ABSTRACT

A common biomarker of damaged DNA, particularly mitochondrial DNA, 8-oxoguanine (8-oxoG) has been identified as a possible contributor to neurodegenerative disorders, such as Alzheimer's disease, Parkinson's disease, preeclampsia, as well as type 1 and type 2 diabetes. Numerous methods have been developed to detect oxidative damage within the genome, including but not limited to immunological techniques, quantitative-polymerase chain reaction (qPCR), and *in situ* imaging. This study explores nanopore sequencing using the MinION Nanopore (Oxford Nanopore Technologies, Oxford, UK) as a more sensitive method of 8-oxoguanine detection, providing proof-of-concept for model training as well as preliminary model development.

DETECTING AND QUANTIFYING OXIDATIVE DNA DAMAGE USING MINION NANOPORE 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 Fulfilment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Alexandra Blessing, BS

Fort Worth, TX, USA

May 2018

ACKNOWLEDGEMENTS

Funding for this work provided by the UNTHSC Institute for Molecular Medicine Intramural Seed Grant RI10075.

I wish to express my deepest appreciation for my major professor, Dr. Nicole Phillips. Her guidance, encouragement, and advice were invaluable throughout the course of this project. I would also like to recognize the other members of our laboratory, Jie Sun and Talisa Silzer for their constant support.

Additionally, I wish to acknowledge Dr. Marcus Stoiber for his support regarding model training, prompt responses to our questions, and multiple updates to Tombo software based on our inquiries and feedback. I would also like to express my gratitude to Dr. Roxanne Zascavage for her guidance regarding the MinION Nanopore sequencing protocols.

Finally, I would like to thank my committee members, Dr. John Planz, Dr. Michael Allen, and Dr. Shaoqing He for their feedback and guidance.

TABLE OF CONTENTS

Page

CHAPTER I. BACKGROUND AND INTRODUCTION ……………………………….………....1

I. Understanding and detecting oxidative damage within the genome………………………………1

II. Limitations in current methods of detecting oxidative damage……………………………………3

III. Oxford Nanopore Technology: an improved method for detecting oxidative damage?..................4

CHAPTER II. EXPERIMENTAL DESIGN AND METHODOLGY……………………………....8

I. Modified and unmodified oligonucleotide synthesis……………………………………….……..8

II. Oligonucleotide resuspension and annealing………………………………………….……...9

III. Sticky end ligation of synthetic oligonucleotides ……………………………………….………...9

IV. Post-ligation cleanup………………………………………………………………….……..10

V. Fragment length assessment……………………………………………………….……..11

VI. Training set development……………………………………………………………….……...12

VII. MinION Nanopore library preparation and sequencing……………………………….……14

VIII. Loading the MinION SpotON Flow Cell and data collection……………………………………18

IX. Basecalling…………………………………………………………………………….…18

X. BWA-MEM sequence alignment and visualization……………………….………………...19

XI. Detection/characterization of 8-oxoguanine..…………………………………….……...…20

XII. Algorithm training……………………………………………………….…….....…21

CHAPTER II. RESULTS AND DISCUSSION…………………………………….….….…24

I. Phase 1: Proof of concept………………………………………….……….….…24

i. Ligation success and fragment length verification……………………………………24

ii. Nanopore sequencing: Rapid protocol vs. $1D^2$ protocol…………………………………27

iii. Examining the test set alignment using Integrated Genomics Viewer…………………..….28

iv. Tombo raw data analysis…………………………………………………………...31

II. Phase 2: 8-oxoguanine alternative model development……………………………………..35

i. The necessity of the training set……………………………………………………….35

ii. Canonical and alternative model training……………………………………………..36

III. CHAPTER III. CONCLUDING REMARKS AND FUTURE DIRECTIONS……..….40

APPENDIX A. Plot most-significant results…………………………………………………43

APPENDIX B. Pipeline for Albacore basecalling and alignment…………………………….47

REFERENCES……………………………………………………………………………..49

CHAPTER I

INTRODUCTION AND BACKGROUND

Understanding and detecting oxidative DNA damage within the genome:

Reactive oxygen species (ROS) occur naturally in the human body, contributing to normal cell signaling pathways. However, in times of environmental stress induced by heat or radiation exposure, for example, these free radicals may increase to a harmful level, capable of causing oxidative damage to DNA molecules, proteins, and lipids [1]. Potentially harmful reactive oxygen species include hydrogen peroxide, peroxide, hydroxyl radical, hydroxyl ions, singlet oxygen, and superoxide anions. These reactive oxygen species have been linked to the initiation and development of cancer and neurodegenerative diseases, such as Alzheimer's [2].

Even intracellularly-originating oxygen species may cause basal levels of damage to normally-functioning cellular components. In the event of an increase in ROS due to a traumatic environmental stressor, the functional ratio of oxidants to antioxidants will be altered and damage to cellular contents will be more extensive. Although damaged lipids and proteins may be digested and exported from the cell and replaced, DNA requires immediate repair.[3]

All four nucleotide bases may be affected by ROS, however, guanine (G), is the most susceptible to oxidative stress due to its low ionization potential.[4, 5] When guanine undergoes oxidative damage, it will commonly form 8-oxoguanine (8-oxoG), an oxidized and highly mutagenic form of the original nucleotide base (Figure 1A). The lesion base remains very similar to the parent; the biochemical alterations include changes at the C8 hydrogen, which is replaced by a keto group, and the protonation of the N7 lone pair [6].

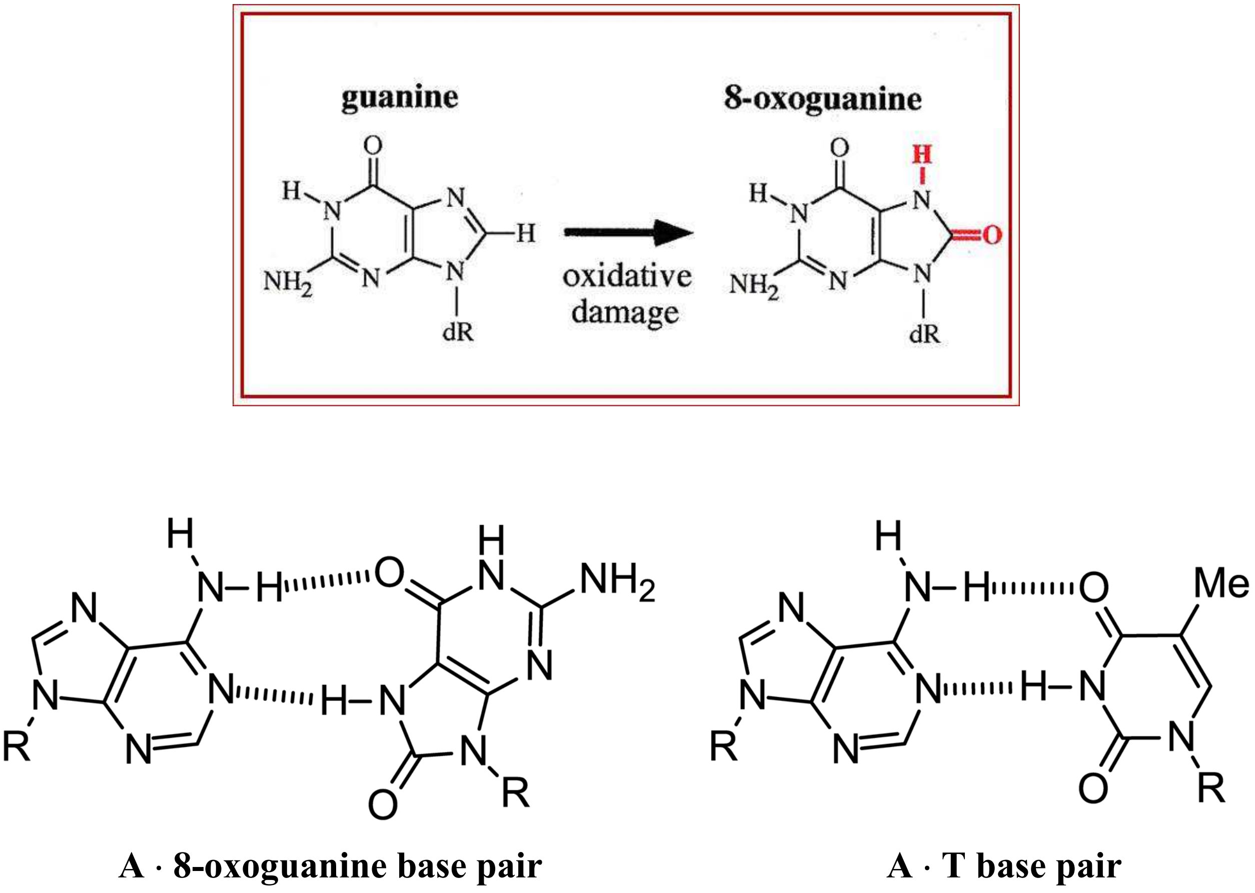


Figure 1: Guanine and 8-oxoguanine: structural differences and base pairing

(A) Oxidative damage is able to cause minute changes to the chemical structure of guanine, resulting in a lesion that can interfere with normal DNA replication [7]
(B) Structure of the A-O8G (8-oxoguanine) base pair, and comparison with the Watson-Crick A-T base pair [8]

Despite overall chemical similarity, 8-oxoguanine differs enough from the undamaged guanine to be recognized by DNA repair systems. The majority of 8-oxoguanine lesions are typically removed via the base excision repair (BER) pathway[9], however, it has been shown that the nucleotide excision repair pathway (NER) is also a suitable repair mechanism[10]. During the repair process, the mutagenic base is excised from its position in the nucleotide chain and removed from the nucleotide pool by 8-oxoguanine-DNA glycosylase, an enzyme responsible for the detection and elimination of damaged bases from the double helix [11]. If glycosylases fail to remove 8-oxoG from the DNA strand prior to replication, the oxidized bases are susceptible to transversion mutations [12]. In this case, DNA polymerases will couple the oxidized bases with adenines (Figure 1B), resulting in a nearly irreversible transversion from guanine to thymine in those positions. If the mutation occurs in exonic regions of a gene, the transversion will typically cause a non-synonymous amino acid substitution.

Several studies[13, 14] have documented increased levels of 8-oxoguanine in mitochondrial DNA when compared to nuclear DNA; this suggests that mtDNA could be especially susceptible to oxidative damage. As mitochondria produce a substantial amount of ROS during the oxidative phosphorylation process, their proximity to these oxidizing agents is consistent with the increased accumulation of 8-oxoguanine. Additionally, mtDNA is more susceptible to oxidative damage because it lacks the robust DNA-repair mechanisms that are present within the nucleus.[15]

Limitations in current methods of detecting oxidative DNA damage:

There are several methods that have been established for the detection and quantification of oxidative damage to the genome. These include high-performance liquid chromatography (HPLC), liquid chromatography-tandem mass spectroscopy (LC-MS/MS), gas chromatography-mass spectroscopy (GC-MS), a modified comet assay, immunoassays, quantitative PCR (qPCR), and imaging techniques. However, comparisons of these approaches have demonstrated an overall lack of reproducibility among assays [16]. In addition, each approach has individual drawbacks, ranging from artificial generation of oxidative damage, to misrepresentation of the extent of oxidative damage within a sample [17, 18]. Nanopore sequencing technology may provide a much-needed alternative to these often unreliable and inaccurate methods.

Oxford Nanopore technology: an improved method for detecting oxidative DNA damage?

Recently hailed as the "Future of DNA Sequencing," [19] nanopore technology has a number of advantages over currently employed Sanger and next-generation sequencing methods. Over a decade ago, Drs. Daniel Branton and David Deamer proposed that individual molecules could be detected and characterized by analyzing current shift when passing through an ion channel [20]. The application of this concept to DNA sequencing requires the use of naturally-occurring, pore-forming proteins which create a minute hole in an electrically resistant polymer membrane, less than 3 nm in diameter. The size of the pore is so restrictive that it is able to uncoil double-stranded nucleotide chains so that the individual nucleotides are passed through the pore sequentially and classified in real-time [21]. This a significant improvement on technologies such as Illumina™ and Ion Torrent™, which require amplification of DNA prior to sequencing. PCR-amplified sequencing libraries have been shown to decrease coverage of -GC and -AT -rich areas of DNA. Factors such as thermocycler and temperature ramp rate have been found to play a role in this amplification bias. A solution to significantly and consistently improve these obstacles has not been identified; it is, therefore, important to avoid amplification of DNA samples intended for sequencing if possible. As each individual nucleotide passes through the nanopore, it creates a unique and identifying disruption in the electrical current [Figure 2], so that each of the four nucleotide bases can be distinguished from one another.

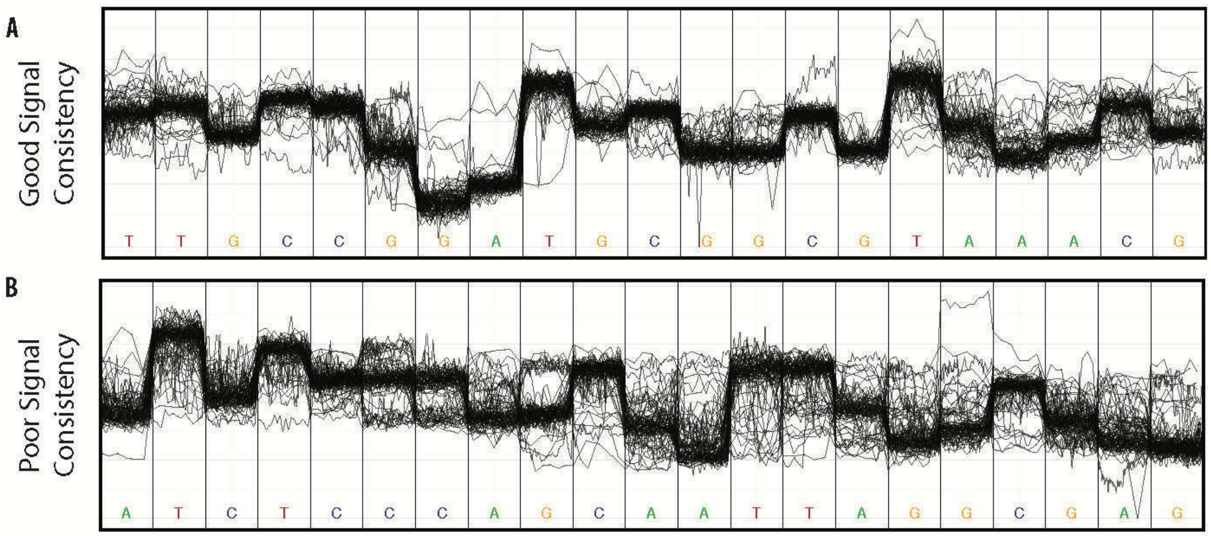


Figure 2: Example loci with good and poor signal correspondence

Raw signal generated by MinION Nanopore can be visualized using software such as *nanoraw* or *Tombo*. Figure 2 shows the difference between good and poor signal consistency, which is crucial for accurate basecalling.[22]

Labeled "fourth-generation DNA sequencing technology,"[23] nanopore devices have the potential to be an accurate and affordable option for genome sequencing. One of the greatest advantages of nanopore sequencing, however, is that amplification is not necessary. There are several benefits to PCR-free sequencing including the removal of amplification bias, the elimination of potential amplification-related errors and sample contamination, the removal of a time-consuming step in the overall sequencing workflow, and especially, the preservation of DNA modifications that would otherwise be lost in the amplification process [24] .

Oxford Technologies MinION is the first commercially available nanopore sequencer and offers several improvements on competing sequencing methods [25]. The MinION is portable, weighing less than 100 grams, generates analyzable data in real time, and has low start-up costs. However, this new technology is not without its challenges. In order to truly replace next-generation sequencing as the standard analysis method of the scientific community, additional studies must be completed in order to identify limitations of nanopore sequencing; some known limitations include a high error rate for individual reads compared to other sequencing technologies, and a relatively large amount of DNA required (only a limitation in certain applications). The long reads produced by the MinION enabled and enhanced sequencing of highly repetitive DNA regions as well as *de novo* genomic assembly in a wide variety of applications, ranging from immunogenetics [26] to Single Nucleotide Polymorphism (SNP) identification for forensic analysis [27].

