples, and often consists of cytosine bases undergoing deamination to uracil, which would be detected as a thymine in DNA sequencing that incorporates PCR enrichment or synthesis. It would be expected that the standard basecaller algorithms used with MinION data would recognize the uracil base as a thiamine as well, unless alternate basecalling rules were developed to specifically identify these base modifications, which incidentally are now being developed. Given the numerous C/T discrepancies randomly detected in both the PCR-enriched and native consensus sequences for sample 449, it is likely that our observations for this sample are originating from cytosine deamination. Being able to train the basecalling algorithm to detect these structural changes of degrading DNA molecules has potentially useful applications in forensic testing, where the age of a sample, or the provenance (e.g. ancient DNA vs contemporary) may be in question.

The NIST-traceable control sample HL60 was utilized to assess the accuracy of the ONT sequencing. Overall, the native sequence produced was concordant with the known HL60 sequence at 16,404 positions out of 16,569, resulting in an error rate of approximately 1.00%. The PCR-enriched consensus sequence for HL60 was concordant with the reference sequence at 16,418 of the 16,569 positions, indicating the error rate of the enriched samples was 0.91% (Table 2). This error rate is consistent with a rate of discordance (0.62%) between the amplified and native samples. The improved accuracy observed with the enriched sample is reflective of the much deeper coverage obtained through PCR and the elimination of pore-competition from nuclear DNA also in sequencing library. These findings are of particular importance considering the common misconception that the Nanopore sequencing system is highly error-prone. Improvements in the library preparation kits, flowcells and base recognition software observed over the course of this study have been successful in grossly reducing the error-frequency of this sequencing system. One feature that changed multiple times during project period was the iterations of basecalling algorithm provided by ONT. Using Albacore version 1.1, which included a homopolymeric correction algorithm, an additional 102 positions of the 165 miscalled bases identified initially from native genome sequencing were correctly called, decreasing the error rate from 1.00% to 0.38%. Similarly, 101 of the 151 errors identified within the amplicon-based HL60 consensus sequence were corrected, bringing the error rate of the system down from 0.91% to 0.30%. These results indicate that the new algorithm was effective in reducing errors, particularly in homopolymeric regions and was also successful in correctly identifying bases that were originally miscalled outside of polynucleotide regions. Improvements to the accuracy of the entire platform have been made throughout the project period that also positively affected the interrogation of STR loci, although the flux of the platform during this time created considerable challenges in assessing consistency in various aspects of the study.

To assess the efficacy of interrogating STR regions from the genomic DNA libraries generated during mtDNA analysis, a pilot project was performed targeting the well characterized forensically-informative identity SNPs developed by Kenneth Kidd's research group at Yale University. Although SNP interrogation was outside of the scope of the current project, it served as a direct, cost-free check to determine what depth of data could be expected for the forensic STR loci targeted for this study. Utilizing the previously generated sequence data, very poor results were obtained utilizing existing alignment and interrogation approaches that typically could be used on next-generation sequence data. The genomic DNA libraries prepared for the mtDNA analysis kit were generated using the ONT Rapid 1D sequencing kit, which through transposase-based tagmentation cuts and inserts sequencing adaptors at 2500-6000 bp intervals throughout the genome depending on sequence composition. These longer read lengths are often difficult to align to specific short target sequences such as SNPs or STRs due to the nature of the alignment algorithms that were developed for short read (150 bp) NGS data. Four alignment strategies were evaluated to mine for the SNP loci, however as a whole, the results indicated that from an individual sequencing run an insufficient coverage of target loci would prevent accurate calling of the intended SNP panel. ON average, 40 percent of the intended loci were detected in the generated data and of that only 10% of heterozygous loci were called correctly. These findings were significant in they did indicate that it was highly unlikely that STR loci could be reliably typed directly from the standard genomic library data.

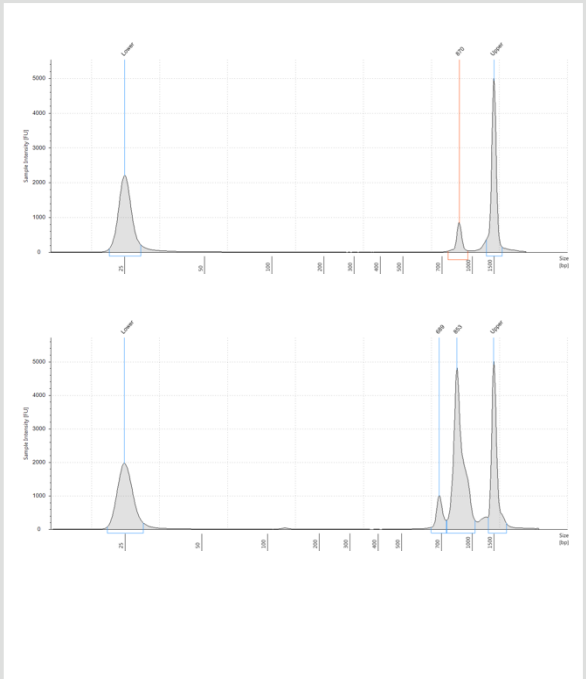
The typical recommendation for genomic DNA input into the ONT Rapid 1D sequencing library kit was at the time approximately 250 ng. Our libraries ranged from 144-259 ng. Although sufficient for generating sequence data, the ratio of content for the targeted loci (SNPs or STRs) is less than 0.1 percent in total. The vast amount of genomic and mtDNA competing with the target loci for pore access severely reduced the ability to generate useable profiles. The similar effect was observed in the mtDNA data discussed above, however the high enrichment of mtGenomes naturally occurring in a genomic DNA preparation of 250 ng has sufficient throughput to provide reliable results. Additionally, the contiguous nature of the mtGenome target, compared to SNPs or STR regions, can be efficiently and accurately aligned by existing softwares. These issues were expected in the original proposal of this project and several approaches to non-PCR based enrichment were discussed as potential options. Several, such as "flossing" (e.g. the ability to cycle the current at individual pores repetitively reading through a target region) while put forth as practical in discussions with ONT did not materialize from them once the project was underway.

