6-10-2020

Toward precision medicine with nanopore technology

Trang Andrea Vu
Rowan University

TOWARD PRECISION MEDICINE WITH NANOPORE TECHNOLOGY

by

Trang Andrea Vu

A Dissertation

Submitted to the
Department of Biomedical Engineering
College of Engineering
In partial fulfillment of the requirement
For the degree of
Doctor of Philosophy
at
Rowan University
May 13, 2020

Dissertation Chair: Jiwook Shim, Ph.D.

Dedication

This dissertation is dedicated to my late grandfathers. Thank you so much for always having faith in me. And to my beloved parents—this journey would not be possible without you

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Dr. Jiwook Shim for his guidance, patience and motivation throughout the course of my Ph.D. study. I would also like to thank my graduate committee members, Drs. Mark Byrne, Peter Galie, Vince Beachley, and Erik Brewer, for their support that goes well beyond the completion of this project. To Dr. Florent Mouliere, thank you for your continuing mentorship and inspiration, even from across the ocean. To Christopher Moran, Joanna Soyring, Julia Borgesi, Joel Goyco, and the many other wonderful students I had the honor to mentor throughout the years, this journey would not be possible without all of your contributions, support, positivity, and humor. I could not ask for a better team.

I would like to give special thanks to Julian Bello, who was there with me since day one of this journey and continues to be a true friend over the years. To Dr. Thomas Christiani, thank you for giving me your unconditional mentorship and friendship, as well as enduring the ridiculous memes I sent every day. I am grateful for all my wonderful graduate friends and colleagues who made this journey a truly unique one.

In addition, I would like to thank my parents and sister for their support and unconditional love. Ba mẹ, cảm ơn vì đã hy sinh vì con và bên con trong suốt những năm qua. Cảm ơn ba mẹ đã ủng hộ con theo đuổi ước mơ và hoài bão. Con thật may mắn khi có cả nhà mình. Hope I made you proud. To all my family and friends, you should know that your support and encouragement were worth more than I can ever express.

Finally, to my partner Ricky, thank you for constantly being my rock, occasionally my bone, and every time my proofreader. This was a long and sometimes bumpy road, but you were always there to support and encourage me.

Abstract

Trang Andrea Vu
TOWARD PRECISION MEDICINE WITH NANOPORE TECHNOLOGY
2016-2020
Jiwook Shim, Ph.D.
Doctor of Philosophy

Currently, when patients are diagnosed with cancer, they often receive a treatment based on the type and stage of the tumor. However, different patients may respond to the same treatment differently, due to the variation in their genomic alteration profile. Thus, it is essential to understand the effect of genomic alterations on cancer drug efficiency and engineer devices to monitor these changes for therapeutic response prediction. Nanopore-based detection technology features devices containing a nanometer-scale pore embedded in a thin membrane that can be utilized for DNA sequencing, biosensing, and detection of biological or chemical modifications on single molecules. Overall, this project aims to evaluate the capability of the biological nanopore, α-hemolysin, as a biosensor for genetic and epigenetic biomarkers of cancer. Specifically, we utilized the nanopore to (1) study the effect of point mutations on *C-kit1* G-quadruplex formation and its response to CX-5461 cancer drug; (2) evaluate the nanopore's ability to detect cytosine methylation in label-dependent and label-independent manners; and (3) detect circulating-tumor DNA collected from lung cancer patients' plasma for disease detection and treatment response monitoring. Compared to conventional techniques, nanopore assays offer increased flexibility and much shorter processing time.

Table Of Contents

Table Of Contents (Continued)

Table Of Contents (Continued)

Table Of Contents (Continued)

List of Figures

Figure Page

List of Figures (Continued)

List of Tables

Table Page

Chapter 1

Introduction

1.1 Cancer and Precision Medicine

Cancer is a group of diseases where abnormal cell growth can potentially spread to the surrounding tissues. It arises from the accumulation of score alterations affecting the structure and function of the genome. Most cancers belong to one of three groups—carcinomas, sarcomas, and leukemias or lymphomas—with carcinomas being accountable for approximately 90% of all human cancers.[1]