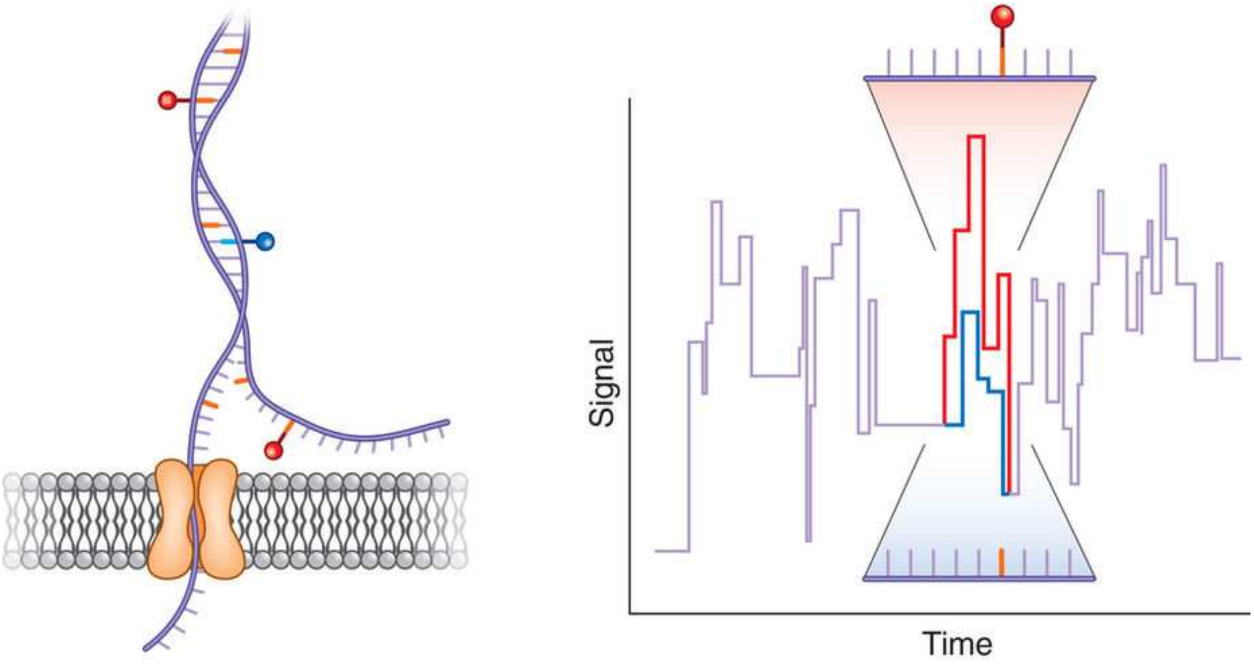


Figure 3: Raw Signal from Methylated Nucleotide Base[28]

Detection of shifting raw current signal due to passage of a modified base (shown as red/blue tag). Oxidized nucleotides, specifically 8-oxogunanine, are expected to generate a similar deviation in signal that is detectibly different from both unmodified bases as well as other modified bases such as 5-methylcytosine.

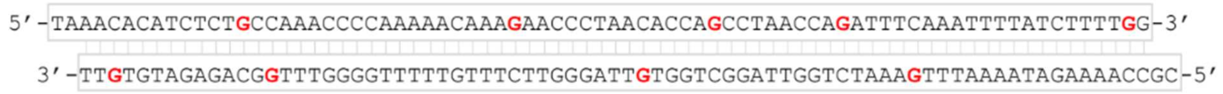
In previous studies, it has been shown that modified nucleotide bases have an electrical current disruption pattern that is distinct from the four canonical nucleotide bases [Figure 3]. This distinction can be detected using the raw sequencing data generated by the MinION Nanopore instrument. [22, 28, 29] Therefore, we hypothesize that the slight structural differences between guanine and 8-oxoguanine will result in a detectible alteration in current disruption within the nanopore. Building upon the method previously established by Simpson *et al*[29], we have developed a test sample set for classifying a fifth nucleotide, 8-oxoguanine, in the basecalling scheme. This will be accomplished via application of a suite tools, Tombo, which is dedicated to modified base detection in nanopore sequencing.

CHAPTER II

EXPERIMENTAL DESIGN AND METHODOLOGY

Modified and Unmodified Oligonucleotide Synthesis:

Due to the selective nature of *Taq* polymerase, standard amplification methods are unable to incorporate the 8-oxoguanine into the nucleotide chain at the same rate as dGTP [30, 31], therefore, a test set was created using four synthetic oligonucleotides [Figure 4] whose sequence was taken from the control region of the human mitochondrial genome: (1) a 74 base-pair mtDNA fragment containing a set percentage and known locations of 8-oxoguanine, (2) the complementary strand, also with known 8-oxoguanine modifications, (3) the equivalent unmodified 74 base-pair mtDNA sequence with no modifications, and (4) the unmodified complementary strand. Five guanine positions in the forward strand and four positons in the reverse strand were chosen to be substituted for 8-oxoguanine in the modified oligonucleotide. The length of the synthetic oligonucleotides and the number of modified bases included were limited due to the difficult nature of oligonucleotide synthesis when incorporating 8-oxoguanine.[32] All oligonucleotides for the test set were synthesized by Sigma-Aldrich (Sigma-Aldrich, Saint Louis, MO)[33].



5′-TAAACACATCTCT**G**CCAAACCCCAAAAACAAA**G**AACCCTAACACCA**G**CCTAACCA**G**ATTTCAAATTTTATCTTTT**G**G-3′

3′-TT**G**TGTAGAGACG**G**TTTGGGGTTTTTGTTTCTTGGGATT**G**TGGTCGGATTGGTCTAAA**G**TTTAAAATAGAAAACCGC-5′

Figure 4: Oligonucleotide mtDNA Sequence

The modified oligonucleotide fragment and its complement synthesized by Sigma-Aldrich are shown above. The guanine bases indicated in red have been replaced with 8-oxoguanine at each location. The unmodified oligonucleotide is the same as the modified oligonucleotide, but without 8-oxoguanine at the indicated positions.

Oligonucleotide Resuspension and Annealing:

Once received, the complementary oligonucleotides were resuspended in Nuclease-Free Duplex Buffer (Integrated DNA Technologies, Coralville, IA)[34] to reach 100uM final stock concentration. The resuspended oligonucleotides were subjected to a standard annealing protocol[35], which involved mixing the oligonucleotides in equimolar concentrations, subsequently heating them at 94 °C for 2 minutes, and then immediately cooling to room temperature. Aliquots of the annealed oligonucleotides were stored as recommended at -20 °C. The annealed oligonucleotides were quantified using Qubit Fluorometer, dsDNA Broad-Range (BR) Assay kit (Thermo Fisher Scientific, Waltham, MA)[36]. Fluorometric quantitation readings reported >1000ng for both annealed oligonucleotides, indicating the success of the annealing process.

Sticky-End Ligation of Synthetic Oligonucleotides:

As MinION nanopore sequencing requires at least 200 base-pairs in length for read initiation and accuracy, it was necessary to ligate the oligonucleotides in tandem to produce fragments of suitable length for sequencing. The synthetic oligonucleotides were designed with compatible ends, allowing for a sticky-end ligation to create alternating, repetitive nucleotide chains capable of detection and sequencing [Figure 5]. Approximately 100ng of DNA was mixed with 5uL of Instant Sticky-End Ligase Master Mix (New England BioLabs, Ipswitch, MA)[37] to facilitate the ligation reaction. These test samples will hereafter be referred to as poly-synthetic$_{\text{unmod}}$ and poly-synthetic$_{\text{mod}}$.

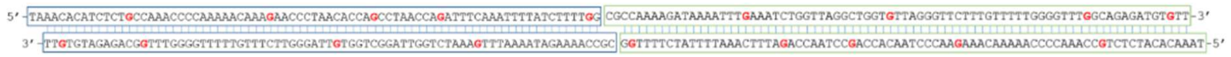

5′-TAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGG CGCCAAAAGATAAAATTTGAAATCTGGTTAGGCTGGTGTTAGGGTTCTTTGTTTTTGGGGTTTGGCAGAGATGTGTT-3′
3′-TTGTGTAGAGACGGTTTGGGGTTTTTGTTTCTTGGGATTGTGGTCGGATTGGTCTAAAGTTTAAAATAGAAAACCGC GGTTTTCTATTTTAAACTTTAGACCAATCCGACCACAATCCCAAGAAACAAAAACCCCAAACCGTCTCTACACAAAT-5′

Figure 5: Ligated Product

Figure 5 shows the double-stranded DNA fragment produced from the sticky-end ligation reaction. Fragment length distribution for ligated samples was verified using Agilent 4200 TapeStation.

Post-Ligation Cleanup:

Typically, ligated products are cleaned using standard PCR cleanup methods to remove enzymes and other components from the ligation reaction that may interfere with downstream reactions. The DNA Clean & Concentrator™-25 Kit[38] by Zymo Research (Irvine, CA) was initially selected for a post-ligation cleanup. This kit is designed for the rapid purification and recovery of up to 25ug of DNA using the Zymo-Spin™ column. The columns themselves contain a silica-based matrix that is able to maximize product recovery while ensuring complete elution with no buffer retention. Recovered DNA fragments range from 50 base-pairs to 23 kilobases; for fragments within this recommended range, users may expect 70-90% sample recovery.

A second cleanup protocol was tested to see if a different method of cleanup would affect fragment length distribution. The Agencourt AMPure XP PCR Purification protocol[39] (Beckman Coulter Inc, Brea, CA) was chosen. The AMPure process utilizes magnetic separation to purify DNA samples [see Figure 6], rather than a series of centrifugation steps. DNA fragments bind to the AMPure XP paramagnetic particles and are separated from impurities in solution. While placed on a magnetic stand, the particles are subjected to two of ethanol wash steps to remove contaminants. After the purified product is eluted from the beads, it is ready for analysis and sequencing. The AMPure XP kit is designed for the purification of fragments 100 base-pairs or longer.

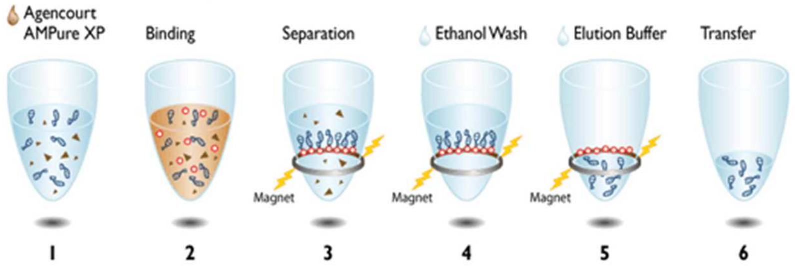


Figure 6: Agencourt AMPure XP Process Overview

The six step process for XP bead cleanup is shown here: 1. PCR amplified product (or DNA) 2. Binding of PCR amplicons to magnetic beads 3. Separation of PCR amplicons bound to magnetic beads from contaminants 4. Washing of PCR amplicons with Ethanol 5. Elution of PCR amplicons from the magnetic particles 6. Transfer away from the beads into a new plate.[39]

Fragment Length Assessment:

In order to assess the ligation efficiency and recovery after clean-up, fragment length distribution for both poly-synthetic$_{\text{unmod}}$ and poly-synthetic$_{\text{mod}}$ were assessed and quantified using the Agilent 4200 TapeStation Instrument[40] (Agilent Technologies, Santa Clara, CA), according to the manufacturer’s protocol. The Agilent 4200 Tapestation offers rapid processing for quality control of samples; DNA samples are automatically loaded, separated via electrophoresis, imaged, and analyzed. This entire process takes roughly 2 minutes per sample. Post-ligation poly-synthetic$_{\text{unmod}}$ and poly-synthetic$_{\text{mod}}$ samples were analyzed using the Agilent Genomic

DNA ScreenTape Assay[41], which was designed for sizing and quantification of genomic DNA with a size range of 200 to 60,000 base-pairs.

Training Set Development:

In order to build a new canonical base model, a human plasma sample was extracted using the Mag-Bind® Blood & Tissue DNA HDQ 96 Kit (Omega Bio-tek Inc., Norcross, GA)[42]. The collection and analysis of human plasma samples is approved under IRB protocol 2015-169. The extracted sample was quantified using Qubit Fluorometer, dsDNA Broad-Range (BR) Assay kit (Thermo Fisher Scientific, Waltham, MA)[36].

Training Set Amplification:

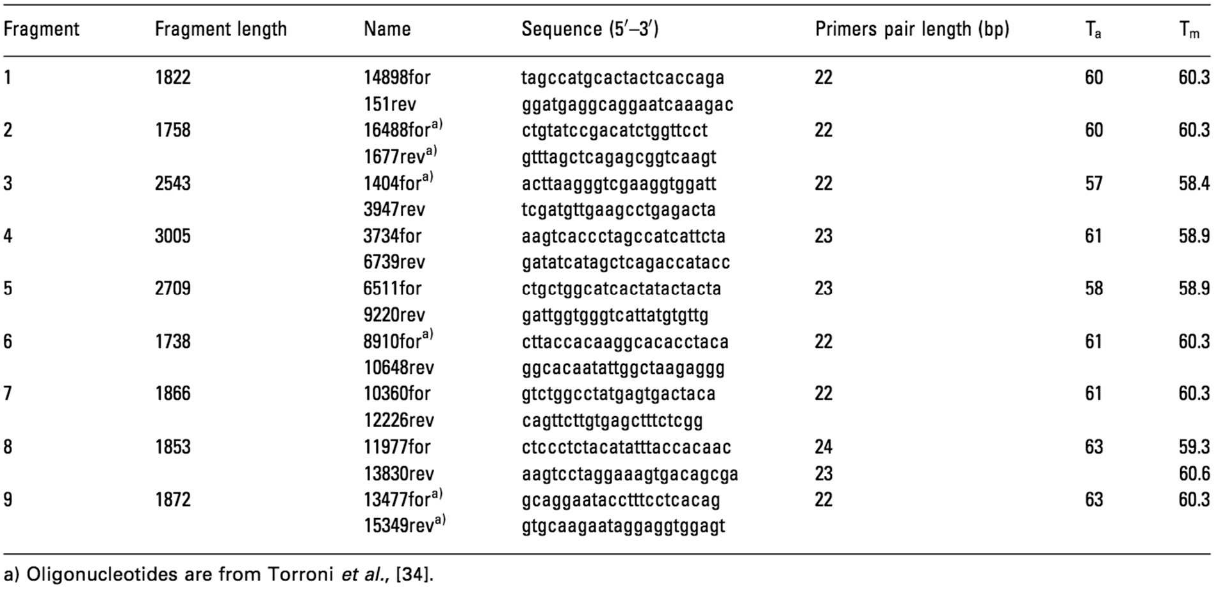
As previously mentioned, 8-oxoguanine is a difficult base to amplify; some percentage of guanine must be included in order to ensure amplification success. To determine the optimal ratio of unmodified dNTPs to 8-oxoguanine, we performed a series of amplification reactions using different concentrations of guanine and 8oxoG (Table 1).

	[INPUT DNTP]	POST-AMP [SAMPLE]
[GUANINE] > [8-OXOGUANINE]	10mM: dATP, dCTP, dTTP, dGTP 5mM: 8-oxoG	160ng/uL
[GUANINE] = [8-OXOGUANINE]	10mM: dATP, dCTP, dTTP, dGTP, 8-oxoG	136ng/uL
[GUANINE] < [8-OXOGUANINE]	10mM: dATP, dCTP, dTTP, 8-oxoG 5mM: dGTP	179 ng/uL
	10mM: dATP, dCTP, dTTP, 8-oxoG 2.5mM: dGTP	112 ng/uL
	10mM: dATP, dCTP, dTTP, dGTP, 8-oxoG 1.25mM: dGTP	54.7 ng/uL
	10mM: dATP, dCTP, dTTP, 8-oxoG 0.625mM: dGTP	43.2 ng/uL

Table 1: [dNTP] Optimization Test Results. We determined that amplification was possible with less dGTP than 8-oxoguanine, and then systematically lowered the dGTP concentration to determine the lowest amount of input dGTP possible to maintain reaction efficacy. The optimal dGTP concentration was determined to be 2.5mM, while maintaining all other dNTPs (including 8-oxoguanine) at the recommended 10mM.

After determining amplification success using less dGTP than 8-oxoguanine, we prepared another series of amplification reactions with increasingly less input dGTP to determine whether it was possible to decrease dGTP concentration and maintain amplification efficiency. The optimal dGTP concentration was determined to be 2.5mM, or one quarter of the recommended individual dNTP concentration.

The sample was then amplified in a single-plex reaction using Takara LA PCR Kit v. 2.1 (Takara Bio USA, Inc., Mountain View, CA)[43] and an amplicon derived from Ramos, et al (Figure 7)[44]. The sample was amplified twice: (1) using equal concentrations of canonical dNTPs and (2) using equal concentrations (10mM) of dATP, dTTP, dCTP, and 8-oxoguanine. A lower concentration (2.5mM) of dGTP was added to this reaction mix as well to increase reaction efficiency. These amplified training samples will hereafter be referred to as mitoAmp$_{unmod}$ and mitoAmp$_{mod}$.