A goal of this project was to evaluate the ability to sequence autosomal and Y STR markers of forensic interest using the MinION device. Our two MS students have made significant progress developing and optimizing a workflow for processing DNA extracts from STR

amplification through nanopore sequencing utilizing buccal swabs from twenty unrelated individuals, one control DNA sample, and three NIST-traceable standards. Following DNA extraction, the GlobalFiler and YFiler Plus PCR Amplification Kits were used to establish genotype and haplotype composition for the twenty buccal swab samples. Single contributor reference sample length-based and sequence data if available for the STRs interrogated were obtained from the respective manufacturers. This information was used to assess concordance between allele designations and determine the accuracy of nanopore sequencing through these regions of low complexity. The workflow for microsatellite sequencing on nanopore platforms was successfully established and the process of optimization was initialized.

Although several hundred copies of autosomal and Y-chromosome markers are present within native genomic DNA, these sequences represent a very small fraction of the total amount of genetic material in a given sample. Analysis of the genomic libraries generated for mtDNA analysis described previously suggested that STR copy number in unenriched libraries is insufficient for accurate genotype determination and identification of SNPs because target strands are unable to outcompete background gDNA for pore access during sequencing runs. Therefore, PCR amplification was performed prior to nanopore sequencing using customized oligonucleotide primers targeting 22 autosomal and 22 Y-chromosome common forensic STRs as well as Amelogenin (**Tables 1 & 2, all tables located at end of section**). The length of amplicons ranged from 790 to 850 bp with melting temperatures between 58.5 and 60.5°C. All target loci were successfully amplified in singleplex reactions, utilizing 0.5 ng input DNA (**Figure 1a**). Amplicons were merged by sample and purified to remove remaining primers and PCR reaction component. Inclusion of this additional step enabled us to eliminate all non-target products and residual primer/primer-dimer content from the amplification process yielding high quality target amplicons of sufficient concentration in 100 percent of the loci. Multiplex PCR reactions for the autosomal and Y STR panels were also successfully created in this project and optimized for 1.0 ng of template (**Figure 1b**). Following amplification, individual multiplexes were barcoded and sequenced on the MinION device in order to ensure all target loci were successfully amplified. The raw read count obtained from these sequencing runs indicates that each locus within the three autosomal and four Y multiplex panels were adequately represented (**Table 3 & 4**).

a.



The results obtained herein demonstrate that amplification prior to library preparation ensures target loci outcompete inherent background DNA for pore access (**Figure 2**) but also generates a considerable number of PCR-induced artifacts that complicate data interpretation (**Figure 3**). TaKaRa Long and Accurate (LA) Taq was utilized in all amplification reactions. Stutter artifacts, however, were still incorporated

The deep coverage achieved from thirty PCR cycles suggests that reliable auSTR and Y-STR typing results can be ascertained from a considerably lower depth using the MinION device. Although enrichment prior to nanopore sequencing is necessary, the use of PCR at thirty cycles is not ideal for the for the STR loci targeted in this project and given that the extremely high level of coverage obtained may be unnecessary. Future studies will evaluate the effect of reducing PCR cycling on stutter artifact formation while retaining sufficient coverage to accurately call STR loci. Optimization of the enrichment process will not only alleviate interpretational difficulties during subsequent data analyses, but also increase current knowledge of the stutter generation process and sensitivity of nanopore sequencing devices. Enrichment through RNA baiting was assessed early in this study, however the captured targets were not amenable to ligation to the ONT adaptors. This was described in an earlier progress report. Recently modifications developed using this process have been developed for a separate study which may prove beneficial for STR and mtDNA enrichment.

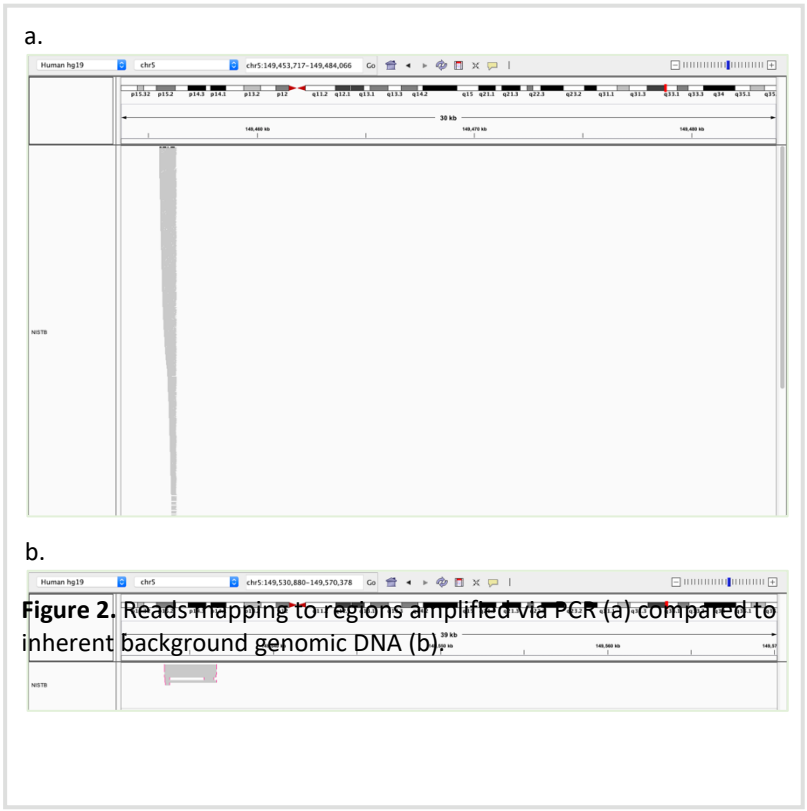


Figure 2. Reads mapping to regions amplified via PCR (a) compared to inherent background genomic DNA (b).