In the United States, there are over one million new cancer cases, and more than 500,000 cancer deaths, each year. With only 5-10% of all cancer cases coming from inherited genetic defects, the majority of cancers are caused by environmental exposures and lifestyles, such as viral infection (15%), tobacco use (22%), poor diet (35%), radiation, stress, and environmental pollution.[2]

1.1.1 Molecular markers of cancer. The development of a malignant cancerous tumor frequently results from a multistep process, rather than just a single genetic change.[3] This multistep process originates from various genetic and epigenetic modifications, which exert their pathological effects by causing defects in genes in one of two ways: (1) an enhanced expression or activation of oncogenes (gain in function) and (2) repression or inactivation of tumor-suppressor genes (loss of function).[4] Several methods have been used in research for detecting various cancer-causing factors and different techniques are applied depending on the particular type of cancer.

1.1.1.1 Genetic alterations. In cancer cells, somatic mutations occur and accumulate at a significantly higher rate than in normal cells. These mutations cover a wide range of structural alterations in the DNA. They consist of large chromosomal alterations that encompass millions of base-pairs (e.g. translocation, deletion, or amplification) as well as smaller changes in nucleotide sequences (e.g. point mutations). These different types of genetic alterations often co-exist within a single tumor.[5] Together, they cause defects in DNA repair pathways, cell-cycle regulation, apoptosis, or create error-prone DNA polymerase.[5] The ability to quickly accumulate mutations is critical for cancer cells' rapid growth and development of resistance to cytotoxic cancer treatments.[6]

1.1.1.2 Epigenetic alterations. In contrast to genetic alterations, "epigenetic alterations" refers to all heritable changes in gene expression and chromatin structure that are not encoded in the DNA sequence itself.[5] Epigenetic inheritance, such as DNA methylation, histone modification and RNA-meditated silencing, are essential mechanisms that allow stable propagation of gene activities.[7-8] Disruption of these mechanisms causes inappropriate gene expression (i.e. activation of oncogenes or inactivation of tumor-suppressor genes) that can lead to cancer development.[9]

A main distinguishing feature of epigenetic alterations compared to genetic changes is that they are reversible, and thus, present an exciting opportunity for the development of novel strategies for cancer prevention.[5] Overall, both genetic and epigenetic alterations are involved in the formation of a field for cancerization (i.e. cancers develop at multiple foci in a tissue). However, their exact contribution is dependent on the major carcinogens involved and their carcinogenic mechanisms.[10]

1.1.2. Toward precision medicine for cancer treatment. Nowadays, when a patient is diagnosed with cancer, they often receive a combination of different treatment approaches to combat the disease, including surgery, chemotherapy, radiation therapy, immunotherapy, or any combination of the foregoing. The type of treatment and dosage are often similar between patients with the same type and stage of cancer. Even so, different people may respond differently. This is because a tumor's genetic changes, which cause cancer to grow and spread, can differ significantly among patients with the same cancer. On the other hand, different types of cancers can also share the same cancer-causing changes.

Precision medicine is an approach that matches patients to therapies depending on their own genomic information, to maximize a patient's response to treatment.[11] Using the concept of massively parallel sequencing, the development of high-throughput Next-generation sequencing (NGS) devices (e.g. Illumina, Complete Genomics, and Roche Applied Science 454) have revolutionized personalized genomic medicines. In fact, for the first time, several clinical samples of patients were obtained and analyzed with NGS, providing an insight into the high complexity of diseases such as cancer.[11-19] However, up-to-date, clinical trials demonstrate that the level of patient benefit from precision medicine in cancer treatment is still low.[11, 20-23]

A possible explanation for this unexpected result is that most current efforts are based on the idea that cancers are just like other rare genetic disorders. Specifically, the presence of epigenetic changes, lineage-specific drivers and non-oncogene-driven vulnerabilities are often ignored.[11] Over 90% of cancer death are caused by metastatic spread of cancer to distant sites.[24-26] During tumor manifestation and progression, treatment

decisions are generally based on results of primary tumor specimens from diagnostic testing. However, this method neglects a tumor's genetic and epigenetic evolutionary changes, which are shown to present a major challenge for treatment selection.[27-30] Thus, longitudinal monitoring of a tumor's genetic and epigenetic variability are essential for correct treatment selection and cancer eradication.[31]

While it is impractical for patients to undergo serial surgical biopsy along tumor progression, liquid biopsy provides a minimally invasive and more sustainable alternative to interrogate cancer cells longitudinally. Liquid biopsy is the sampling of non-solid biological tissue, to extract tumor-derived materials (e.g. circulating tumor cells, circulating tumor DNAs, and exosomes).[31] Up-to-date, plasma-derived circulating tumor DNA (ctDNAs) are the most commonly used blood-based biomarker in clinical practice, and are utilized most frequently to (1) monitor a patient's response to therapy, (2) detect minimal residual disease along treatment, and (3) assess therapy resistance development.