Fragment	Fragment length	Name	Sequence (5′–3′)	Primers pair length (bp)	T_a	T_m
1	1822	14898for	tagccatgcactactcaccaga	22	60	60.3
		151rev	ggatgaggcaggaatcaaagac			
2	1758	16488for[a)]	ctgtatccgacatctggttcct	22	60	60.3
		1677rev[a)]	gtttagctcagagcggtcaagt			
3	2543	1404for[a)]	acttaagggtcgaaggtggatt	22	57	58.4
		3947rev	tcgatgttgaagcctgagacta			
4	3005	3734for	aagtcaccctagccatcattcta	23	61	58.9
		6739rev	gatatcatagctcagaccatacc			
5	2709	6511for	ctgctggcatcactatactacta	23	58	58.9
		9220rev	gattggtgggtcattatgtgttg			
6	1738	8910for[a)]	cttaccacaaggcacacctaca	22	61	60.3
		10648rev	ggcacaatattggctaagaggg			
7	1866	10360for	gtctggcctatgagtgactaca	22	61	60.3
		12226rev	cagttcttgtgagctttctcgg			
8	1853	11977for	ctccctctacatatttaccacaac	24	63	59.3
		13830rev	aagtcctaggaaagtgacagcga	23		60.6
9	1872	13477for[a)]	gcaggaatacctttcctcacag	22	63	60.3
		15349rev[a)]	gtgcaagaataggaggtggagt			

a) Oligonucleotides are from Torroni *et al.*, [34].

Figure 7: Table adapted from Ramos, et al.[44] Validated primers to amplify the complete mtDNA in nine overlapping fragments. Melting temperatures and annealing temperatures for each pair of primers are also presented. Amplicon 7 was chosen for our training set.

MinION Nanopore Library Preparation and Sequencing:

Rapid Sequencing Protocol (SQK-RAD003)

The poly-synthetic$_{\text{unmod}}$ sample was prepared for sequencing following the Oxford Nanopore Technologies Rapid Sequencing Protocol (SQK-RAD003). This sequencing method requires approximately 400ng of high-molecular weight gDNA with a length greater than 30,000 base-pairs. This protocol is an expedited, two-step approach that generates sequencing libraries using a transposase which randomly fragments the original DNA strands and attaches tags and sequencing adaptors to the cleaved ends[45]. The Rapid Sequencing Protocol was initially chosen for this project after noting the successful sequencing of bacterial genomes using this method.[46] Results of the Rapid Sequencing protocol were analyzed to determine if sequencing of the modified oligonucleotide should proceed.

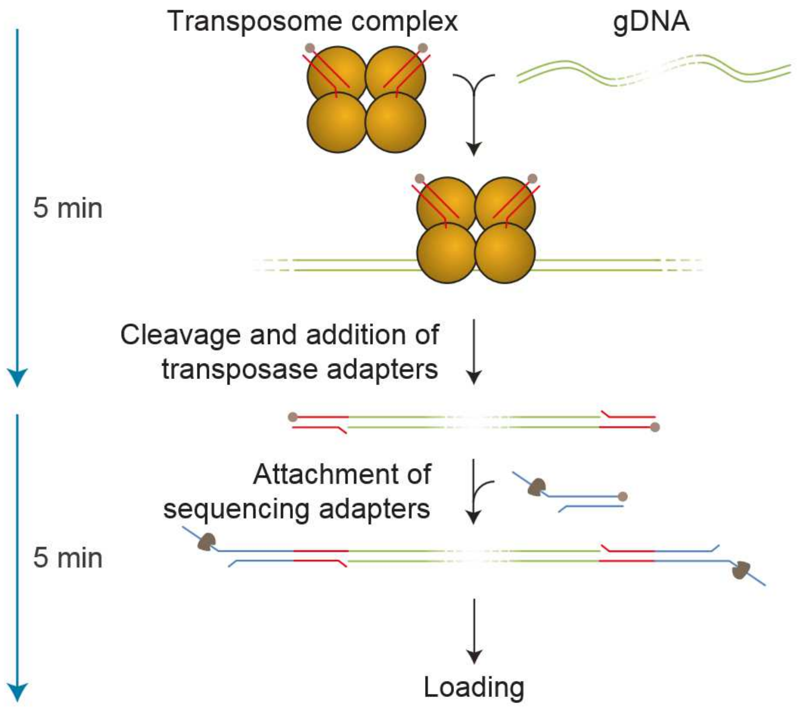

Transposome complex
gDNA
5 min
Cleavage and addition of transposase adapters
Attachment of sequencing adapters
5 min
Loading

Figure 8: Rapid Sequencing Kit Workflow

The Rapid Sequencing Kit generates sequencing libraries from extracted gDNA in 10 minutes using a simple 2-step protocol. At the heart of the kit is a transposase which simultaneously cleaves template molecules and attaches tags to the cleaved ends. Rapid Sequencing Adapters are then added to the tagged ends. [45]

$1D^2$ Sequencing Protocol (SQK-LSK308)

The test sample set and training sample set were prepared for sequencing via the ID^2 protocol[47]. The $1D^2$ Sequencing Kit provides the highest raw read accuracy of Oxford Technologies Nanopore Sequencing kits. $1D^2$ sequencing is designed to promote the consecutive sequencing of the template followed by the complement strand and the DNA fragmentation step is optional. This method (Figure 9) requires approximately eighty minutes of library preparation time with an optimal input of 1000ng of dsDNA.

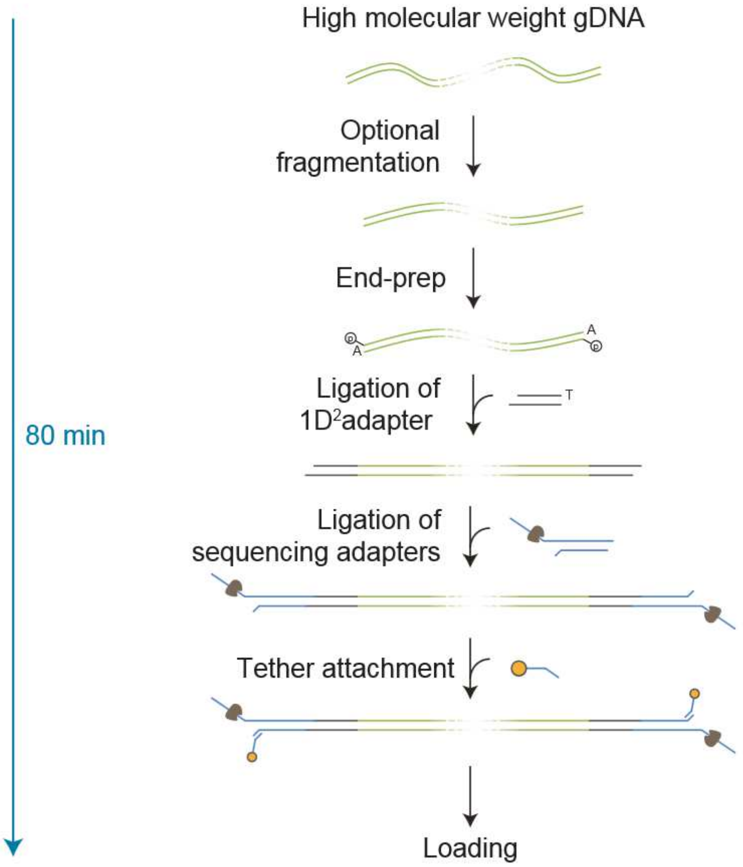
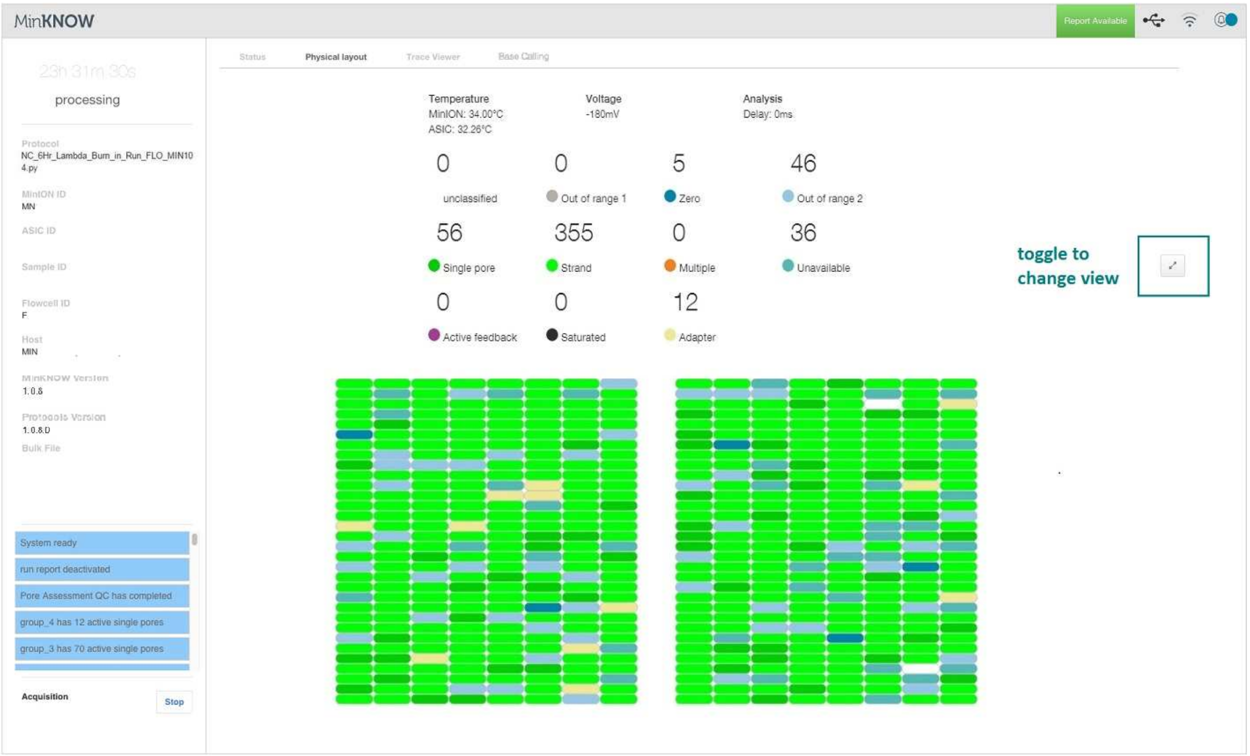


Figure 9: $1D^2$ Sequencing Kit Workflow

If necessary, the genomic DNA is fragmented in a Covaris g-TUBE. However, if your experiment requires long reads, fragmentation is not advised. The DNA ends are then end-repaired and dA-tailed using the NEBNext End Repair/dA tailing module. Adapters supplied in the Ligation Sequencing Kit are then ligated onto the DNA. Adapters introduce the components needed for the DNA to enter the pore, and the $1D^2$ Adapter allows the pore to capture the complement strand immediately after the template. One strand of the duplex is sequenced at a time, producing $1D^2$ reads.[48]

Immediately before the sequencing run is initiated, MinKNOW, software responsible for data acquisition and real-time analysis, activates a quality control procedure to evaluate the number of active pores. It is crucial to have at a high percentage of active pores available (near 1000) within the flow cell. After sequencing is initiated, MinKNOW will report the number of events, or data points in which the contents of the pore are consistent. The events are then grouped into reads using the signal change resulting from shifting pore contents. The main screen of the

MinKNOW Graphical User Interface (GUI) displays the activity of the flow cell's pores, color coding according to sequencing status.[49] The status will either indicate one of three possibilities: (1) the channel is actively sequencing, or is ready for sequencing, (2) data is not currently being collected. This can be due to overlapping signals from multiple active pores, or the channel is blocked by strand of DNA, (3) no significant data is being produced. This could be due to misloading of the flow cell, i.e., too much or too little DNA has been added to the sequencing library.



23h 31m 30s

processing

Protocol
NC_6Hr_Lambda_Burn_in_Run_FLO_MIN104.py

MinION ID
MN

ASIC ID

Sample ID

Flowcell ID
F

Host
MIN

MinKNOW Version
1.0.8

Protocols Version
1.0.8.D

Bulk File

System ready

run report deactivated

Pore Assessment QC has completed

group_4 has 12 active single pores

group_3 has 70 active single pores

Temperature
MinION: 34.00°C
ASIC: 32.26°C

Voltage
-180mV

Analysis
Delay: 0ms

0	0	5	46
unclassified	● Out of range 1	● Zero	● Out of range 2
56	355	0	36
● Single pore	● Strand	● Multiple	● Unavailable
0	0	12	
● Active feedback	● Saturated	● Adapter	

Figure 10: MinKNOW Sequencing Checks and Monitoring

A good library will be indicated by a high proportion of light and dark green channels. The combination of light and dark green indicate the number of active pores at any point in time and the dark green indicate the proportion of pores in strand (or sequencing) at a particular time point. A low proportion of dark green channels will reduce the throughput of the sequencing[49].

Loading the MinION SpotON Flow Cell and Data Collection:

Both Rapid Sequencing and 1D^2 Sequencing methods were used during the course of this project. After the priming of the SpotON Flow Cells and the preparation and loading of the respective DNA libraries, the sequencing protocol was initiated and left to run for at least 24-30 hours. The poly-synthetic$_{\text{unmod}}$ sample was sequenced on a single flow cell until sufficient coverage was obtained. After adequate reads were acquired for the poly-synthetic$_{\text{unmod}}$ sample, the poly-synthetic$_{\text{mod}}$ sample was sequenced on a different flow cell. Both samples were sequenced without pooling or barcoding. For both runs, new SpotON flow cells were assessed for QC protocol performance and utilized upon passing.

The acquired data were sorted into /pass and /fail directories based on preliminary quality analysis by MinKNOW. The data produced are stored in FAST5 (.fast5) file format, with one FAST5 file created per read. The raw FAST5 files are archived in directories in batches of 4,000 FAST5 files (reads). The raw FAST5 files were "unpacked" into a single directory, and were utilized by the base-calling and alignment software.

Basecalling:

Albacore software (v.2.0.2) was utilized for the basecalling of acquired reads. Oxford Nanopore Technologies provides a data processing pipeline called Albacore for basecalling which is utilized both by MinKNOW and external applications for processing DNA libraries generated by all available sequencing kits. Prior to initiating Albacore, a directory was created for the storage of base-caller output. Albacore generates called/processed FAST5 files (one per read) and a FASTQ file (one per raw FAST5 directory). FASTQ files contain information from each read within the directory along with a corresponding quality score. The quality of each base in a single read is represented by a series of characters with ASCII codes that correspond to the quality score. The FASTQ files from each raw-read directory were unpacked and merged prior to alignment.

BWA-MEM sequence alignment and visualization:

After the FASTQ files were merged, the sequences were aligned to a reference, in this case, the known mtDNA sequence of the unmodified oligonucleotide. There are many alignment tools available, each with its advantages based on characteristics of the sequencing data (such as read length, error rate) and characteristics of the reference genome (such as size, complexity, structure). BWA-MEM (Burrows-Wheeler Aligner)[50], a software package for the rapid and accurate alignment of nanopore sequencing data, was chosen for the alignment process. Using SAMtools, the reference sequence was indexed prior to initiating the alignment; the FASTA file (typically .fasta or .fa) is a text file which contains the actual reference sequence. BWA-MEM alignment was generated from the merged FASTQ file, generating the .bam file (see Appendix A for detailed command). Subsequent processing using SAMtools included (1) `-samtools sort`, sorting the alignment according to the uploaded reference sequence, and (2) `samtools index-`, generating the index file for the alignment (.bai), and (3) `samtools stats-,` generating summary information and statistics describing the alignment run. Integrated Genomics Viewer (ITG)[51], an imaging tool for sequencing data, was then used to visualize the aligned reads and assesses consensus basecalling differences between the poly-synthetic$_{\text{unmod}}$ and the poly-synthetic$_{\text{mod}}$.

Detection/characterization of 8-oxoG:

To train an algorithm/model for the recognition of the modified base 8-oxoG, the raw current data must be accessed and analyzed. This was accomplished using a relatively new python-based software suite called Tombo (v.1.2.1), a software package created by Dr. Marcus Stoiber and designed specifically for the detection and visualization of modified nucleotides. Tombo is the next-generation of Nanoraw, the original software suite designed for the visualization of raw nanopore sequencing data (Nanoraw being phased out currently). After configuring environmental requirements for Tombo, it was installed on a local server (running Python 2.7), using `pip install ont-tombo[full].` Tombo generates internally-required index files by "resquiggling" the Albacore base-called FAST5 reads. The following command structure was executed to resquiggle both poly-synthetic$_{unmod}$ and the poly-synthetic$_{mod}$:

```
tombo resquiggle
<mod/unmod_FAST5_pass_directory> <SyntheticDNA_reference.fa> --minimap2-index
<indexfile> --processes 4.
```

Tombo has a built-in aligner, minimap2, which (1) indexes the reference, and (2) performs an alignment. Minimap2 is the newest aligner that handles noisy, long-reads with highest efficiency, replacing BWA-MEM in many cases as the ideal aligner for nanopore reads. Per correspondence with Dr. Stoiber (https://github.com/nanoporetech/tombo/issues/38) minimap2 was used to generate an alignment index with better-optimized word size (kmer -k) and window (-w): `minimap2 -k 3 -w 3 -d SyntheticDNA_reference_minimap2.idx.` The alignment/mapping parameters are not accessible to the user; however, the user can generate an index for the alignment outside of Tombo, and feed the index files (.idx) to Tombo for its alignment. This is helpful for mapping to shorter references since the built-in indexing parameters for Tombo are optimized for genomic scale alignment.

```
Tombo resquiggle –
minimap2-index SyntheticDNA_reference_minimap2.idx
```

`mod/unmod_olgio/FAST5/pass/ SyntheticDNA_reference.fa –processes 4.` The resquiggled data was then tested for regions of significant deviation: `tombo test_significance --fast5-basedirs mod_oligo/FAST5/pass/--control-fast5-basedirs unmod_olig/FAST5/pass/ --statistics-file-basename mod_unmod_compare.` Squiggle plots were then generated for the most-significant regions of signal difference throughout the reference sequence: `tombo plot_most_significant --tombo_model_filename/path/mod_unmod_compare.`

Algorithm Training:

Oxford Nanopore Technologies provides the expected values for each change in nucleotide bases; the alterations in electrical current are continuously collected as each base passes through the pore. This shift in current is expressed as a squiggle plot. The inclusion of a damaged base within the nucleotide chain is expected to generate a unique change in the electrical current. As 8-oxoguanine enters and exits the pore, the resulting current should be fundamentally different when compared to all possible combinations of unmodified nucleotides when occupying the pore.