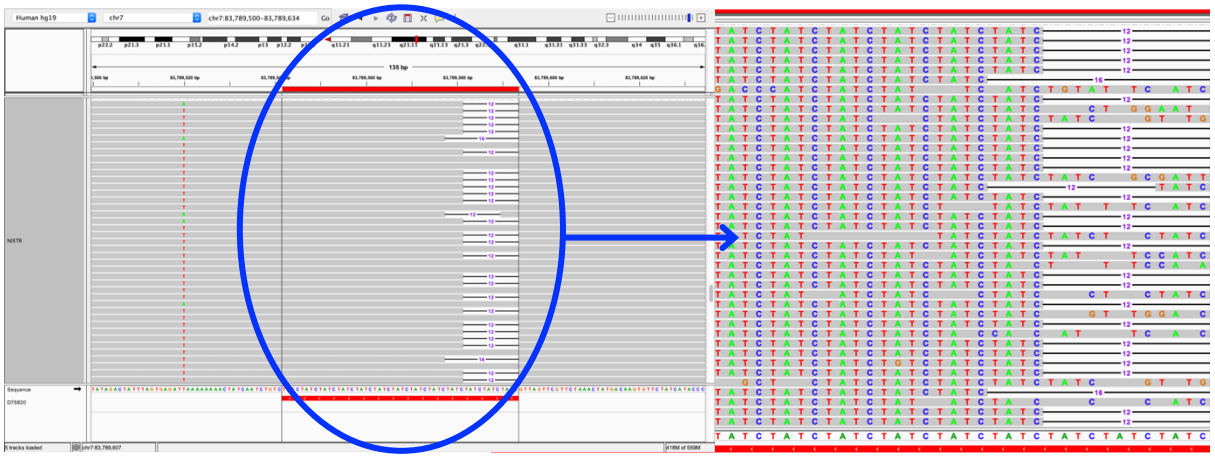


Figure 3. The 12 bp deletion observed (black line) is consistent with the 10 allele of NIST Component B given that the reference sequence is 3 repeat units (12bp) longer. The 16 bp deletions indicates PCR-induced minus stutter.

Some challenges encountered throughout the course of this project stemmed directly from the rapid evolution of the library preparation reagents, flow cell sensors, and base recognition software. Techniques developed by previous students in our laboratory to improve DNA recovery during library preparation actually complicated processing of singleplex amplification of autosomal loci for three samples. During this library preparation, barcode ligation was terminated by a 10-minute incubation at 65°C, thereby eliminating a previously required Agencourt AMPure XP bead cleanup known to result in significant DNA loss. The pooled and barcoded samples were then concentrated using a Microcon DNA Fast Flow Filter Device and eluted to a volume of 50 μL. Quantification of the pooled and barcoded samples on the Qubit 2.0 fluorometer using the Qubit dsDNA BR Assay Kit indicated roughly 65% sample loss. The results obtained were confirmed on a D1000 ScreenTape using the Agilent TapeStation 4200. In order to recover DNA presumably retained by the filter, nuclease-free water was heated to 98°C, applied directly to the membrane, and incubated at room temperature for 10 minutes. Following a brief vortex at full speed, the filter device was inverted into a clean collection tube and centrifuged at 1000 ×g for 3 minutes. Comparison of the results obtained before and after attempted recovery suggests that the barcoded samples were, in fact, trapped within the membrane. Subsequently, bead purification was performed as per manufacturer's protocol in all subsequent library preparations with an additional 2.5 × wash after pooling 700 ng of the barcoded samples to be sequenced in a single run. To decrease loss during these steps, the magnetic force of the DynaMag-2 was increased by addition of circular magnets at the point of tube contact. Quantification checkpoints were performed on a D1000 ScreenTape using the Agilent TapeStation 4200 and sample input was based on amplicon rather than total DNA concentration from Qubit readings.

The continuous evolution of kit reagents, as well as in the hardware and software utilized, resulted in notable differences between sequencing runs, complicating comparison of the data generated throughout this project. For instance, Revision C flow cells (FLO-MIN106) were replaced by Revision D (Rev D) flow cells

Because the libraries were prepared, normalized, and sequenced in the exact same manner, the disproportionally high number of reads obtained were likely due to inherent differences between the recently released flow cells utilized. In addition to flow cell issues, Native Barcode 03 (BC03) was not performing as expected. The relatively low number of reads for all samples barcoded with BC03 (**Figure 4**) from the same kit suggests that the issue was caused by the barcode, itself, rather than DNA loss or other human errors during library preparation. ONT Customer Support confirmed failure of NB03 after reviewing the sequencing run information. Although intended to reduce the amount of DNA required for sequencing and improve quality of the resultant data, unforeseen issues in the updates released can impact sequencing and even result in unrecoverable data.

Concordance between allele designations generated using traditional STR typing techniques and the longer-read nanopore sequencing data were assessed via a customized bioinformatics pipeline developed in collaboration with Dr. Sedlazeck. The longer-read data generated allowed for accurate alignment and identification of single nucleotide polymorphisms (SNPs) within and around microsatellite loci. In addition to length-based genotype or haplotype determinations and SNP identification, the accuracy of nanopore sequencing through these highly repetitive microsatellite regions was determined using Components A, B, and C of NIST Standard Reference Material 2391c. Through continued collaboration with Dr. Sedlazeck the data analysis pipeline developed will be further streamlined for