Currently, the majority of liquid biopsy samples are analyzed with NGS, which can be costly and require extensive processing times of up to several weeks, or digital PCR that are restricted to a single locus (a fixed position on a chromosome). Furthermore, NGS technologies employ several enrichment, amplification, and labeling steps, such as polymerase chain reaction (PCR) and bisulfite conversions, which can cause the performance to be time and cost intensive, and can also increase the possibility of false positive results.[32] Due to the need for a label-free, high throughput system, there has been a growing interest toward using third- and fourth-generation sequencing, specifically nanopore technologies, in cancer prevention and detection in the past decades.

1.2 Nanopore Technology

Since its development and publication in 1996, the nanopore has become an emergent and powerful technology for a direct and inexpensive method for DNA sequencing, biosensing, and detecting biological or chemical modifications on single molecules, as well as the kinetics of DNA and protein folding.[33-37]

1.2.1 Definition of nanopore technology. Formal definitions of nanopore technology typically feature devices that contain a nanometer-scale pore embedded in a thin membrane. Originating from the Coulter counter and ion channels, nanopore-based devices can detect various charged biomolecules that are slightly smaller than the diameter of the pore. In the nanopore-based analysis, a biological or a solid-state membrane separates the experimental chamber into two compartments, referred to as the *cis* and *trans* sides, to which a cathode and anode are attached, respectively. Negatively charged biomolecules, such as DNA, are then introduced into the *cis* side of the chamber. Under the electrophoretic force exerted by the external current, the biomolecule transports through the nanopore to the *trans* chamber. As the molecule moves through the nanopore, it interrupts the current signal, causing ionic current blockages. Physical and chemical properties of the targeted molecule can be analyzed using the amplitude and duration of current blockages through the nanopore.[38-39]

1.2.2 Types of nanopore. The two main types of nanopore—biological and solid-state nanopores—can be obtained or fabricated in numerous ways[40-45] and offer a wide range of biomolecule detection. Biological nanopores are secreted from different bacteria, of which the two most popular types come from *α-hemolysin* and *MspA* porin. These biological nanopores are then usually inserted into different biological substrates, such as

a phospholipid bilayer, liposomes, or polymer films. Biological membranes are structurally well-defined and easily reproducible. Biological nanopores are mostly used for the detection of single-stranded DNA (ssDNA), microRNA (miRNA), and disease diagnostics.[39] Most solid-state nanopores are fabricated in membranes made of silicon oxide (SiO), silicon nitride (SiN_x), hafnium oxide (HfO_2), graphene, aluminum oxide (Al_2O_3) and hybrid materials.[46-48] With controllable pore size and membrane thickness, solid-state nanopores have been beneficial for use in RNA sequencing, single-stranded and double-stranded DNA sequencing, DNA-protein complex detection, and other biomolecule detection.

1.2.3 Experimental setup. Detailed reviews on the experimental setups and uses of the nanopore technology have been described in literature before.[39-40, 47, 49-51] Figure 1A shows a schematic of the biological nanopore experiment setup. Briefly, a Teflon film divides the testing chamber into two separate compartments, named *cis* and *trans* chambers, to which negative and positive electrodes are attached, respectively (as described in 2.1 above). When α-HL protein is introduced to the *cis* chamber, it can insert a mushroom-shaped channel into lipid bilayer, connecting the *cis* and the *trans* sides (Figure 1B). α – hemolysin (α-HL), a toxin released by the *Staphylococcus aureus* bacterium, is well known for its ability to form a protein channel with a well-defined structure as well as dimensions. Specifically, the α-HL α–hemolysin is a water-filled, mushroom-shaped channel consists of a *cis*-opening (2.6nm), a nano-constriction (1.4nm), and a *trans* opening (2.0nm).[52] DNA or any other charged biomolecules are then injected to the *cis* chamber. Under an electrophoretic force, charged molecules (e.g. DNA) can translocate through the nanopore, resulting in a transient current blockage with