The raw sequencing data produced by the MinION Nanopore is visualized in squiggle plots, or line graphs depicting electrical current fluctuations vs. time. Tombo uses the reference sequence as a standard base model; the software identifies the localized clusters of bases that exhibit significant deviation from the canonical-base sequence reads, thereby identifying modified nucleotide positions. It is also able to retain this information to detect modified nucleotides from future sequencing run data sets. Tombo does not require the event calls (i.e., signal shifts that result as each base exits the pore) generated by MinKNOW, but discovers events directly from the raw signal for more efficient basecalling. The events are determined by detecting large alterations in current level and defining the distance between neighboring signal shifts. The largest shifts in current are chosen as the dividing points between events, and the raw signal is normalized. The algorithm is able to take the given reference sequence and the defined signal segments and assign the most likely pairing between the two.

The `estimate_reference` command was used to establish a model for canonical bases, called *unmod_estimate_reference*. This was accomplished using the data from the poly-synthetic$_{\text{unmod}}$ samples and amplicons generated without 8-oxoG. The alternate reference, according to Tombo, must contain the four canonical bases along with a single, known, alternative base incorporated randomly instead of one canonical base into a sample with a known genome.[52] The poly-synthetic$_{\text{mod}}$ sample fulfills these criteria, and has the added benefit of analyzing signal deviations at known modified positions. Additionally, amplification of DNA in the presence of 8-oxoguanine will result in randomly modified product; these samples can also be used for algorithm training. The `estimate_alt_reference` was used to model the alternative sample by taking a number of reads from both modified samples grouped by -kmer at the location assigned by the re-squiggle algorithm. Once all reads have been processed, a kernel density estimate is calculated for each -kmer for both reference and alternative samples. The alternative distribution was then isolated. This is accomplished using an algorithm that assumes a portion of the alternative density represents the canonical base density. The density attributed to canonical bases was removed and the remainder is determined to be the alternative base distribution. The alternate base model was stored as an HDF5 file (called *8oxoG_model_file*) and was called for subsequent plotting and re-squiggling commands.

Preliminary 8-oxoG model testing was conducted after model development by creating an *in silico* mixture of the poly-synthetic$_{\mathrm{mod}}$:polysynthetic$_{\mathrm{umod}}$ reads of 50:50. Reads were re-squiggled using our newly developed model for canonical base calling, *unmod_estimate_reference*. The re-squiggled and aligned data were then tested for detection of 8-oxoG using the `test_significant` command and the `--alternative-model filename` option. Individual reads were assessed for likelihood of canonical guanine versus 8-oxoguanine. Plots of these individual reads with statistics were generated using `plot-per-read`.

CHAPTER III

RESULTS AND DISCUSSION

Phase 1: Proof of Concept

<u>*Ligation Success and Fragment Length Verification*</u>

The sticky-end ligation for both the poly-synthetic$_{\text{unmod}}$ sample and poly-synthetic$_{\text{mod}}$ sample produced DNA fragments >100 to <48500 base-pairs, thereby surpassing the minimum length requirement for MinION Nanopore sequencing. (Figure 11).

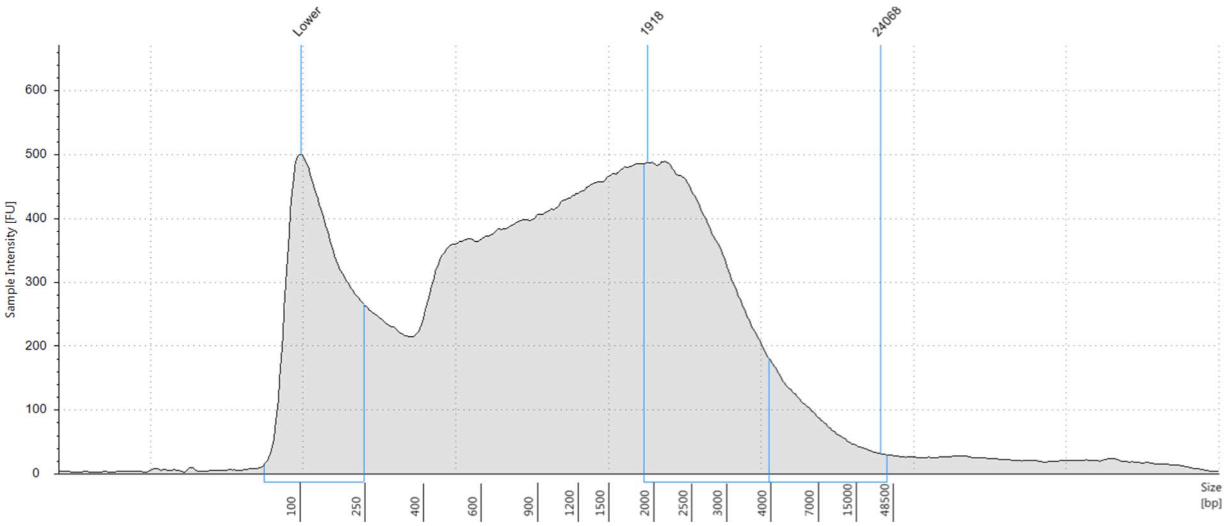


Figure 11: Agilent 4200 TapeStation electropheregram for poly-synthetic$_{\text{unmod}}$ ligated sample

Fragment length distribution displays a high degree of variability, indicating the success of the ligation reaction. Fragments range from <100 to >50,000 bases, with peaks called at 1,918 and 24,068 base-pairs.

Immediately after confirming fragment length, the ligation product was subjected to a cleanup to remove potential contaminants using the Zymo DNA Clean & Concentrator™-25 kit. Fragment length was assessed for the post-cleanup samples. Fragment length distribution results from the Agilent indicated that the majority of DNA fragments present after the cleanup were roughly 100bp in length, with very low amounts of longer fragments (Figure 12).

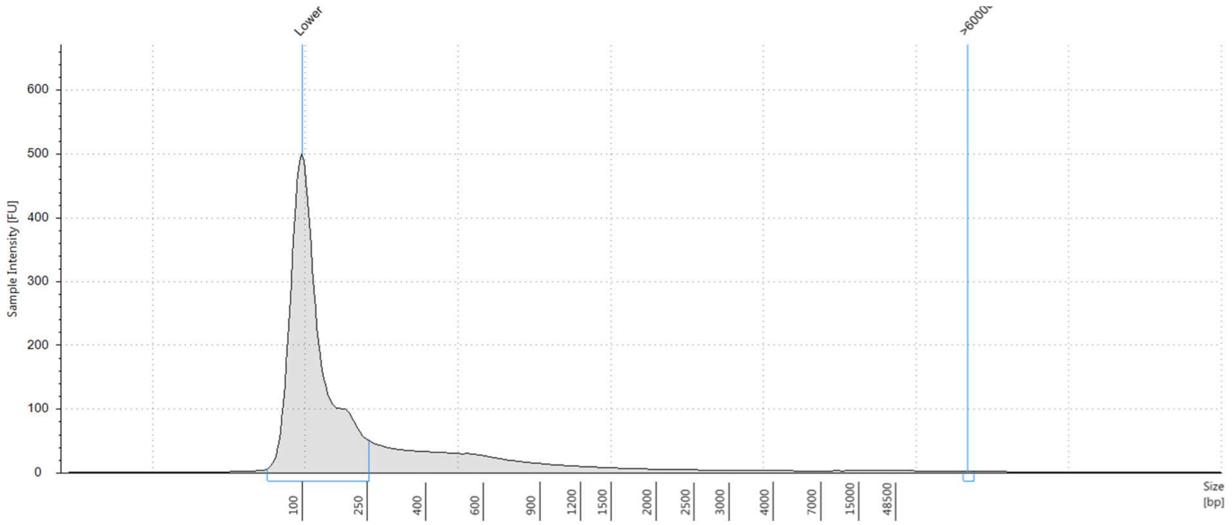


Figure 12: Agilent 4200 TapeStation electropherogram for poly-synthetic$_{\text{unmod}}$ ligated sample with Zymo Clean & Concentrator™ -25 Cleanup

Genomic DNA kit contains a lower marker with a defined concentration for accurate quantification calculation. The cleaning process appears to have removed the majority of DNA fragments >1,000 bases in length and most of the remaining fragments appear to be close to the lower marker. One peak is called at >60,000 base-pairs, but the sample intensity is >10 FU, indicating the low concentration of sample at this length.

After unsuccessfully attempting to fine-tune the Zymo DNA Clean & Concentrator™-25 kit protocol to retain the larger DNA fragments, the Agencourt AMPure XP PCR Purification protocol was chosen to see if different cleanup techniques could improve longer fragment recovery.

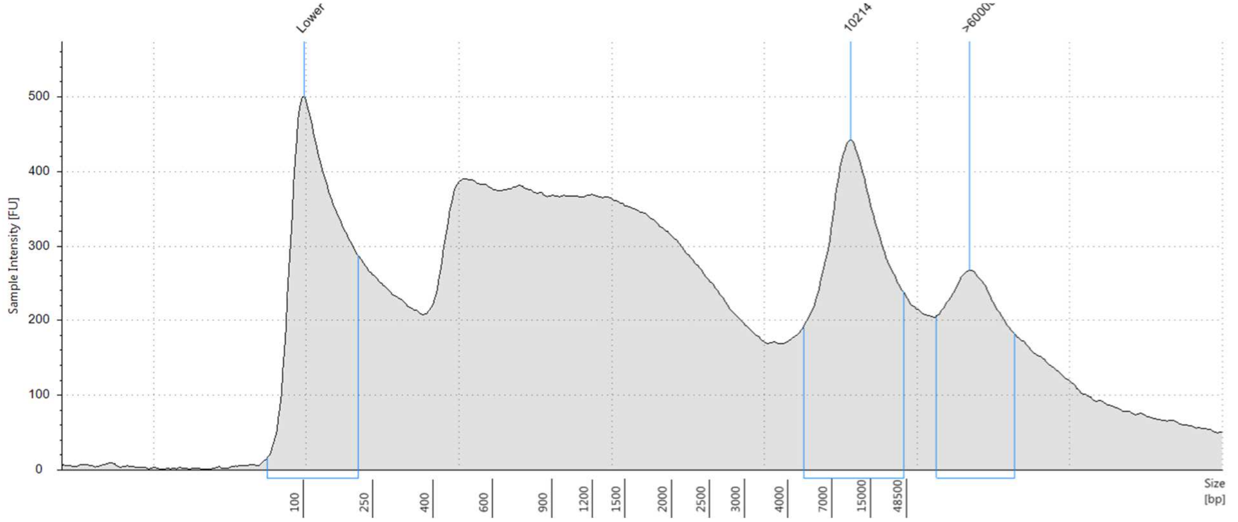


Figure 13: Agilent 4200 TapeStation electropherogram for poly-synthetic$_{\text{unmod}}$ ligated sample

Fragment length distribution displays a high degree of variability, indicating the success of the ligation reaction. Fragments range from >60,000 to <100 bases, with peaks called at 10,214 and 60,000 base-pairs.

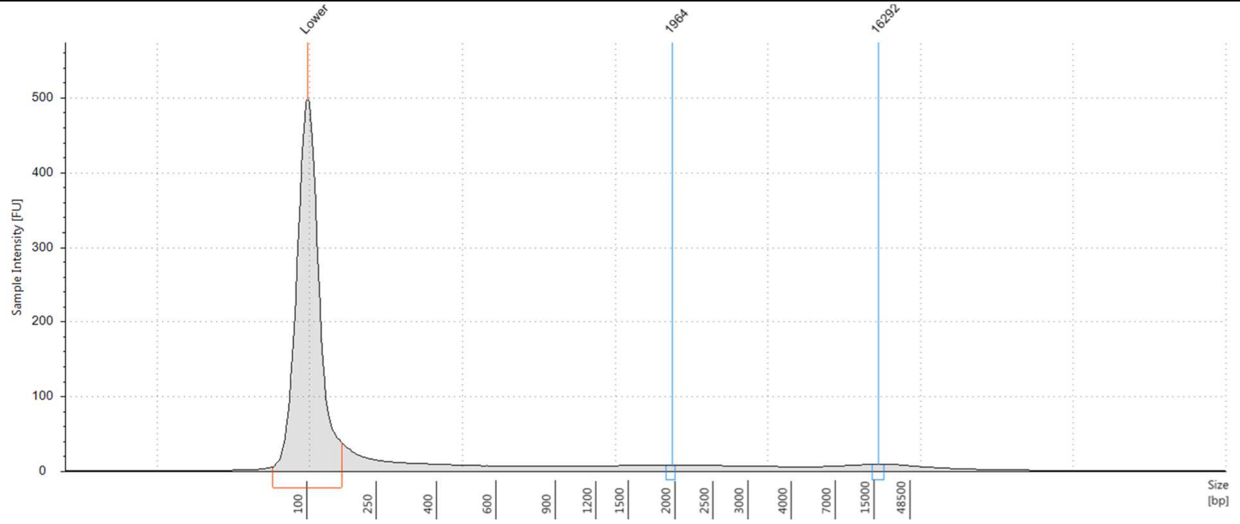


Figure 14: Agilent 4200 TapeStation electropherogram for poly-synthetic$_{\text{unmod}}$ ligated sample with Agencourt AMPure XP Cleanup

The cleaning process appears to have removed the majority of DNA fragments >1,000 bases in length and most of the remaining fragments appear to be close to the lower marker. Small peaks were detected at 1,964 base-pairs and 16,292 base-pairs, but the intensity is > 10 FU, indicating very low sample concentration at these lengths.

The fragments imaged following the AMPure XP cleanup yielded similar results to the Zymo DNA Clean & Concentrator™-25 cleanup (Figures 13 & 14). It was decided to forego the cleanup step and proceed directly to the sequencing process with the ligation reaction. We are unsure as to the reason for the severe product loss in the cleanup attempts. Although we took precautions, loss of long fragments can occur due to overdrying the sample prior to elution, or difficulty in eluting the long fragments from the beads or filter.

Nanopore Sequencing – Rapid Protocol vs. 1D2 Protocol

The first attempts at sequencing the poly-synthetic$_{\text{unmod}}$ sample was accomplished using the Rapid Sequencing Kit (SAQ-RAD003). Acquired reads numbered around 500 total for these initial sequencing attempts. At least half of the acquired reads were labeled “failed reads” by the

MinKNOW software and were unable to be utilized in basecalling attempts. It became apparent that the random fragmentation was too harsh_for the samples, perhaps due to the repetitive nature of the synthetic oligonucleotide fragments. After two unsuccessful sequencing attempts using the Rapid Sequencing Protocol, we elected to use the 1D^2 kit, designed for maximum read accuracy and, most importantly, lacks the mandatory fragmentation. The 1D^2 Protocol contains a series of purification steps using the Agencourt AMPure XP particles. As demonstrated by the unsuccessful cleanup attempts, the sample concentration was (roughly) halved after each AMPure XP bead purification. $1D^2$ Sequencing Runs yielded approximately 150,000 reads for the poly-synthetic$_{unmod}$ test sample and 54,000 reads for the poly-synthetic$_{mod}$ test sample. Of these total reads, the reads that passed were sufficient to successfully basecall, align, and visualize the data in Integrated Genomics Viewer[53].

Examining the Test Set Alignment using Integrated Genomics Viewer:

The resulting base-called sequencing reads for both the poly-synthetic$_{unmod}$ sample and the poly-synthetic$_{mod}$ sample were aligned against the reference sequence using IGV. The reads for the poly-synthetic$_{unmod}$ sample were mostly concordant, with the typical inaccuracies that accompany nanopore sequencing. Interestingly, the greatest area of dissimilarity between the poly-synthetic$_{unmod}$ sample and the reference was at the point of fragment ligation.