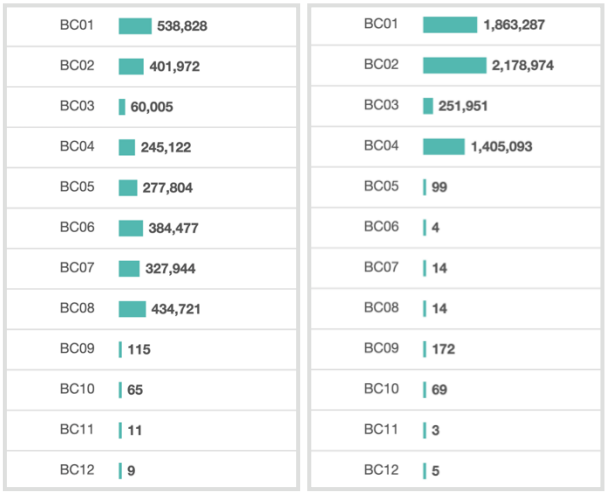


Figure 4. Resultant reads separated by barcode using EPI2ME for two separate nanopore sequencing runs. Note the significantly lower number of reads for BC03 relative to the barcodes utilized.

application in our continuing research in nanopore sequencing applications. Furthermore, we will be evaluating strategies of sequencing STRs with minimal enrichment to reduce the confounding effects of STR stutter and aid in generating sufficient read counts for accurate assessment of genotypes. Once the working protocols are fully resolved for the STR approaches, we will conduct initial sensitivity assays for the STR libraries (autosomal and Y) to assess the efficiency of different approaches in extracting and retaining the forensically informative loci we are targeting.

Although optimization of the sequence analysis pipeline is still in progress, the nanopore sequencing results we obtained for the STR loci have been robust and concordant with length-based electrophoretic results for the NIST standards we utilized for calibration and quality assessment **(Tables 5-9)**. In all cases the dominant read(s) observed at each locus corresponded to the expected allele(s) in the sample. Stutter (plus and minus 1 and 2 by repeat motif definition) were observed as expected, however given the 30 amplification cycles utilized in the PCR (comparable to commercial CE kits) our initial analyses indicate a generally lower percentage of stutter than typically reported for CE and MPS-based evaluation of these same loci. It is possible that the use of longer amplicons in this study may have some stabilizing effect on the polymerase. We are continuing to further refine the analysis of allele calling software and these observations may change once the analytical pipeline is optimized for these data and we can determine ratios more accurately. We are also currently working with the ONT $1D^2$ chemistry in which the leading strand and its synthesized complement are linked in analysis providing confirmatory sequence data similar to the original 2D chemistry. One result of stutter events during synthesis is removal of paired reads to the failed folders when the complementary strands are in disagreement. These failed reads are still accessible but do not contribute to the standard strand alignment, however we are harnessing this capability to explore the mechanisms behind microsatellite expansion/contraction in current projects.

The data generated during this project is still under interrogation with alternate analytical methods and will be we more fully described in publications under development for submission mid-2019. The technology is evolving rapidly, and this has its disadvantages when trying to optimize or validate a system. At present the ONT technologies and analytical pipelines are not ready for classical forensic validation studies, however the work initiated in this study sets the baseline for further optimizations which have been suggested to ONT and acted upon as we observed with homopolymer correction capabilities in the mtDNA workflows. This technology has the ability to produce accurate sequence-based STR allele results comparable to current 2^{nd} generation sequencing technologies, however the manner in which the process is performed, and data interrogated, are radically different. The suite of loci examined can be easily expanded and data forms, such as methylation state, can be assessed during standard sequencing, unlike any other platform. As such, nanopore sequencing can evolve in step with discovery.

Table 1. **Designed primer sequences and amplicon lengths for 22 autosomal STRs and Amelogenin.**

Locus	Orientation	Primer sequence (5′ to 3′)	Amplicon length
TPOX	Forward Reverse	CATGCCTGCACACACCTAC CGCTCAAACGTGAGGTTGAC	800
D2S1338	Forward Reverse	ACAACAGAATATGGGTTCTTGCG CTGTGCTATGGAAAAGCCGTG	801
FGA	Forward Reverse	TCGTTTCATATCAACCAACTGAGC TAGGAAACATTGCATTCACTCTGG	795
CSF1PO	Forward Reverse	GCCACCTCCATCTCCCATAAC GAGGAACATATGCAAGGCTCAAAG	799
D8S1179	Forward Reverse	ATCCAAAGTGGATCCTCAGCC CCGGCCCAAGTTTTTATTTTCC	794
TH01	Forward Reverse	TCAGCTTCATCCTGAGCTTCC TTGAATCTTAACGATCGGAATGTGG	806
D12S391	Forward Reverse	GAATATATGAAGTCGCCTGTAATCCC CTCCTGGAGCTGCGACAC	795
Penta E	Forward Reverse	CTCAATGCCAACATTACAGGGC GCTTAAAGTTGACGTCTCATTGC	801
D18S51	Forward Reverse	CTAACAATAGGCCAAGCGTGATG ATTGAGTCAGGTAACATTTATGCCC	813
D21S11	Forward Reverse	CAAATTTCCCCTCTCACTTCTGG AGAATTAGAGAGCTGTGTCGAGG	810
D22S1045	Forward Reverse	GAGCAGCTTAAGGCATCCTG GCTTCTCACCGTTGCATTAGG	820

Table 2. Designed primer sequences and amplicon lengths for 22 Y-STRs.