characteristic amplitude (%I/Io), dwell time (Δt), and event signature. Depending on the specific sequence, ssDNA strands can adopt several common conformations, including hairpin, G-quadruplexes and i-motif (among many others).[47, 53-55] These secondary structures, while different, can coexist at the same time in the solution, causing various types of current-blockages as shown in Figure 1C. Depending on the size, conformation, and kinetics of each biomolecule species, several types of current blockage can be observed. These current blockages are characterized by their specific amplitudes and dwell times, allowing different biomolecules to be distinguished.

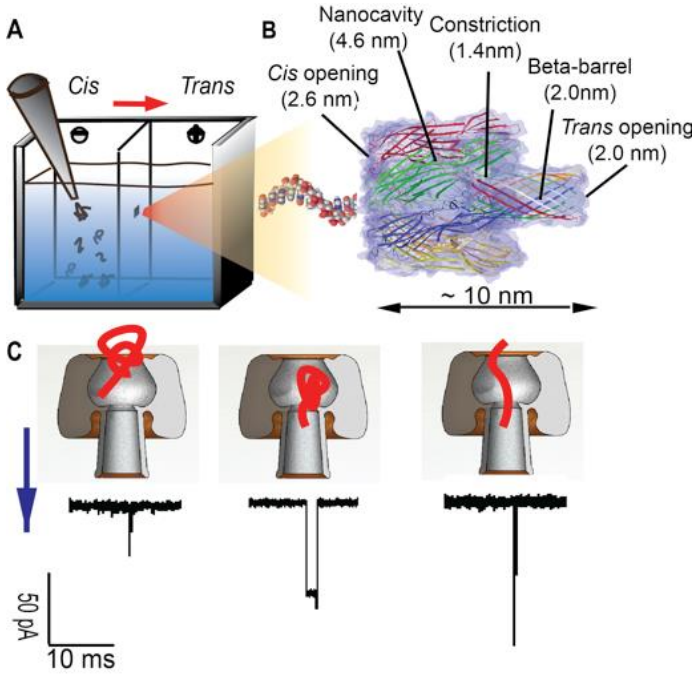


Figure 1. The experimental setup of α-hemolysin nanopore. The experimental setup of the $\boldsymbol{\alpha}$-HL nanopore experiment. (A) The testing chamber contains two compartments, cis and trans, separated by a Teflon film. (B) $\boldsymbol{\alpha}$-HL, a mushroom shape nanopore which connects the cis and trans chambers. (C) When DNA molecules, or other charged biomolecules, are inserted into the cis chamber, the applied current forces the biomolecules through the $\boldsymbol{\alpha}$-HL nanopore, causing a current-drop (termed a blockage). Current blockages are analyzed and characterized by their amplitudes and dwell times. Each blockage trace is unique to the biomolecule species passing through the channel, depending on the size, conformation and kinetics of the interaction between the biomolecule and the nanopore.

1.2.4 Nanopore technology for precision medicine. Nanopore technology offers many advantages that NGS devices are incapable of: for instance, the nanopore has demonstrated the ability to detect *CpGs* methylation (one of the earliest epigenetic biomarkers in cancer hallmarks) without the need of PCR amplification and bisulfite conversion.[56-57] Thus, nanopore technology strives to be a potential genomic tool that is label-free, has a high throughput, a small sample volume requirement, flexible runtime, and minimal footprint.[39] However, despite the past twenty years of significant progress in single molecular sequencing and analysis, nanopore technologies have not yet been translated into even distantly comparable advances in clinical settings. The aim of this thesis is to explore the ability of α-hemolysin nanopore to detect genetic and epigenetic changes on DNA, specifically ctDNAs. Through studying physical and electrical behavior of short DNA fragments, we can enhance the nanopore readout resolution, allowing the detection of DNA modifications. We believe that α-hemolysin nanopore is a good candidate, as the pore size stays consistent, yielding replicable results**.**

1.3 Thesis Outline

As described above, despite its significant potential and technological advancements in the past decades, applications of nanopore technology in clinical settings still come up short. The **objective of this research** is to explore and enhance the ability of α-hemolysin nanopore to study DNA fragments for cancer detection and precision medicine. To establish this purpose, we developed protocols to optimize the ability of both lab-based and commercially available nanopore assays in detecting DNA genetic and epigenetic alterations. A more detailed description of the studies performed in this thesis is listed below.