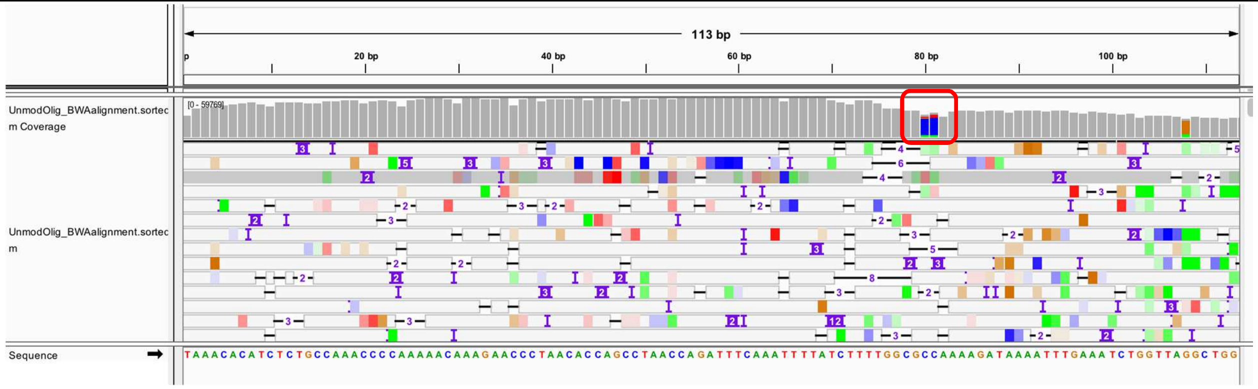

113 bp
p
20 bp
40 bp
60 bp
80 bp
100 bp
UnmodOlig_BWAalignment.sortec m Coverage
[0 - 59769]
UnmodOlig_BWAalignment.sortec m
Sequence
TAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGCCAAAAGATAAAATTTGAAATCTGGTTAGGCTGG

Figure 15: Image generated using IGV showing the alignment of MinION reads for poly-synthetic$_{\text{unmod}}$ test sample to synthetic oligonucleotide reference sequence

The mapped reads for poly-synthetic$_{\text{unmod}}$ display very little deviation from the reference sequence. With the exception of the point of ligation (shown boxed in red), the majority of reads are concordant, with <10% alternate basecalls for each base.

The bars shown at the top of the graph are the read coverage track, and represent coverage of all usable reads. IGV uses colors and other visual markers to highlight variation in reads against the reference sequence. Read bases that match the reference are shown in gray; those that do not match are color coded. The color blue designates cytosine, gold is guanine, green is adenine, and red is thymine. Insertions are marked by purple "I" and deletions are marked with black dashes "-." The accompanying digit indicates the number of bases inserted or deleted at that position. Additionally, mismatched bases are also assigned a transparency value proportional to the read quality known as a phred score. The bolder the color, the greater the degree of mismatch.[54]

The comparison of the poly-synthetic$_{\text{mod}}$ sample and the reference (Figure 16) revealed many more points of low confidence for base-calling, corresponding to variations caused at the 8-oxoG locations.

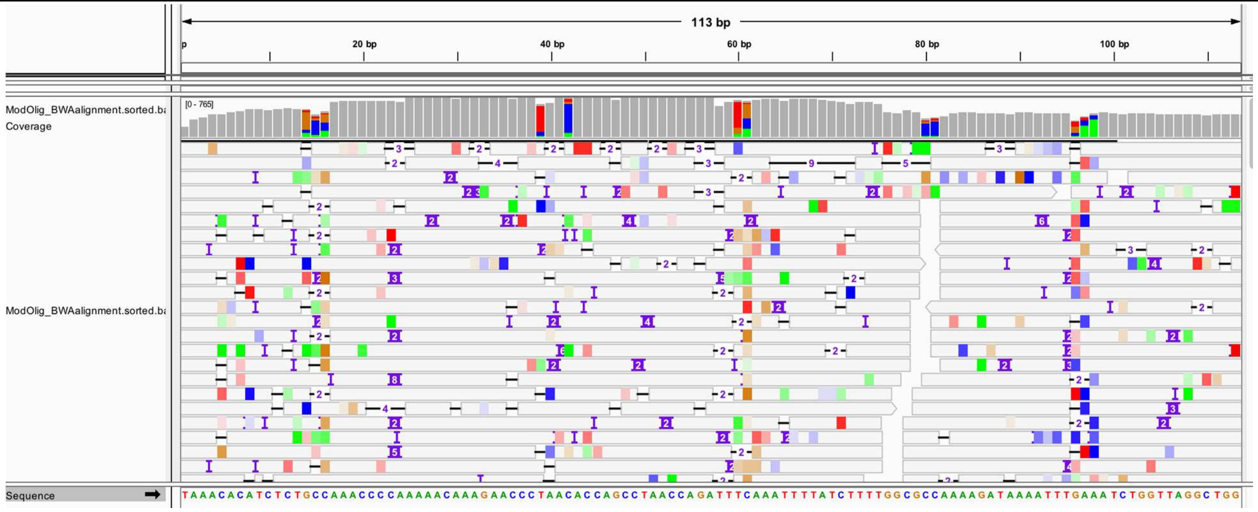
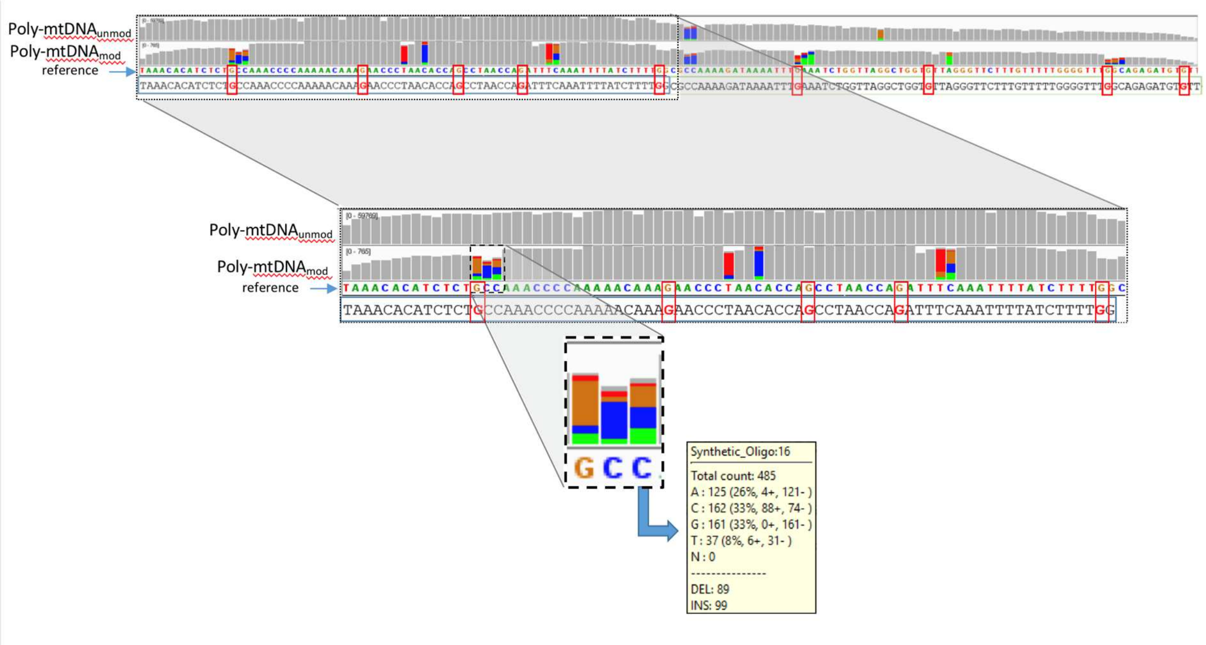

113 bp
p
20 bp
40 bp
60 bp
80 bp
100 bp
ModOlig_BWAalignment.sorted.ba
Coverage
[0 - 765]
ModOlig_BWAalignment.sorted.ba
Sequence
TAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGCCAAAAGATAAAATTTGAAATCTGGTTAGGCTGG

Figure 16: Image generated using IGV showing the alignment of MinION reads for poly-synthetic$_{mod}$ test sample to synthetic oligonucleotide reference sequence

The mapped reads show many points of low-confidence for basecalling, particularly around 8-oxoguanine locations. While areas shown in gray indicate high-consensus basecalling for the majority of aligned reads, areas shown in color indicate a relatively high percentage of alternate basecalls (≥ 10%).

As DNA strands translocate through the pore, a kmer of approximately 5 nucleotides occupies the pore at any point in time during the sequencing process. Therefore, it was expected to see a significant number of alternate basecalls caused by the oxidized base since the presence of 8-oxoguanine affects the reads for all bases simultaneously sharing the pore space. The results shown here provide strong evidence that 8-oxoG disrupts the current flow through the nanopore significantly, and that further characterization of the signal disruption may allow for calling of 8-oxoG in unknown samples.


Poly-mtDNA$_{unmod}$
Poly-mtDNA$_{mod}$
reference
TAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGCCAAAAGATAAAATTTGAAATCTGGTTAGGCTGGTGTTAGGGTTCTTTGTTTTTGGGGTTTGGCAGAGATGTGTT
Poly-mtDNA$_{unmod}$
Poly-mtDNA$_{mod}$
reference
TAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGG
G C C

Synthetic_Oligo:16

Total count: 485
A : 125 (26%, 4+, 121-)
C : 162 (33%, 88+, 74-)
G : 161 (33%, 0+, 161-)
T : 37 (8%, 6+, 31-)
N : 0

DEL: 89
INS: 99

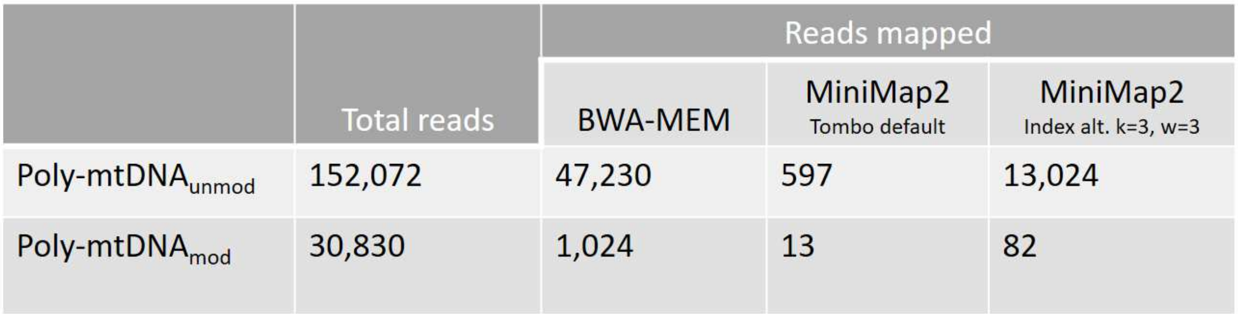
Figure 17: IGV Comparison of poly-synthetic$_{unmod}$ and poly-synthetic$_{mod}$ reads

The coverage of poly-synthetic$_{unmod}$ reads (top) and poly-synthetic$_{mod}$ reads (middle) is shown compared to the reference sequence (bottom). The expanded portion indicates an area of low consensus for the poly-synthetic$_{mod}$ sample attributed to the presence of 8-oxoguanine. Out of 485 total reads at the C location, only 33% of these reads called this base a cytosine; 33% called an guanine, while 26% and 8% called adenine and thymine, respectively.

Tombo Raw Data Analysis:

Minimap2, as built into Tombo, is not optimized for our reference size or read structure. The built-in settings for alignment resulted in poor read mapping efficiency in our data. For example, the Tombo default mapping parameters of the poly-synthetic$_{mod}$ dataset yielded a single successfully aligned read out of the thousands of useable reads we were able to align in BWA-MEM. Minimap2 was designed for the alignment of large reads to sizeable sections of the genome, and specifically, to reduce the error rate associated with the alignment of long sequencing reads. The relatively short read lengths generated for our test set combined with the small and repetitive reference sequence were likely the causes of the alignment issues.

Given the successful alignment using BWA-MEM, we attempted to alter the index for minimap2. Minimap2 utilizes indexing and alignment options to refine the alignment process, but these parameters are not accessible to the user in Tombo. We were able to manipulate two parameters: -k and -w. Reducing -k allows the user to decrease the length of the k-mer to be aligned. -w decreases the window size used to index the reference sequence. By altering the kmer size and window size (reduced to kmer of 3 and window of 3), mapping was drastically improved, although still far reduced compared to BWA-MEM (Table 1).



		Reads mapped		
	Total reads	BWA-MEM	MiniMap2 Tombo default	MiniMap2 Index alt. k=3, w=3
Poly-mtDNA$_{unmod}$	152,072	47,230	597	13,024
Poly-mtDNA$_{mod}$	30,830	1,024	13	82

Table 2: Number of test set reads mapped using BWA-MEM vs minimap2

Out of 152,072 passed reads for poly-synthetic$_{unmod}$ and 30,830 reads for poly-synthetic$_{mod}$, 47,230 (unmod) and 1,024 (mod) were able to be aligned against the reference using BWA-MEM. Initial alignment attempts in Tombo resulted in 597 and 13 mapped reads for poly-synthetic$_{unmod}$ and poly-synthetic$_{mod}$ samples respectively. After manipulating the -k and -w parameters of minimap2, we were able to align 13,024 reads both samples.

After modifying the minimap2 settings, 82 reads were able to be aligned against the reference for the modified sequence (Table 2), which was sufficient for the purpose of viewing the raw data in Tombo. The reason for the poor alignment efficiency in BWA-MEM is unclear; many of the nanopore reads, specifically the longer reads, contained repeated elements (*e.g.*, ATTATTATTATTATT…) of unknown origin. Some reads may not have been well-aligned because of the nature of the reference sequence. Further analysis of the unmapped reads is underway.

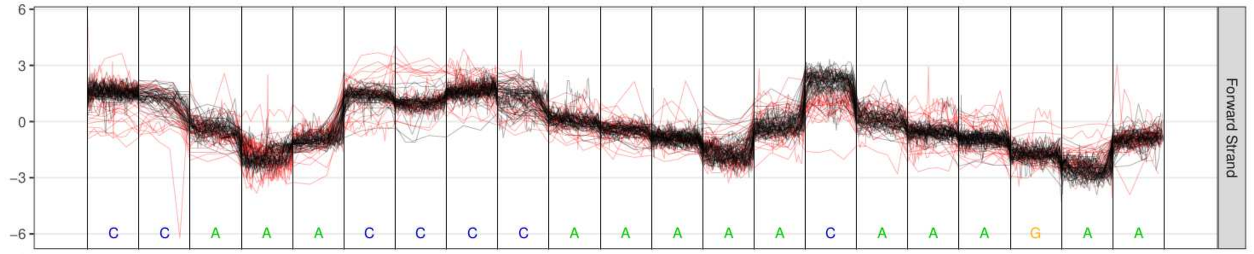


Figure 18: Squiggle plot of region with no modified guanine.

The canonical base model, poly-synthetic$_{\text{unmod}}$ is shown in black and the poly-synthetic$_{\text{mod}}$ mapped reads are shown in red. Little variation in signal is observed between the two.

The basecalled and aligned data shown in Figure 18 displays the canonical base model mapped against the unmodified oligonucleotide sequencing reads. The current distribution qualitatively appears to be similar for both poly-synthetic$_{\text{mod}}$ and the reference, poly-synthetic$_{\text{unmod}}$. This contrasts with the current distribution seen in bases with close proximity to 8-oxoguanine (Figure 19 and 20).

Increased variation is seen in the alternative model when compared to the canonical base model (Figure 19). Significant current shifts, as identified by Tombo, are clear in the raw data for both the modified base itself as well as the three bases preceding and following 8-oxoguanine. The variation for significant current changes indicative of 8-oxoguanine are shown in the box plots below. The regions with most-significant deviation from poly-synthetic$_{\text{unmod}}$ are centered around the modified 8-oxoG bases (for boxplots, see Appendix A).