Locus	Orientation	Primer sequence (5' to 3')	Amplicon length
DYS19	Forward	TGTCATAGCTCTGAACATTTGGTTG	825
	Reverse	AGTTTCATGAAGCATTCTCTACAGT	
DYS385a/b	Forward	CAAGTTCTGTGTGCATCATTTTCTC	791/803
	Reverse	TTGAACCTGAAATGTAAAGGGTGTC	
DYS389I/II	Forward	ACCAGATGCGAGAACACCC	796
	Reverse	CCCCAGATACAGATGCTGCC	
DYS390	Forward	ACAAATAAAACCTACCTGCAGCC	806
	Reverse	TACTGGAACACCTGTTGGAGAG	
DYS392	Forward	CTCCCTGTGAGAATGAATTGTTCC	800
	Reverse	ATTTGGGGTAGTGCTTAGTCCC	
DYS391	Forward	ACCACAGATTAGCATTCATCTTGC	844
	Reverse	ACCAGATGCCAGCAGTATCC	
DYS393	Forward	CCCCTACACAAGTTCTCCTGC	799
	Reverse	GTGGTGGTGCATGCTGTTAA	
DYS435	Forward	TGCAGTGCTGCATAATAAATGACC	807
	Reverse	CAGCCCACCAGATATCTCTATCTAC	
DYS437	Forward	AATCCATCCATTCATTCATCTCCCT	838
	Reverse	CTGAGGTTGGAATTGCTTGAGTC	
DYS438	Forward	ACCATTTTCTGATGAAAAGGAACCAG	802
	Reverse	TACCTGGCTGCTCCCCTAG	
DYS439	Forward	GTACTTCCTAGGTTTTCTTCTCGAGT	839
	Reverse	TCTGGTGTCCTCTTCCAACTTG	
DYS448	Forward	GGAGGATATGTCAAAGGATTCAAGG	796
	Reverse	TTCTTCCCTTCTCTGACATTTTTGG	
DYS456	Forward	TTGCTTCCACAAGATTTCACACTG	795
	Reverse	TGGGGTCTCATTATGTTGTTTATGC	
DYS458	Forward	CTTTAACAGTTTGGGACAGCTGAG	816
	Reverse	CTTAAAAAGTTTCCCCATGTTGGTG	
DYS481	Forward	AAGTCTGGCATCCAAACCTGC	904
	Reverse	TCTCCTTGGTTTGGCCATAGG	
DYS533	Forward	TCTCAAACCCCTACTCATTCAAAAC	802
	Reverse	AGCTTTTGGAAACACCAGTTAGAG	
DYS549	Forward	CAATGAACCCCACCACAAGC	808
	Reverse	TGTGGATTAGTTTTTGACGTGCTAC	
DYS570	Forward	TCCTTGGAATTGCAACTTGGC	845
	Reverse	TCCATCTCAGAATCAAGAAGGGC	
DYS576	Forward	AAAGGTGAAAATGTGCCTTCCC	795
	Reverse	TACCTTTGGGTGTGTATGTTAGTCC	
DYS643	Forward	CCAACTGAGTGGATATTTCCTTGC	827
	Reverse	ACGTTGCCTGTTAACTCTGATAATAC	
GATAH4	Forward	CTATGCCCAGAATATTGCATTAGGC	841
	Reverse	AATGTCAAGCCTTCTGATTGCC	

Table 3. **Number of reads mapping to the autosomal STR loci amplified via multiplex PCR. Shading indicates primer pairs in each panel.**

Locus	Multiplex 1	Multiplex 2	Multiplex 3
D1S1656	12672	9	40
TPOX	9683	8	30
D2S441	5767	15	52
D2S1338	8552	3	31
D3S1358	11078	13	34
FGA	1003	3	5
D5S818	11499	13	26
CSF1PO	10230	12	36
D7S820	10857	19	32
D8S1179	7	13512	52
D10S1248	9	18954	40
TH01	3	21600	40
vWA	1	30744	36
D12S391	3	20420	44
D13S317	5	34676	37
Penta E	13	20997	59
D16S539	14	32490	46
D18S51	6	6	17076
D19S433	6	14	19910
D21S11	4	9	22950
Penta D	5	14	10735
D22S1045	1	4	5279
AMEL X	2	2	9112
AMEL Y	1	3	8371

Table 4. **Number of reads mapping to the Y STR loci amplified via multiplex PCR. Shading indicates primer pairs in each panel.**

Locus	Multiplex 1	Multiplex 2	Multiplex 3	Multiplex 4
DYS19	28718	2	2	5
DYS438	77050	2	4	22
DYS448	38794	10	5	13
DYS456	58998	4	5	18
DYS458	25727	3	5	4
DYS385a	6	7499	0	7
DYS385b	1	8867	4	7
DYS391	13	25829	0	27
DYS393	12	35715	8	26
DYS549	8	28408	3	20
DYS533	8	10410	6	15
DYS392	21	7	49299	18
DYS435	–	–	–	–
DYS439	11	7	22135	7
DYS570	17	8	50518	13
DYS643	13	9	38551	19
DYS389I	15	4	11	41755
DYS389II	15	4	11	41994
DYS390	6	7	5	13636
DYS437	9	3	7	1128
DYS576	7	2	11	25727
GATAH4	8	3	4	34181

Table 5. NIST Component A length-based genotypes and sequencing data for autosomal STR loci interrogated in this project.