In **Chapter 2** we provide an overview of the recent developments and applications of nanopore technology in studying cancer biomarkers. The development and progression of cancer are influenced by numerous genetic and epigenetic factors (e.g. structural variants, transcription factors, G-quadruplexes, methylations, histone modifications and expression of microRNAs). While nanopore technology has grown rapidly as a versatile DNA sequencer,[39-40, 49, 58-61] its potential as a biosensor for cancer markers is still significantly underdeveloped.

DNA is a flexible polymer that can adopt various structures depending on the sequence and surrounding environment. It is essential to evaluate the ability of α-hemolysin nanopore as a biosensor to detect and distinguish these structural changes. Therefore, in **Chapter 3**, we set out a study of the structural changes of an i-motif sequence with the surrounding environment's pH level. Our results demonstrated that at high ionic concentration, C-rich DNA can dynamically fold into an i-motif or compacted structure according to the solution pH. These secondary structures, whose sizes were different, generated distinct electrical signal events when captured by the nanopore.

The most appealing strength of biological nanopores, such as the α-hemolysin, is their consistent dimensions and size selectivity. As the α-hemolysin nanopore only allows single-stranded DNA to translocate, all secondary structure must first dissociate into a linear strand at different rates, depending on its specific structural stability. In **Chapter 4**, we performed a comparative study on the formability and stability of Ckit1 G-quadruplex in K^+ and Na^+ environments. We further evaluated the effect of CX-5461 (a potential cancer drug) on stabilizing Ckit1 G-quadruplex structure. Using the event translocation time obtained from nanopore study coupled with K-nearest neighbor clustering analysis,

we found that cations have a significant impact on the G-quadruplex structural stability and volume. Our results demonstrate that combining machine learning classification with nanopore results would allow different G-quadruplexes to be distinguished with over 80% accuracy and sensitivity.

While CX-5461 has an amplified effect on unmodified *C-kit1* DNA sequence to halt cancer cell growth, activating mutations on *C-kit1* have been observed in several types of malignancy, notably leukemia, melanoma, and gastrointestinal tumors.[62-64] Thus, in **Chapter 5**, we set out to evaluate the stability of G-quadruplex structure formed by mutated *C-kit1* DNA sequences and the altered efficacy of CX-5461 on these structures. We designed the sequences containing between zero and six mutated guanines, to partially or completely disrupt the G-quadruplex. In this study, we found that while G-quadruplex stability depends strongly on the number of mutations present, the position of mutated guanine was the main deciding factor for CX-5461 efficacy.

Another type of epigenetic marker we focused on in this thesis is methylated cytosine, which is among the most studied epigenetic markers for cancer diagnosis, therapeutics, and prognosis.[65-69] Despite the enormous number of studies and papers published on promoter methylation as potential cancer biomarkers, currently, there are only two FDA-approved, DNA methylation-based biomarker assays available on the market. A major challenge in the study of epigenetics is that most commonly used techniques, including bisulfite sequencing, cannot directly identify 5-methylated cytosine (5-mC) from native DNA. In **Chapter 6**, our study evaluates the feasibility of label-free detection of methylated cytosine using the nanopore. Through employing a salt concentration gradient between the two experimental chambers, we were able to decrease the DNA mobility and

velocity through the pore; thus, increasing the detection resolution. This method allowed us to distinguish methylated from unmethylated cytosine in mixtures with different composition.

In **Chapter 7**, we employed methyl-binding protein (MBD2) as a specific tag, as a method to enhance methylation detection with nanopore technology. Label-free detection, while attractive for its simplicity, can be affected greatly by the local context of the DNA sequence. Conjugation of MBD2 to methylated DNA resulted in a significant increase in both event translocation time and current blockage amplitude. Coupling nanopore data output with commonly employed machine learning algorithms, we were able to distinguish methylated DNA from unmethylated DNA with optimal sensitivity and specificity of 88% and 86%, respectively.