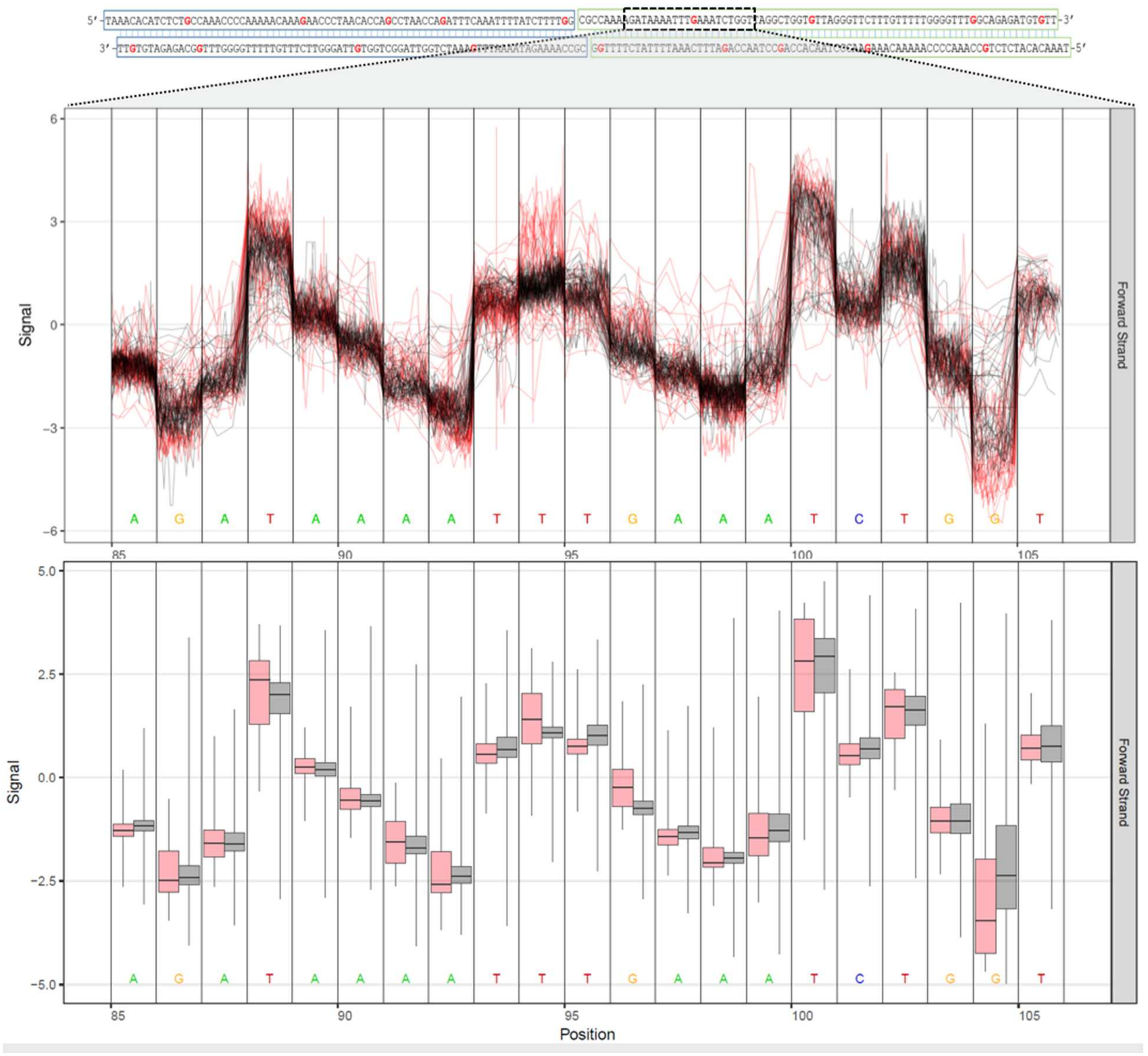


Figure 19: Squiggle plot of significantly modified region around 8-oxoguanine

The squiggle plot (top) showing an excerpt of the raw data aligned by minimap2. The black "squiggles" indicate the change in electrical signal from canonical bases, and serves as a reference. The red "squiggles" indicate the change in electrical signal from the poly-synthetic$_{\text{mod}}$ sample. The bar graphs (bottom) show the variation in signal between the reference (black), and the modified (red) sample. The modified guanine within the sequence is shown in red.

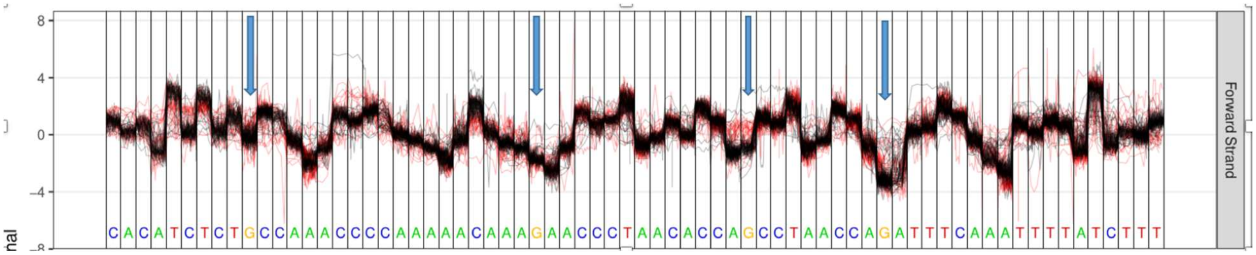


Figure 20: Expanded view of variation between modified and unmodified oligonucleotide

Reference reads are shown in black; poly-synthetic reads are shown in red. The arrows displayed above point directly to 8-oxoguanine locations within the sequence. Resulting signal shifts are attributed to the influence of the modified base.

The expanded view of the squiggle plots for unmodified and modified reads (Figure 20) points to four locations of 8-oxoguanine with corresponding signal changes. Over the 60 base-pair stretch, each of the four modified bases shown are accompanied by numerous significant shifts in current signal. Using these reads, the Tombo algorithm is able to differentiate the canonical base signal from the modified base signal to identify the distinctive signal indicative of 8-oxoguanine.

Phase 2: 8-oxoguanine Alternative Model Development

<u>The Necessity of the Training Set</u>

The test set of synthetic oligonucleotides, while instrumental in confirming the feasibility of the project, was not enough to generate an alternative model for 8-oxoguanine. In order to build an alternative model, Tombo must assess the modified base in every possible context. The modified sequence must be very diverse, and display all k-mer combinations with the modified base over some designated length. Cost limitations prevented us from obtaining a synthetic test set that fulfills these requirements. Therefore, we developed a training set, using mtDNA amplified with

8-oxoguanine to obtain sequencing data capable of generating a canonical and alternative model in Tombo.

Canonical and Alternative Model Training

Both mitoAmp$_{unmod}$ and mitoAmp$_{mod}$ were successfully sequenced using the 1D^2 kit. mitoAmp$_{unmod}$ and mitoAmp$_{mod}$ training samples produced 3644 total reads and 1819 reads, respectively. Of these, 3357 mitoAmp$_{unmod}$ reads and 1582 mitoAmp$_{mod}$ reads mapped using BWA-MEM, respectively. Visualization of mapped reads in IGV (Figure 21) illustrate alignment through the correct genomic region of the rCRS reference sequence..

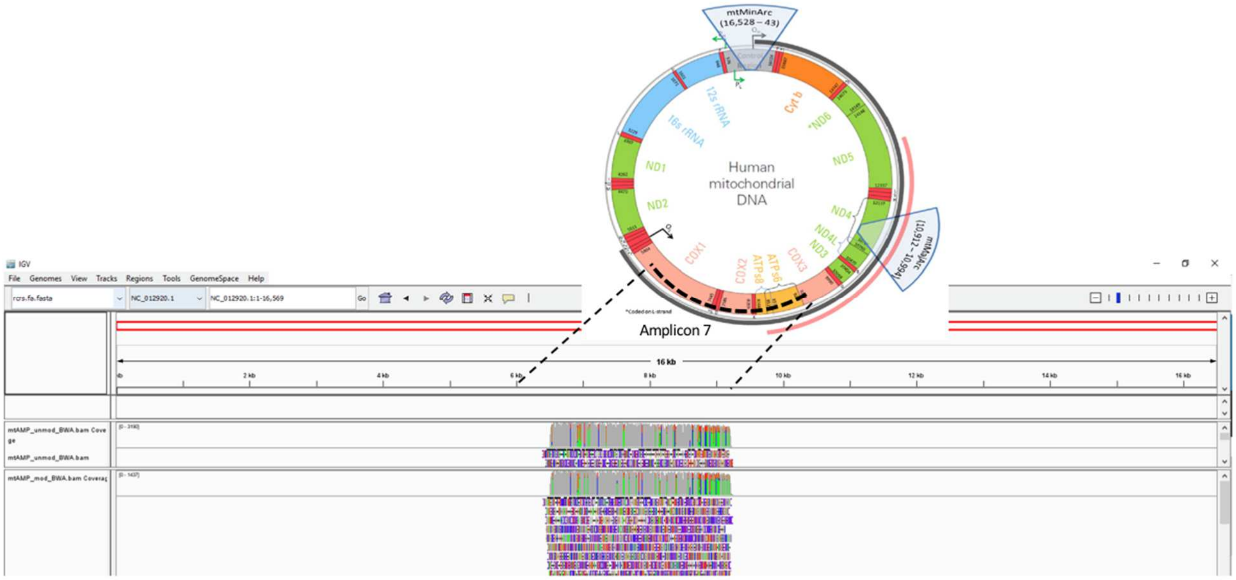


Figure 21: Visualization of mitoAmp$_{unmod}$ and mitoAmp$_{mod}$ reads in IGV.
Of the total reads, 3357 mitoAmp$_{unmod}$ and 1582 mitoAmp$_{mod}$ reads were able to be mapped using BWA-MEM and visualized in Integrated Genomics Viewer. The entire mitochondrial genome is shown above the IGV window, with dotted lines indicated the selected amplicon used for training set generation.

The Albacore-called FAST5 files were resquiggled using a built-in Tombo canonical model and aligned using minimap2. This alignment resulted in 3209 mitoAmp$_{unmod}$ and 1505 mitoAmp$_{mod}$

mapped reads. Using the aligned mitoAmp$_{\text{unmod}}$ data, we were able to build a new canonical base model to use as the reference. The Albacore-called FAST5 files for mitoAmp$_{\text{unmod}}$ and mitoAmp$_{\text{mod}}$ were then resquiggled using the newly-built canonical base model and aligned. The alternative model for 8-oxoguanine was generated from these reads and saved. As with the synthetic oligonuclelotide sequences, we were able to identify regions of the amplicon with significant deviation from the canonical model (Figure 22). Upon completion of model training, the estimated rate of 8-oxoG incorporation was 7.67%, concordant with estimates of the mutagenesis rate of 8-oxoG when amplifiying DNA with *Taq* polymerase [reference] (Figure 23).

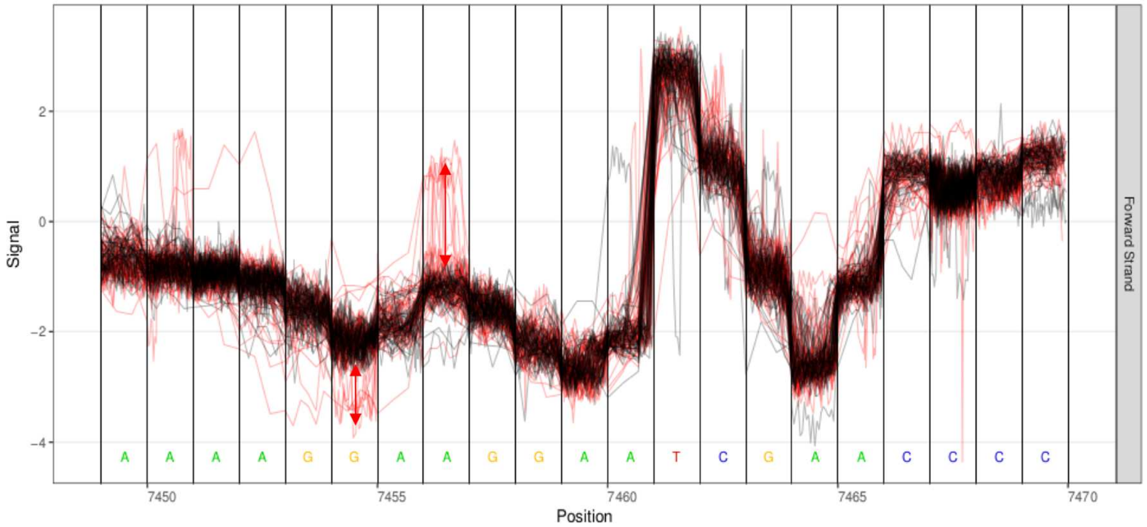


Figure 22: Squiggle plot of mtDNA amplicon 7 modified (red) and unmodified (black).

Clear deviations in signal for the mitoAmp$_{\text{mod}}$ (red) sequence are observed in this squiggle plot of a region of the aligned reads with significant deviation from the canonical model (black). Since 8-oxoG incorporation is random here, we are not sure of the exact location of the oxidized base(s). Red arrows indicate signal variation that differs from the canonical model and may be attributed to the presence of 8-oxoguanine.

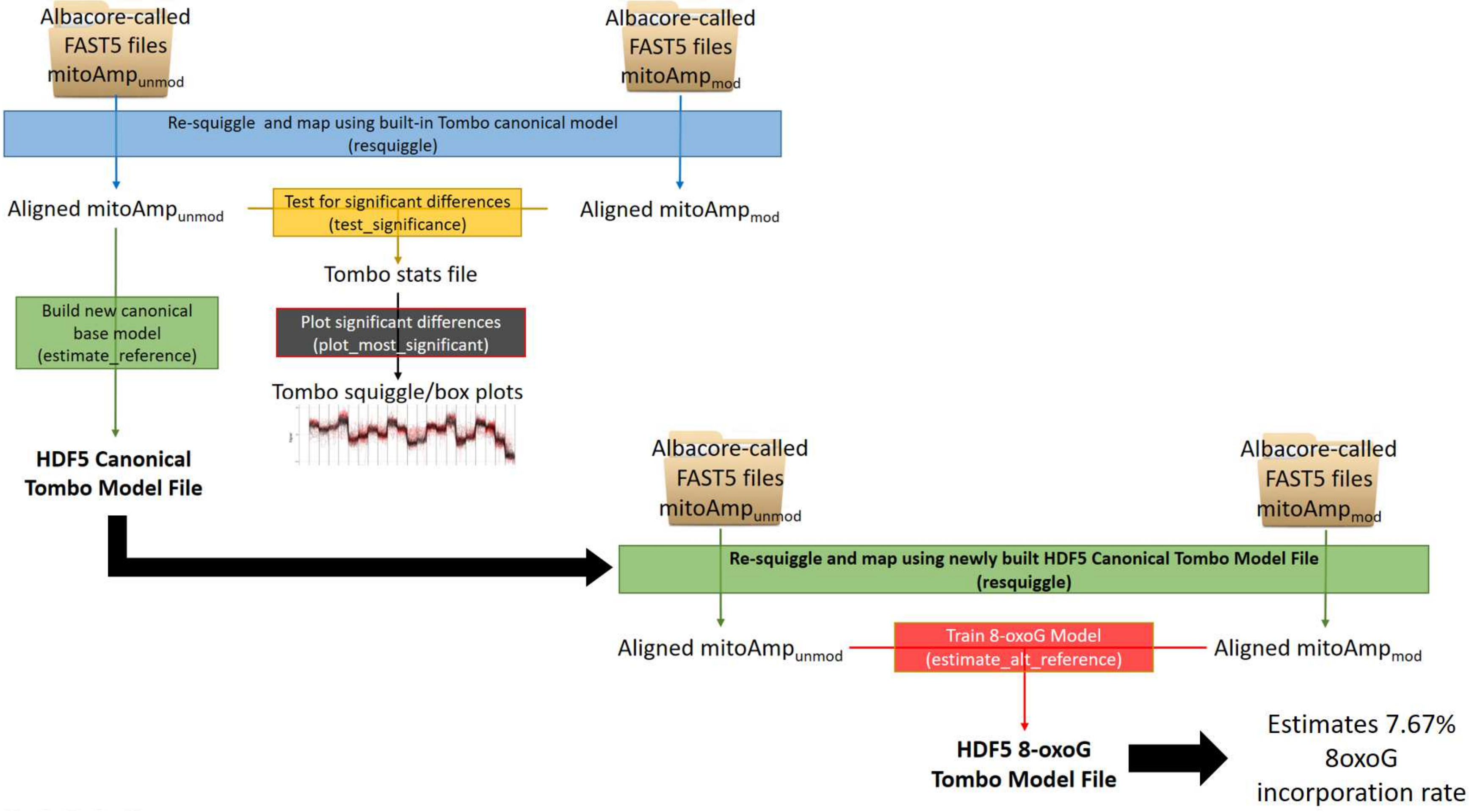

Albacore-called FAST5 files mitoAmp$_{unmod}$
Albacore-called FAST5 files mitoAmp$_{mod}$
Re-squiggle and map using built-in Tombo canonical model (resquiggle)
Aligned mitoAmp$_{unmod}$
Test for significant differences (test_significance)
Aligned mitoAmp$_{mod}$
Tombo stats file
Build new canonical base model (estimate_reference)
Plot significant differences (plot_most_significant)
Tombo squiggle/box plots
HDF5 Canonical Tombo Model File
Albacore-called FAST5 files mitoAmp$_{unmod}$
Albacore-called FAST5 files mitoAmp$_{mod}$
Re-squiggle and map using newly built HDF5 Canonical Tombo Model File (resquiggle)
Aligned mitoAmp$_{unmod}$
Train 8-oxoG Model (estimate_alt_reference)
Aligned mitoAmp$_{mod}$
HDF5 8-oxoG Tombo Model File
Estimates 7.67% 8oxoG incorporation rate

Figure 23.

Overview of pipeline used for generation of alternate basecalling model using Tombo. The estimated 8-oxoG incorporation rate for the mitoAmp$_{mod}$ sample is 7.67%.