Locus	Genotype	Repeat sequence	
		Allele 1	Allele 2
D1S1656	17.3, 17.3	[TAGA]$_4$ TGA [TAGA]$_{12}$ TAGG **[TG]$_5$**	[TAGA]$_4$ TGA [TAGA]$_{12}$ TAGG
TPOX	8, 8	[AATG]$_8$	[AATG]$_8$
D2S441	10, 10	[TCTA]$_{10}$	[TCTA]$_8$ TCTG [TCTA]$_1$
D2S1338	18, 23	[TGCC]$_6$ [TTCC]$_{12}$	[TGCC]$_7$ [TTCC]$_{13}$ GTCC [TTCC]$_2$
D3S1358	15, 16	TCTA [TCTG]$_2$ [TCTA]$_{12}$	TCTA [TCTG]$_3$ [TCTA]$_{12}$
FGA	21, 23	[TTTC]$_3$TTTT TTCT [CTTT]$_{13}$ CTCC [TTCC]$_2$	[TTTC]$_3$TTTT TTCT [CTTT]$_{15}$ CTCC [TTCC]$_2$
D5S818	11, 12	[AGAT]$_{11}$	[AGAT]$_{12}$
CSF1PO	10, 10	[AGAT]$_{10}$	[AGAT]$_{10}$
D7S820	11, 11	[GATA]$_{11}$	[GATA]$_{11}$
D8S1179	13, 14	[TCTA]$_{13}$	[TCTA]$_2$ TCTG [TCTA]$_{11}$
D10S1248	15, 16	[GGAA]$_{15}$	[GGAA]$_{16}$
TH01	8, 9.3	[AATG]$_8$	[AATG]$_6$ ATG [AATG]$_3$
vWA	18, 19	TCTA [TCTG]$_4$ [TCTA]$_{13}$	TCTA [TCTG]$_4$[TCTA]$_{13}$
D12S391	18.3, 22	AGAT GAT [AGAT]$_9$ [AGAC]$_7$ AGAT	[AGAT]$_{13}$ [AGAC]$_8$ AGAT
D13S317	8, 8	[TATC]$_8$	[TATC]$_8$
Penta E	5, 10	[AAAGA]$_5$	[AAAGA]$_{10}$
D16S539	10, 11	[GATA]$_{10}$	[GATA]$_{10}$
D18S51	12, 15	[AGAA]$_{12}$	[AGAA]$_{15}$
D19S443	13, 14	[AAGG] **AAAG** [AAGG] **TAGG** [AAGG]$_{11}$	[AAGG] **AAAG** [AAGG] **TAGG** [AAGG]$_{12}$
D21S11	28, 32.3	[TCTA]$_4$ [TCTG]$_6$ {[TCTA]$_3$ **TA** [TCTA]$_3$ **TCA** [TCTA]$_2$ **TCCATA**} [TCTA]$_{10}$	[TCTA]$_5$ [TCTG]$_6$ {[TCTA]$_3$ **TA** [TCTA]$_3$ **TCA** [TCTA]$_2$ **TCCATA**} [TCTA]$_{12}$ TA TCTA
Penta D	9, 13	[AAAGA]$_9$	[AAAGA]$_{13}$
D22S1045	15, 15	[ATT]$_{12}$ ACT [ATT]$_2$	[ATT]$_{12}$ ACT [ATT]$_2$

Note: Bases in **bold** do not contribute to length-based genotype determinations.

Table 6. NIST Component B length-based genotypes and sequencing data for autosomal STR loci interrogated in this project.

Locus	Genotype	Repeat sequence	
		Allele 1	Allele 2
D1S1656	11, 14	[TAGA]$_{11}$ **[TG]$_5$**	[TAGA]$_{14}$ **[TG]$_5$**
TPOX	8, 11	[AATG]$_8$	[AATG]$_{11}$
D2S441	10, 14	[TCTA]$_{10}$	[TCTA]$_{11}$ TCTG [TCTA]$_2$
D2S1338	17, 17	[TGCC]$_6$ [TTCC]$_{11}$	[TGCC]$_6$ [TTCC]$_{11}$
D3S1358	15, 19	TCTA [TCTG]$_3$ [TCTA]$_{11}$	TCTA [TCTG]$_3$ [TCTA]$_{15}$
FGA	20, 23	[TTTC]$_3$TTTT TTCT [CTTT]$_{12}$ CTCC [TTCC]$_2$	[TTTC]$_3$TTTT TTCT [CTTT]$_{15}$ CTCC [TTCC]$_2$
D5S818	12, 13	[AGAT]$_{12}$	[AGAT]$_{13}$
CSF1PO	10, 11	[AGAT]$_{10}$	[AGAT]$_{11}$
D7S820	10, 10	[GATA]$_{10}$	[GATA]$_{10}$
D8S1179	10, 13	[TCTA]$_{10}$	[TCTA]$_{13}$
D10S1248	13, 13	[GGAA]$_{13}$	[GGAA]$_{13}$
TH01	6, 9.3	[AATG]$_6$	[AATG]$_6$ ATG [AATG]$_3$
vWA	17, 18	TCTA [TCTG]$_4$ [TCTA]$_{12}$	TCTA [TCTG]$_4$[TCTA]$_{13}$
D12S391	19, 24	[AGAT]$_{12}$ [AGAC]$_6$ AGAT	[AGAT]$_{15}$ [AGAC]$_9$
D13S317	9, 12	[TATC]$_9$	[TATC]$_{12}$
Penta E	7, 15	[AAAGA]$_7$	[AAAGA]$_{15}$
D16S539	10, 13	[GATA]$_{10}$	[GATA]$_{13}$
D18S51	13, 16	[AGAA]$_{13}$	[AGAA]$_{16}$
D19S443	16, 16.2	[AAGG] **AAAG** [AAGG] **TAGG** [AAGG]$_{14}$	[AAGG] **AA--** [AAGG] **TAGG** [AAGG]$_{15}$*****
D21S11	32, 32.2	[TCTA]$_4$ [TCTG]$_6$ {[TCTA]$_3$ **TA** [TCTA]$_3$ **TCA** [TCTA]$_2$ **TCCATA**} [TCTA]$_{14}$	[TCTA]$_5$ [TCTG]$_6$ {[TCTA]$_3$ **TA** [TCTA]$_3$ **TCA** [TCTA]$_2$ **TCCATA**} [TCTA]$_{12}$ TA TCTA
Penta D	8, 12	[AAAGA]$_8$	[AAAGA]$_{12}$
D22S1045	15, 17	[ATT]$_{12}$ ACT [ATT]$_2$	[ATT]$_{14}$ ACT [ATT]$_2$

Note: Bases in **bold** do not contribute to length-based genotype determinations. *16.2 allele results from 2 base pair deletion in uncounted repeat unit.

Table 7. NIST Component B length-based haplotypes and sequencing data for Y-STR loci interrogated in this project.