Having determined the sensitivity of nanopore technology, we evaluated the feasibility of applying nanopore technology in a clinical setting. Recent studies found that circulating-tumor DNA in the bloodstream exhibits the same genetic and epigenetic alterations as the solid tumor they originated from,[70] and thus ctDNA can reflect such tumor type and stage. However, as circulating-tumor DNA are highly fragmented and present in low concentrations in blood circulation, it is still challenging for nanopore assay to effectively capture them. In **Chapter 8**, we developed and evaluated two novel protocols that could be integrated: commercially available nanopore assay for sequencing and analyzing circulating-tumor DNAs. We employed rolling-circle amplification and blunt-end ligation methods to create long DNA fragments that could be captured by the nanopore.

Finally, in **Chapter 9** we present a summary of our findings and discuss the potential directions and future applications.

Chapter 2

Piecing Together the Puzzle: Nanopore Technology in Detection and Quantification of Cancer Biomarkers

2.1 Abstract

Nanopore-based technology has grown rapidly in recent years, which have seen the wide application of biosensing research to a number of life sciences. In this chapter, we present a comprehensive outline of various genetic and epigenetic causal factors of cancer at the molecular level, as well as the use of nanopore technology in the detection and study of those specific factors. With the ability to detect both genetic and epigenetic alterations, nanopore technology would offer a cost-efficient, labor-free, and highly practical approach to diagnosing pre-cancerous stage and early-staged tumors in both clinical and laboratory settings. This chapter was adapted from our previously published paper (Vu, Trang, et al. "Piecing together the puzzle: nanopore technology in detection and quantification of cancer biomarkers." *RSC advances* 7.68 (2017): 42653-42666.)2.2 Detection of Cancer Biomarkers

2.2 Current State of the Art

2.2.1 Structural variants. Structural variants (SVs) are one of the first recognized causal factors of cancer. A structural variant is a form of somatic DNA mutation, whereby the SV promotes the development and progression of cancer while contributing to all the important hallmarks of the instability in cancer genomes.[71] The four main types of SVs are large deletions, amplifications, inversions and translocations of nucleotides within a DNA sequence. They are often responsible for the creation of fusion genes, copy number, and other regulatory changes that lead to activation or overexpression of oncogenes, as well as inactivation of tumor suppressor genes.[72-76] In many cases, different SVs occur

simultaneously in a specific pathway that amplifies their genetic effects on cell instability. For example, with head and neck cancers, it was found that when the deletion of CDKN2A and amplification of CCND1 happen together, there is a higher risk of recurrence, metastasis, and death rather than when either genetic alteration occurs alone. [62-63]

SVs are important indicators of human cancers.[64-67] Complex SVs have been found to cause approximately half of nucleotide deletions in pancreatic ductal adenocarcinoma (PDAC).[68-70] Furthermore, *CDKN2A/p16* and *SMAD4/DPaazC4* have been identified as two of the most common deleted tumor suppressor genes. The ability to detect these mutations is critically important to the healthcare industry, allowing the monitoring of cancer patients for early detection of possible relapse.[67-71] In mammalian cells with highly repetitive genomes, studies of SVs frequently use a resequencing approach, in which the read from the target genome is independently aligned from the reference genome to search for SVs.[72] In general, besides specificity and sensitivity, when detecting SVs, a method's quality is further judged by its ability to accurately predict breakpoint locations, the size of variants, and changes in copy count.[67, 73]

As shown in Figure 2, Norris et al. demonstrated the value of detecting long SVs using Oxford MinION™, to detect a series of well-characterized SVs, including large deletions, inversions, and translocations that inactivate the *CDKN2A/p16* and *SMAD4/DPC4* tumor suppressor genes in pancreatic cancer.[67] Using Oxford Nanopore barcodes, the Norris et al. produced libraries for all 12 PCR amplicons in one run, yielding reads with PHRED scores of 10.9-11.50. PHRED, invented back in 1998 by Ewing and Green, was originally a base-calling program for automated sequencer traces. In later research, the term "PHRED score" has been used for the determination of quality and