Phase 3: Alternative Model Testing

Preliminary testing of the 8-oxoG model indicates successful detection of modified bases. The *in silico* test of 50:50 poly-synthetic$_{\text{mod}}$ and poly-synthetic$_{\text{unmod}}$ per read plot (Figure 24) depicts the reads aligned through the synthetic oligonucleotide reference sequence, bases 1 through 76. Red circles indicate 8-oxoG calls at the respective positions, with the shade of red indicating the likelihood ratio for the assessment, in log likelihood ratio. Black circles indicate canonical, unmodified guanine calls at the position, with the shade of black also indicating likelihood of the call. While not quantitative, there appears to be approximately 50% of calls red (or 8-oxoG), and 50% black (or unmodified guanine), which is concordant with the *in silico* mixture. Further tests with additional ratios as well as quantification tests are currently ongoing.

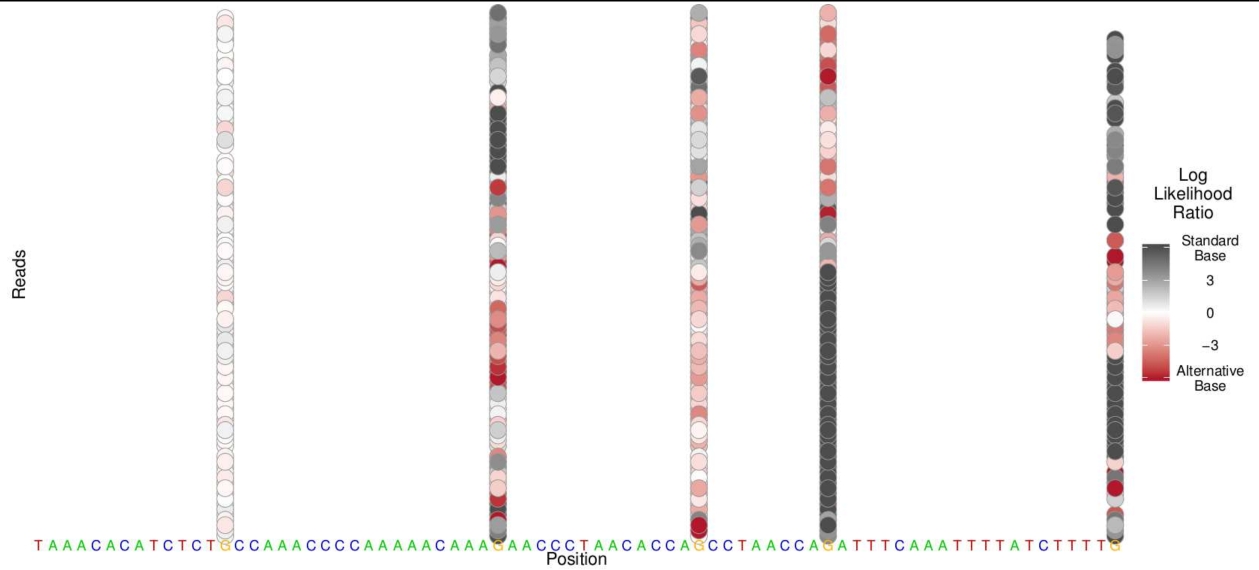


Figure 24. Plots per read for 50:50 *in silico* mixture of poly-synthetic modified and unmodified test samples.

Individual base calls for the mixed reads are indicated at the modified base positions. Alternative base calls (8-oxoG) as ascertained using the trained model for 8-oxoG detection are indicated in red, with the shade indicating the likelihood or confidence in the call. Black circles indicate canonical or unmodified base calls.

CHAPTER IV

CONCLUDING REMARKS AND FUTURE DIRECTIONS

The sensitive and accurate quantification of oxidized DNA has extremely broad applications, and many fields of study will benefit from improved methodology. Oxidized DNA has been acknowledged as a contributor to diseases characterized by inflammation, including preeclampsia, diabetes, and neurodegenerative disorders [1-4]. Cell-free mtDNA has been implicated as a particular stimulant to the body's immune response[55]. Nanopore sequencing of clinically relevant samples may provide an innovative and more sensitive method for detecting oxidative damage within the genome. In addition, it has been proven that several common DNA extraction protocols may induce oxidative damage, typically as a result of harsh lysis steps[56, 57]. This novel sequencing approach may permit quantification of the degree of damage produced by each protocol, as well as identify preferential extraction procedures to be used when experimentally-induced oxidative damage must be minimized.

The preliminary data from this study indicates that 8-oxoguanine can be successfully distinguished from canonical bases using MinION Nanopore sequencing technology. The basecalled reads were able to aligned to the reference sequence using both BWA-MEM and minimap2. The alignments for both poly-synthetic$_{unmod}$ and poly-synthetic$_{mod}$ were examined using Integrated Genomics Viewer; this analysis indicated areas of low consensus for the bases surrounding the 8-oxoguanine substitutions. These low-confidence areas demonstrated a high degree of variability for basecalls, demonstrating the effects of the unique electrical signal disruption caused by 8-oxoG. Raw signal analysis was accomplished in Tombo, which generated squiggle plots based on minimap2 alignments. The squiggle plots and their corresponding bar charts indicate increased local variation in electrical signal between the canonical reference base and the modified base, and regions of statistically significant variation between poly-synthetic$_{unmod}$ and poly-synthetic$_{mod}$ were centered around the 8-oxoG modified bases. Further, a model for 8-oxoG detection was trained using a random base incorporation strategy. Initial tests of the 8-oxoG model indicate that the model can detect modified bases at known positions. Further tests and model optimization is warranted to validate the accuracy and sensitivity of this approach for detecting and quantifying oxidative damage to this particular base.

While this study was, overall, successful, several limitations and difficulties exist which are worth mentioning. Firstly, the software tools used to analyze nanopore data, Tombo in particular, are still in development and are constantly changed updated. The creator of Tombo, Dr. Marcus Stoiber, updated the software numerous times over the course of our data analysis, which complicated our alignment attempts and squiggle plot generation. Secondly, 8-oxoguanine is very difficult to work with, and traditional PCR incorporation attempts have proven unsuccessful. In order to better train the established model, higher rates of 8-oxoG incorporation may be required.

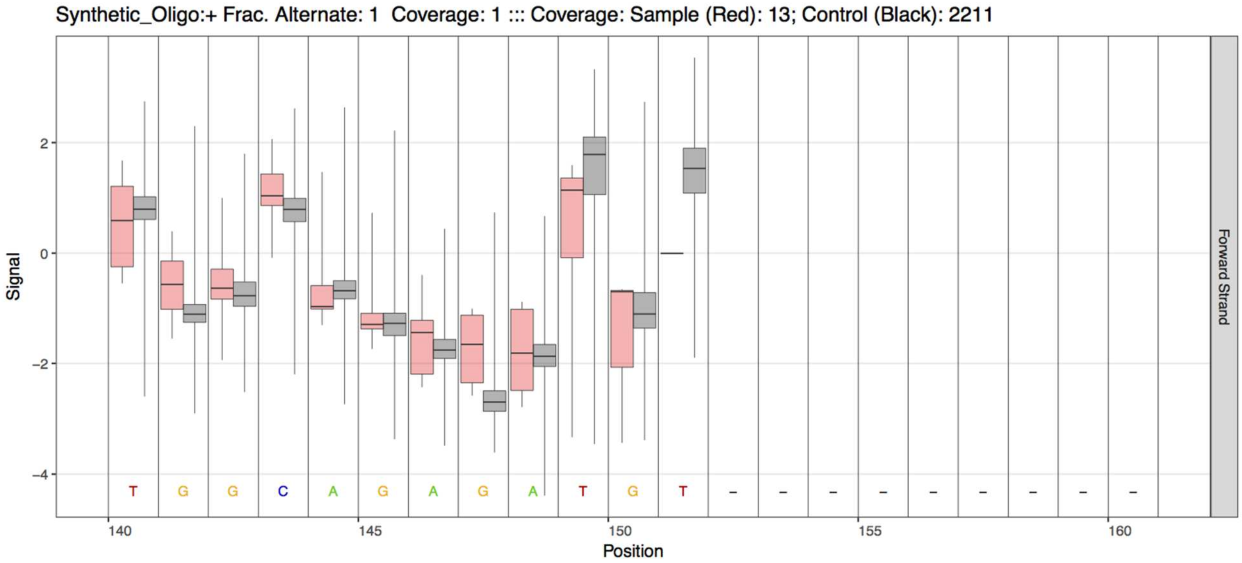
Once validated, this method will be applied to clinically relevant samples for assessment of oxidative damage; these sample types include plasma from pregnant women with preeclampsia as well as plasma and buffy coat DNA from individuals who suffer from diabetes and/or cognitive impairment. Further characterization of oxidative DNA damage in cell-free mtDNA as well as in organellar mtDNA may be indicative of disease risk, state, or progression.

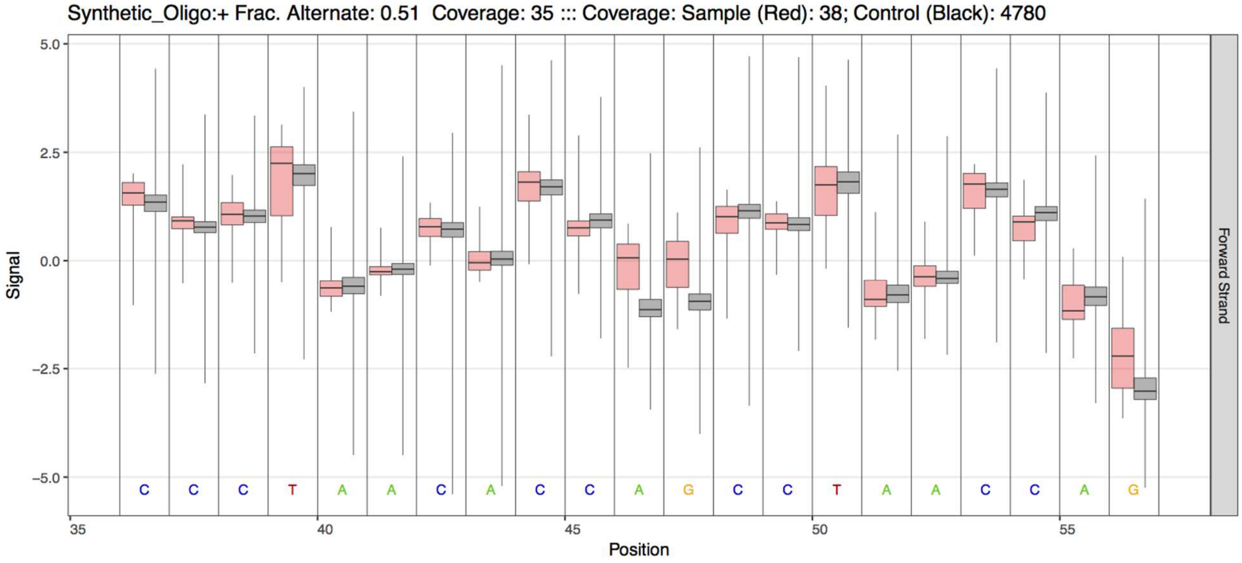
APPENDIX A

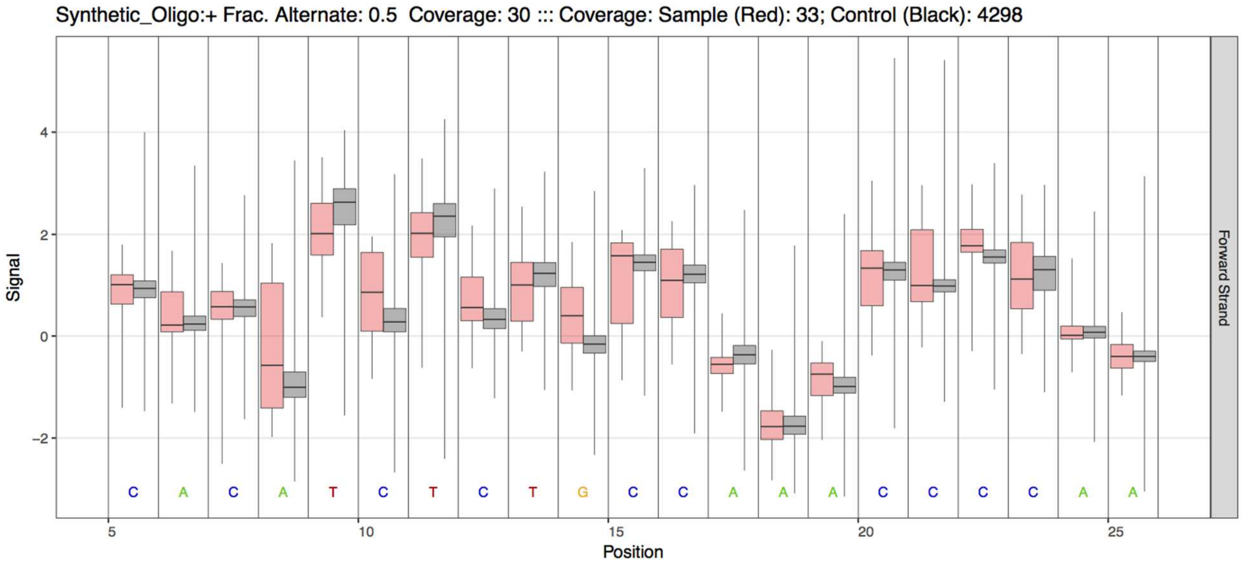
PLOT MOST-SIGNIFICANT RESULTS

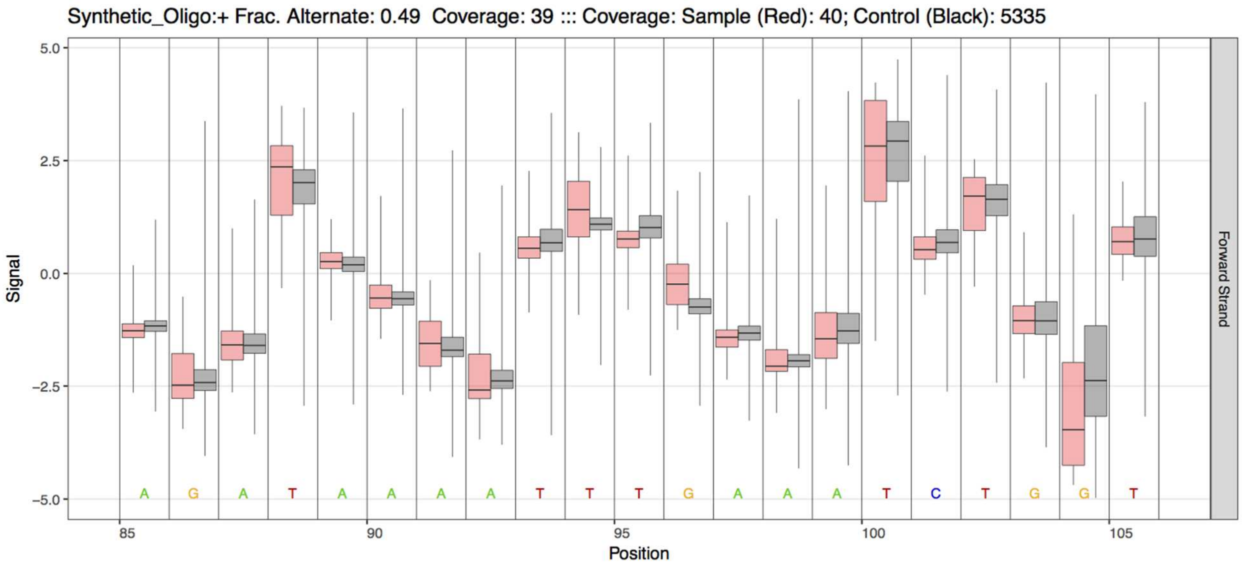
Appendix A: Plot most-significant results

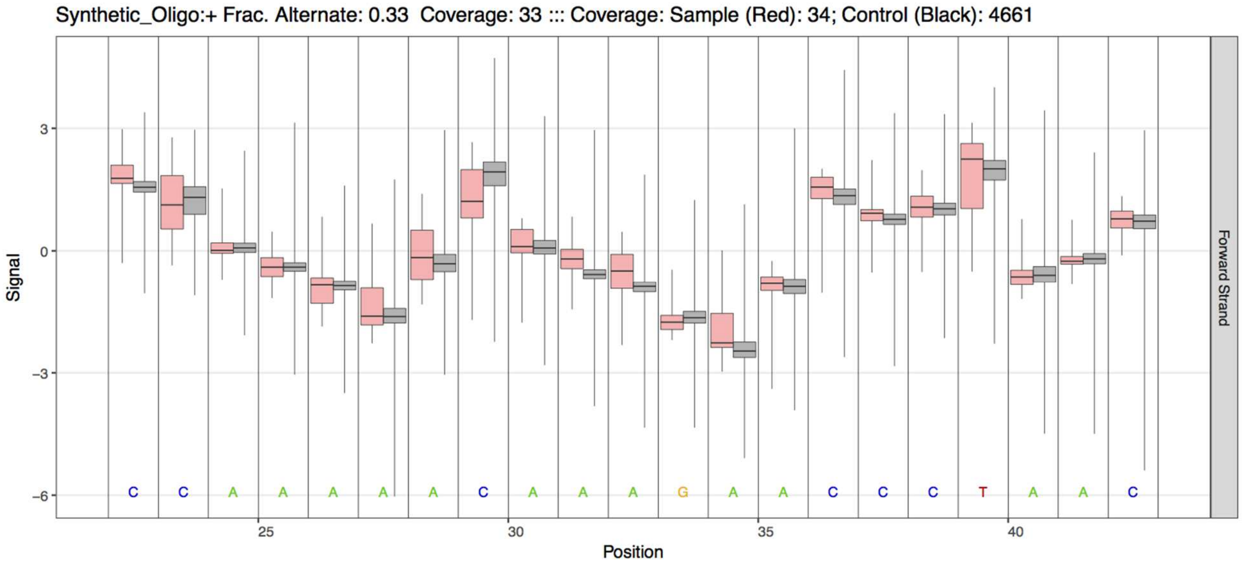
Boxplots indicate variation between reference signal (black) and significant regions of poly-synthetic$_{mod}$ signal (red). Significance is determined using a hypothesis test against a normal distribution estimated from the signal level observed from the control sample reads at each position. A Fisher's method is then used to combine test values over a moving window extending several positions in either direction. This is helpful in locating the exact location of the 8-oxoguanine locations since the signal variations can be observed in any bases within close proximity to the modified base.[58]