Locus	Haplotype	Repeat sequence
DYS19	14	[TAGA]$_3$ **TAGG** [TAGA]$_{11}$
DYS385a	13	[GAAA]$_{13}$
DYS385b	17	[GAAA]$_{17}$
DYS389I	13	[TCTG]$_3$ [TCTA]$_{10}$
DYS389II	31	[TCTG]$_6$ [TCTA]$_{12}$ **N$_{48}$** [TCTG]$_3$ [TCTA]$_{10}$
DYS390	23	[TCTG]$_8$ [TCTA]$_{10}$ TCTG [TCTA]$_4$
DYS391	10	[TCTA]$_{10}$
DYS392	11	[TAT]$_{11}$
DYS393	12	[AGAT]$_{12}$
DYS437	14	TCTA]$_8$ [TCTG]$_2$ [TCTA]$_4$
DYS438	10	[TTTTC]$_{10}$
DYS439	11	[AGAT]$_{11}$
DYS448	20	[AGAGAT]$_{12}$ **N$_{42}$** [AGAGAT]$_8$
DYS456	15	[AGAT]$_{15}$
DYS458	17.2	[GAAA]$_{15}$ AA [GAAA]$_2$
DYS481	25	[CTT]$_{25}$
DYS533	11	[ATCT]$_{11}$
DYS549	12	[GATA]$_{12}$
DYS570	18	[TTTC]$_{18}$
DYS576	17	[AAAG]$_{17}$
DYS643	9	[CTTTT]$_9$
GATAH4	11	[TAGA]$_{11}$

Note: Bases in **bold** do not contribute to length-based genotype designations. Haplotype and repeat sequence data for DYS435 are not available.

Table 8. NIST Component C length-based genotypes and sequencing data for autosomal STR loci interrogated in this project.

Locus	Genotype	Repeat sequence	
		Allele 1	Allele 2
D1S1656	11, 15	[TAGA]$_{11}$ **[TG]$_5$**	[TAGA]$_4$ TGA [TAGA]$_{12}$ TAGG **[TG]$_5$**
TPOX	11, 11	[AATG]$_{11}$	[AATG]$_{11}$
D2S441	10, 10	[TCTA]$_8$ TCTG [TCTA]$_1$	[TCTA]$_8$ TCTG [TCTA]$_1$
D2S1338	19, 19	[TGCC]$_7$ [TTCC]$_{12}$	[TGCC]$_7$ [TTCC]$_{12}$
D3S1358	16, 18	TCTA [TCTG]$_3$ [TCTA]$_{12}$	TCTA [TCTG]$_3$ [TCTA]$_{14}$
FGA	24, 26	[TTTC]$_3$TTTT TTCT [CTTT]$_{16}$ CTCC [TTCC]$_2$	[TTTC]$_3$TTTT TTCT [CTTT]$_{18}$ CTCC [TTCC]$_2$
D5S818	10, 11	[AGAT]$_{10}$	[AGAT]$_{11}$
CSF1PO	10, 12	[AGAT]$_{10}$	[AGAT]$_{12}$
D7S820	10, 12	[GATA]$_{10}$	[GATA]$_{12}$
D8S1179	10, 17	[TCTA]$_{10}$	[TCTA]$_2$ TCTG [TCTA]$_{14}$
D10S1248	12, 16	[GGAA]$_{12}$	[GGAA]$_{16}$
TH01	6, 8	[AATG]$_6$	[AATG]$_8$
vWA	16, 18	TCTA [TCTG]$_4$ [TCTA]$_{11}$	TCTA [TCTG]$_4$[TCTA]$_{13}$
D12S391	19, 23	[AGAT]$_{13}$ [AGAC]$_5$ AGAT	[AGAT]$_{12}$ [AGAC]$_{11}$
D13S317	11, 11	[TATC]$_{11}$$^+$	[TATC]$_{11}$$^+$
Penta E	12, 13	[AAAGA]$_{12}$	[AAAGA]$_{13}$
D16S539	10, 10	[GATA]$_{10}$	[GATA]$_{10}$
D18S51	16, 19	[AGAA]$_{16}$	[AGAA]$_{19}$
D19S443	13.2, 15.2	[AAGG] **AA--** [AAGG] **TAGG** [AAGG]$_{12}$*	[AAGG] **AA--** [AAGG] **TAGG** [AAGG]$_{14}$*
D21S11	29, 30	[TCTA]$_4$ [TCTG]$_6$ {[TCTA]$_3$ **TA** [TCTA]$_3$ **TCA** [TCTA]$_2$ **TCCATA**} [TCTA]$_{11}$	[TCTA]$_6$ [TCTG]$_5$ {[TCTA]$_3$ **TA** [TCTA]$_3$ **TCA** [TCTA]$_2$ **TCCATA**} [TCTA]$_{11}$ TA TCTA
Penta D	10, 11	[AAAGA]$_{10}$	[AAAGA]$_{11}$
D22S1045	16, 16	[ATT]$_{13}$ ACT [ATT]$_2$	[ATT]$_{13}$ ACT [ATT]$_2$

Note: Bases in **bold** do not contribute to length-based genotype designations. *13.2 and 15.2 alleles result from 2 base pairs deletion in uncounted repeat unit. $^+$A to T single nucleotide polymorphism located 1 base pair downstream from repeat unit.

Table 9. NIST Component C length-based haplotypes and sequencing data for Y-STR loci interrogated in this project.

Locus	Haplotype	Repeat sequence
DYS19	15	$[TAGA]_3$ **TAGG** $[TAGA]_{12}$
DYS385a	13	$[GAAA]_{13}$
DYS385b	15	$[GAAA]_{15}$
DYS389I	12	$[TCTG]_3$ $[TCTA]_9$
DYS389II	27	$[TCTG]_5$ $[TCTA]_{10}$ **N_{48}** $[TCTG]_3$ $[TCTA]_9$
DYS390	24	$[TCTG]_8$ $[TCTA]_{11}$ TCTG $[TCTA]_4$
DYS391	11	$[TCTA]_{11}$
DYS392	13	$[TAT]_{13}$
DYS393	13	$[AGAT]_{13}$
DYS437	16	$[TCTA]_{10}$ $[TCTG]_2$ $[TCTA]_4$
DYS438	11	$[TTTTC]_{11}$
DYS439	12	$[AGAT]_{12}$
DYS448	19	$[AGAGAT]_{11}$ **N_{42}** $[AGAGAT]_8$
DYS481	26	$[CTT]_{26}$
DYS456	15	$[AGAT]_{15}$
DYS458	17	$[GAAA]_{17}$
DYS533	10	$[ATCT]_{10}$
DYS549	13	$[GATA]_{13}$
DYS570	20	$[TTTC]_{20}$
DYS576	16	$[AAAG]_{16}$
DYS643	12	$[CTTTT]_{12}$
GATAH4	11	$[TAGA]_{11}$

Note: Bases in **bold** do not contribute to length-based genotype designations. Haplotype and repeat sequence data for DYS435 are not available.

3) What opportunities for training and professional development has the project provided?

The project has provided professional training one post-doctoral fellow, four Masters students from our Forensic Genetics graduate program, one STEM science teacher and in May-June 2017 we hosted a SMART program scholar out of the Forensic Biology program from Saint Mary's University in San Antonio, TX. These individuals were trained in PCR development for STRs and mtDNA, DNA library preparation, Nanopore sequencing methods, and data analysis through meetings with the PI, conference calls with other Nanopore researchers and webinars and conference calls with Oxford Nanopore. One Masters Student, Kelcie Thorson graduated in May 2017 and the results of her work on enrichment-free mtDNA sequencing produced a publication [Zascavage et al 2018]. Two additional Masters students (Courtney Hall and Kiana Valenti) joined the project team in Fall 2017 focusing on STR analysis and adaptations of the software tools to enhance efficient evaluation of STR sequence data. A manuscript is currently in preparation describing that component of the project [Hall et al., in preparation]. Additionally, another Forensic Genetics Masters student (Whitney Ludwick) interrogated the native genome sequence datasets produced in this project (from the amplification -free mtGenome analysis described previously) assessing efficacy of directly typing SNPs form the unamplified libraries human identity SNPs (based on the NIJ-funded research of Dr. Kenneth Kidd's laboratory). The team has learned a great deal about the computer programming associated with data analysis for the Nanopore data as commercially optimized analysis suites for this technology are not available. Through travel to England, New York and San Francisco to participate in training and research conferences we have been able to collaborate with the leaders in Nanopore sequencing based research who have contributed greatly to our progress in analytical approaches.

4) How have the results been disseminated to communities of interest?

The approaches and progress on our forensic based approaches to date have been informally discussed with collaborators and researchers through the Nanopore yearly meetings. Dr. Planz has presented on the mtDNA portion of the research outcomes from this project at the Green Mountain Conference in Burlington, VT in July 2017. Dr. Zascavage presented an overview of our progress on this project at the Nanopore Community Meeting, San Francisco, CA November 2018. The manuscript reported on previously entitled Nanopore sequencing: An enrichment-free alternative to mitochondrial DNA sequencing was published in *Electrophoresis* 2019, 40:272-280. We are currently preparing manuscripts encompassing the STR analyses for submission Summer 2019 to Forensic Science Review, Forensic Science International Genetics, Current Protocols in Genetics, and PLoS Genetics.

5) What do you plan to do during the next reporting period to accomplish the goals?

The project ended December 31, 2018. No further work is being performed on this project except the preparation of manuscripts resulting from the work. The process development, data,

and several team members working on this project have transitioned to addressing the goals of next phase of nanopore evaluation under award NIJ 2018-DU-BX-0179.

PRODUCTS

Hall, C.L., R.R. Zascavage, F.J. Sedlazeck, J.V. Planz. Sequencing long amplicon microsatellite loci using the Oxford Nanopore Technologies MinION™ device. RAD Conference, Fort Worth, TX March 2019.

Zascavage, R.R., K. Thorsen, J.V. Planz. 2018. Nanopore Sequencing: A PCR-free alternative to mitochondrial DNA sequencing. Electrophoresis, 40: 272-280. doi:10.1002/elps.201800083

Zascavage, R. R., F.J. Sedlazeck, J.V. Planz. 2018. Applications of Nanopore Sequencing for Forensic Analysis. Nanopore Community Meeting, New York, NY November 30, 2018. *Invited Speaker*

Planz, J.V. and Zascavage, R.R. 2017. The potential of PCR-free Nanopore sequencing for forensic use. Green Mountain Conference, Burlington VT August 25, 2017 *Invited Speaker*

Thorson, Kelcie. 2017. Approaches to mitochondrial genome sequencing using the Oxford Nanopore MinION device. Master of Science (Biomedical Sciences, Forensic Genetics). April 2017, 42 pages, 7 tables, 29 figures, references.

Thorsen, K., Zascavage, R.R., J.V. Planz. 2017. PCR-free mitochondrial genome sequencing using the Oxford Nanopore MinION device. Research Appreciation Day Poster, April 15, 2017 UNTHSC.

Zascavage, R.R. 2016. Forensic applications of nanopore sequencing. Nanopore Pore Camp. Exeter, UK.

IMPACT

Although we have no formal way of assessing the present impact of this project, considerable interests from individuals in the forensic DNA discipline has been shown and the PI has been asked to provide an update on the ONT sequencing research conducted under this project and others in August 2019 at the Green Mountain Conference, as a follow-up to the presentation given there in 2017. Awareness of the technology is spreading in the forensic community in the US and abroad, however, there are to date only 6 forensic related publications including MinION sequencing. The technology in its present form is not in a state of completion to justify forensic developmental or internal validation studies and until additional fundamental research is completed on the platform it would be folly to undertake that level of evaluation.