Synthetic_Oligo:+ Frac. Alternate: 0.5 Coverage: 30 ::: Coverage: Sample (Red): 33; Control (Black): 4298


Synthetic_Oligo:+ Frac. Alternate: 0.49 Coverage: 39 ::: Coverage: Sample (Red): 40; Control (Black): 5335


Synthetic_Oligo:+ Frac. Alternate: 0.33 Coverage: 33 ::: Coverage: Sample (Red): 34; Control (Black): 4661
Signal
3
0
−3
−6
C
C
A
A
A
A
A
C
A
A
A
G
A
A
C
C
C
T
A
A
C
25
30
35
40
Position
Forward Strand

APPENDIX B

PIPELINE FOR ALBACORE BASECALLING AND ALIGNMENT

Appendix B: Pipeline for Albacore Basecalling and Alignment:

Step 1: raw data--> unpack all fast5
Command:

```
find . -type f -print0 | xargs -0 -I file mv --backup=numbered file .
```

Step 2: Creating a directory for basecaller output
Command:

```
mkdir Olig_modified_fast5basecaller
```

Step 3: Execute Basecaller (Albacore):
Command Example:

```
cd Oligo_modified_fast5basecaller
read_fast5_basecaller.py --flowcell FLO-MIN107 --kit SQK-LSK308 --output_format fast5,fastq
--input /var/lib/MinKNOW/data/reads/20180227_2243_AMB022718qc/fast5_unpacked/ --
save_path ~/Phillips_NanoData/Olig_modified_fast5basecaller/ --worker_threads 4
```

Step 4: Merge Fastq Files:
Command Example:

```
cat *.fastq > ModOligo_02272018_pass_merged.fastq
```

Step 5: Create the reference .fasta File
Command Example:

```
>Synthetic_Oligo
TAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGATTTC
AAATTTTATCTTTTGGCGC
CAAAAGATAAAATTTGAAATCTGGTTAGGCTGGTGTTAGGGTTCTTTGTTTTTGGGGTTTGG
CAGAGATGTGTT
```

Step 6: Index the Reference
Command Example:

```
bwa index /home/lab-
nanopore/Phillips_NanoData/SyntheticDNA_reference.fa
```

Step 7: Align using BWA-MEM
Command Example:

```
bwa mem -x ont2d -t 8 /home/lab-nanopore/Phillips_NanoData/SyntheticDNA_reference.fa
/home/lab-
nanopore/Phillips_NanoData/Olig_modified_fast5basecaller/workspace/pass/ModOligo_02272
018_pass_merged.fastq |samtools sort -o ModOlig_BWAalignment.sorted.bam -T
ModOligo_BWAalignment.tmp -
```

Step 8: Index .bam Alignment File
Command Example:

```
samtools index UnmodOlig_BWAalignment.sorted.bam
```

Step 9: Generate Alignment Statistics
Command Example:

```
samtools stats UnmodOlig_BWAalignment.sorted.bam >
UnmodOligo_BWAalignment.stats.txt
```

Step 10: Import to Integrated Genomics Viewer to look at Alignments and Basecalls

REFERENCES

1. Aguiar PH, F.C., Repolês BM, Ribeiro GA, Mendes IC, Peloso EF, Gadelha FR, Macedo AM, Franco GR, Pena SD, Teixeira SM, Vieira LQ, Guarneri AA, Andrade LO, Machado CR, *Oxidative Stress and DNA Lesions: The Role of 8-Oxoguanine Lesions in Trypanosoma cruzi Cell Viability.* PLoS Neglected Tropical Diseases, 2013. **7**(6): p. e2279.
2. Mark D. Evans, M.D., Marcus S. Cooke, *Oxidative DNA damage and disease: induction, repair and significance.* Mutation Research/Reviews in Mutation Research, 2004. **567**(1): p. 1-61.
3. Pickering, A.M., and Kelvin. J.A. Davies, *Degradation of Damaged Proteins - The Main Function of the 20S Proteasome.* Progress in molecular biology and translational science, 2012. **109**: p. 227-228.
4. Vladimír Palivec, E.P., Isaak Unger, Bernd Winter, Pavel Jungwirth, *DNA Lesion Can Facilitate Base Ionization: Vertical Ionization Energies of Aqueous 8-Oxoguanine and its Nucleoside and Nucleotide.* Journal of Physical Chemistry B, 2014. **118**(48): p. 13833-13837.
5. Xin Yang, X.-B.W., Erich R. Vorpagel and Lai-Sheng Wang, *Direct experimental observation of the low ionization potentials of guanine in free oligonucleotides by using photoelectron spectroscopy.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(51): p. 17588-17592.
6. Crenshaw CM, W.J., Arthanari H, Frueh D, Lane BF, Núñez ME, *Hidden in Plain Sight: Subtle Effects of the 8-Oxoguanine Lesion on the Structure, Dynamics, and Thermodynamics of a 15-Base-Pair Oligodeoxynucleotide Duplex.* Biochemistry, 2011. **50**(39): p. 8463-8477.
7. Kantharaj, D.G.R., *DNA Damage and Repair.*
8. Prof. Tom Brown, D.T.B.J., *Mutagenesis and DNA Repair*, in *Nucleic Acids Book.* atdbio.
9. Fortini, P., Pascucci, B., Parlanti E., D'Errico, M., Simonelli, V., Dogliotti, E., *8-Oxoguanine DNA damage: at the crossroad of alternative repair pathways.* Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 2003. **531**(1-2): p. 127-139.
10. Grigory Dianov, C.B., Jason Piotrowski, Vilhelm A. Bohr, *Repair Pathways for Processing of 8-Oxoguanine in DNA by Mammalian Cell Extracts.* Journal of Biological Chemistry, 1998. **273**: p. 33811-33816.
11. Gold, S.K.S.M.W.S.M.G.I.K.M.P.S.L.A.M.B., *Characterization of DNA with an 8-oxoguanine modification.* Nucleic Acids Research, 2011. **39**(15): p. 6789-6801.
12. Sheila S. David, V.L.O.S., Sucharita Kundu, *Base Excision Repair of Oxidative DNA Damage.* Nature, 2007. **447**(7147): p. 941-950.

13. Yakes FM, V.H.B., *Mitochondrial DNA damage is more extensive and persists longer than nuclear DNA damage in human cells following oxidative stress.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(2): p. 514-519.

14. Hang Cui, Y.K., Hong Zhang, *Oxidative Stress, Mitochondrial Dysfunction, and Aging.* Journal of Signal Transduction, 2012(2012).

15. Cline, S.D., *Mitochondrial DNA Damage and its Consequences for Mitochondrial Gene Expression.* Biochimica et biophysica acta, 2012. **1819**(9-10): p. 979-991.

16. Gedik, C.M., Collins, A., *Establishing the background level of base oxidation in human lymphocyte DNA: results of an interlaboratory validation study.* The FASEB Journal 2005. **19**(1): p. 82-84.

17. Halliwell, B., *Why and how should we measure oxidative DNA damage in nutritional studies? How far have we come?* The American Journal of Clinical Nutrition 2000. **72**(5): p. 1082-1087.

18. Lunec, J.H., K. E.; Jones, G. D.; Dickinson, L.; Evans, M.; Mistry, N.; Mistry, P.; Chauhan, D.; Capper, G.; Zheng, Q., *Development of a quality control material for the measurement of 8-oxo-7, 8-dihydro-2'-deoxyguanosine, an in vivo marker of oxidative stress, and comparison of results from different laboratories.* Free radical research 2000. **33**: p. 27-31.

19. James D. Watson, A.B.a.K.D. *Nobel Laureate: The Future of DNA Sequencing Will Be in the Palm of Your Hand*. 2017; Available from: http://time.com/4971220/future-dna-sequencing/.

20. Kasianowicz JJ, B.E., Branton D, Deamer DW, *Characterization of individual polynucleotide molecules using a membrane channel.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(24): p. 13770-13773.

21. Daniel Branton, D.W.D., Andre Marziali, Hagan Bayley, Steven A Benner, Thomas Butler, Massimiliano Di Ventra, Slaven Garaj, Andrew Hibbs, Xiaohua Huang, Stevan B Jovanovich, Predrag S Krstic, Stuart Lindsay, Xinsheng Sean Ling, Carlos H Mastrangelo, Amit Meller, John S Oliver, Yuriy V Pershin, J Michael Ramsey, Robert Riehn, Gautam V Soni, Vincent Tabard-Cossa, Meni Wanunu, Matthew Wiggin, and Jeffery A Schloss, *The potential and challenges of nanopore sequencing.* Nature biotechnology, 2008. **26**(10): p. 1146-1153.

22. Marcus H Stoiber, J.Q., Rob Egan, Ji Eun Lee, Susan E Celniker, Robert Neely, Nicholas Loman, Len Pennacchio, James B Brown, *De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Sequencing.* bioRxiv.

23. Yanxiao Feng, Y.Z., Cuifeng Ying, Deqiang Wang, Chunlei Du, *Nanopore-based Fourth-generation DNA Sequencing Technology.* Genomics, Proteomics & Bioinformatics, 2015. **13**(1): p. 4-16.

24. *Versatile sequencing library preparation methods for MinION, GridION and PromethION*. Available from: https://nanoporetech.com/resource-centre/posters/versatile-sequencing-library-preparation-methods-minion-gridion-and.

25. T. Laver, J.H., P.A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, D.J. Studholme, *Assessing the performance of the Oxford Nanopore Technologies MinION.* Biomolecular Detection and Quantification, 2015. **3**: p. 1-8.

26. J. Tilanus, M., *The power of Oxford Nanopore MinION in human leukocyte antigen immunogenetics.* Annals of Blood, 2017. **2**(6).

27. Senne Cornelis, Y.G., Lieselot Deleye, Dieter Deforce & Filip Van Nieuwerburgh, *Forensic SNP Genotyping using Nanopore MinION Sequencing.* Scientific Reports, 2017. **7**.
28. Schatz, M.C., *Nanopore sequencing meets epigenetics.* Nature Methods, 2017. **14**: p. 347-348.
29. Jared T Simpson, R.E.W., P C Zuzarte, Matei David, L J Dursi & Winston Timp, *Detecting DNA Methylation using the Oxford Nanopore Technologies MinION sequencer.* Nature Methods, 2017. **14**: p. 407-410.
30. Mohan, U., Kaushik, Shubhangi, Banerjee, Uttam Chand, *PCR Based Random Mutagenesis Approach for a Defined DNA Sequence Using the Mutagenic Potential of Oxidized Nucleotide Products.* Open Biotechnology Journal, 2011. **5**: p. 21-27.
31. Hanes, J.W., Thal, D.M., and Johnson, K.A., *Incorporation and replication of 8-oxo-deoxyguanosine by the human mitochondrial DNA polymerase.* J. Biol. Chem., 2006. **281**: p. 36241-36248.
32. *Personal Correspondence with Karen Juntunen van Delft, Product Specialist, Sigma Aldrich.*
33. *Sigma-Aldrich Custom DNA and RNA Oligos*. [cited 2018; Available from: https://www.sigmaaldrich.com/catalog/product/sigma/oligo?lang=en&region=US.
34. *Integrated DNA Technologies Duplex Buffer Product Page*. [cited 2018; Available from: https://www.idtdna.com/site/order/stock/index/nfd.
35. Prediger, D.E. *Annealing oligonucleotides*. [cited 2018; Available from: https://www.idtdna.com/pages/education/decoded/article/annealing-oligonucleotides.
36. *Thermo Fisher Scientific Qubit dsDNA BR Assay Kit Product Page*. 2018]; Available from: https://www.thermofisher.com/order/catalog/product/Q32850.
37. *New England BioLabs Sticky-end Ligase Master Mix Product Pace*. 2018]; Available from: https://www.neb.com/products/m0370-instant-sticky-end-ligase-master-mix#Product%20Information.
38. *Zymo Research Clean & Concentrator -25 Kit Product Page.*
39. Inc, B.C. *Beckman Coulter Agencourt AMPure XP - PCR Purification Kit Product Page*. [cited 2018; Available from: https://www.beckmancoulter.com/wsrportal/wsrportal.portal?_nfpb=true&_windowLabel=UCM_RENDERER&_urlType=render&wlpUCM_RENDERER_path=%252Fwsr%252Fresearch-and-discovery%252Fproducts-and-services%252Fnucleic-acid-sample-preparation%252Fagencourt-ampure-xp-pcr-purification%252Findex.htm#2/10//0/25/1/0/asc/2/A63880///0/1//0/%2Fwsrportal%2Fwsr%2Fresearch-and-discovery%2Fproducts-and-services%2Fnucleic-acid-sample-preparation%2Fagencourt-ampure-xp-pcr-purification%2Findex.htm/.
40. Technologies, A. *Agilent Technologies 4200 TapeStation Instrument Product Page*. Available from: https://www.genomics.agilent.com/en/TapeStation-System/4200-TapeStation-Instrument/?cid=AG-PT-181&tabId=prod2420037.
41. Agilent Technologies, I. *Agilent Genomic DNA ScreenTape System Quick Guide*. September 2015.
42. *Omega Bio-Tek Mag-Bind Blood & Tissue DNA HDQ 96 Kit Product Page*. Available from: http://omegabiotek.com/store/product/mag-bind-blood-dna-hdq-96-kit-2/.
43. *Takara LA PCR Kit Product Page*. Available from: http://www.clontech.com/VN/Products/PCR/Long_PCR/LA_PCR_Kit.

44. Amanda Ramos, C.S., Luis Alvarez, Ramon Nogués, Maria Pilar Aluja, *Human mitochondrial DNA complete amplfiication and sequencing: A new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification.* Electrophoresis, 2009. **30**(9): p. 1587-93.
45. Technologies, O.N. *Rapid Sequencing Kit Workflow and Product Page*. Available from: https://store.nanoporetech.com/catalog/product/view/id/219/s/rapid-sequencing-kit/category/28/.
46. Barata, C.d.C.B.R., *Comparative genomic analyses of cyanobacteria.* Mestrado em Biologia Evolutiva e do Desenvolvimento, 2017.
47. Technologies, O.N., *1D^2 Sequencing Kit Product Page and Workflow.*
48. *1D^2 Sequencing Protocol*. Available from: https://store.nanoporetech.com/catalog/product/view/id/175/s/1d-2-sequencing-kit/category/28/.
49. Technologies, O.N. *MinKNOW Protocol - Checks and Monitoring*. Available from: https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revah_11apr2016/checks-and-monitoring.
50. Heng, L., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* PREPRINT 2013.
51. James T. Robinson, H.T., Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov, *Integrative Genomics Viewer.* Nature biotechnology, 2011. **29**(1): p. 24-26.
52. *Tombo (1.2.1b) Model Training (Advanced Users Only)*. Available from: https://nanoporetech.github.io/tombo/model_training.html.
53. *Integrated Genomics Viewer Software Home*. Available from: http://software.broadinstitute.org/software/igv/.
54. *IGV User Guide > Viewing Alignments*. Available from: http://software.broadinstitute.org/software/igv/alignmentdata.
55. Boyapati, R.K., Tamborska, A., Dorward, D. A., & Ho, G.-T, *Advances in the understanding of mitochondrial DNA as a pathogenic factor in inflammatory diseases.* F1000Research, 2017. **6**(169).
56. M.T.V. Finnegan, K.E.H., M.D. Evans, H.R. Griffiths, J. Lunec, *Evidence for sensitisation of DNA to oxidative damage during isolation.* Free Radical Biology and Medicine, 1996. **20**(1): p. 93-98.
57. Egil Kvam, R.M.T., *Artificial background and induced levels of oxidative base damage in DNA from human cells.* Carcinogenesis, 1997. **18**(11): p. 2281-2283.
58. *Tombo (v 1.2.1b) Modified Base Detection*. Available from: https://nanoporetech.github.io/tombo/modified_base_detection.html#statistical-testing.