accuracy between consensus sequences. The higher the PHRED score, the higher the accuracy. For example, a PHRED score of 10 stands for a 90% base call accuracy, and a PHRED score of 20 is correlated with 99% base call accuracy.[74] For this specific study, the readings were averaged at 640 bps long with a PHRED score of 11.50. It was also found that these reads are consistent for the entire bp length. The amplicons mapped with an overall percentage of 99.6% for regions of *hg19*, while 79% of aligned reads accurately matched to bases. Notably, the representation of amplicons does not change accuracy based on the complexity of the sequence. Additionally, the researchers wanted to test their method with low frequency SVs. In a 1:100 dilutions, the run produced 4,058 2D reads from 270 of 512 channels. The average read length was 650 bps and had a PHRED score of 10.9. Overall, the researchers proved their methods can be conducted in a timely manner. For the two sequences (*CDKN2A/p16* and *SMAD4/DPC4*) in this study, it took 15 minutes and 33 minutes respectively, to generate 450 reads.[67] In comparison, 2nd generation sequencers can generate millions of reads simultaneously, but it can take hours to days to complete. The experiment indicated the ability of nanopores to serve as a reliable and efficient method of sequencing, allowing rapid detection of tumor-associated structural variants. The two limitations of MinION™, as noted by the researchers, were (1) a relatively high mismatch and index error rate and (2) a limited yield (on the scale of Mb or Gb).

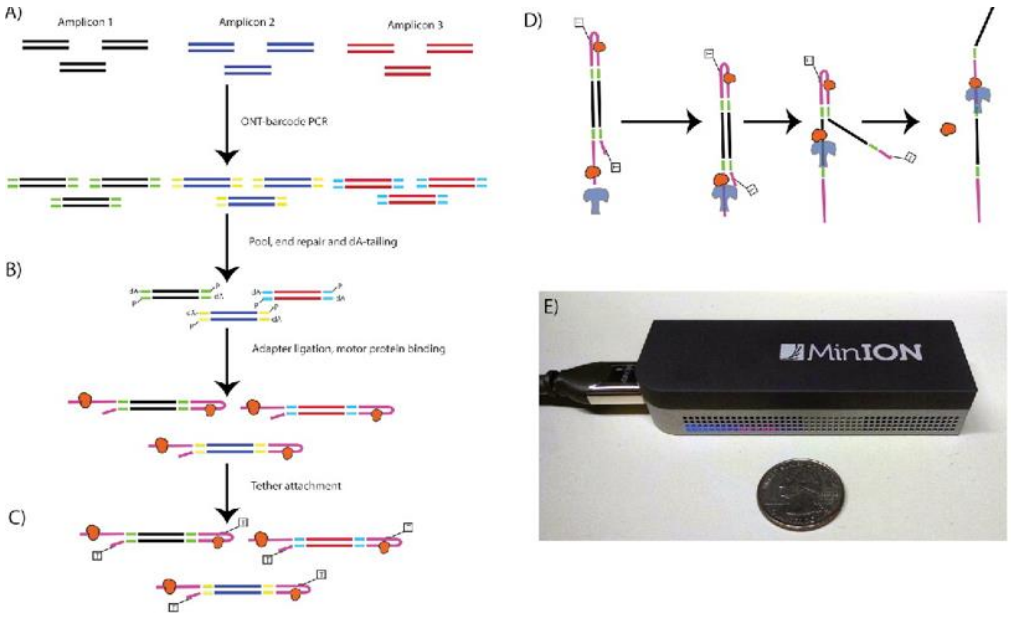


Figure 2. Nanopore Library Prep Workflow. Oxford Nanopore barcodes were incorporated into amplicons by PCR- individually for each SV, then resultant reactions were pooled (A). After NEB End Repair and dA-tailing modules (B), hairpin and leader adapters were ligated on, each containing a motor protein. Only the hairpin protein contained a his-tag, which was used to enrich for molecules containing a leader adapter and his-tag (his-tag selection step not shown). Tether attachment (C) allowed for direct attachment of the molecules to the flow cell membrane. Within the MinION flowcell (D), DNA molecules are pulled through a protein pore (blue), with motor protein (orange) affecting speed of DNA translocation through the pore. One side of the DNA molecule is read, then the hairpin, then the second side. Both reads were aligned to produce a 2D consensus read.

Compared to conventional genome-based methods, such as fluorescence *in situ* hybridization (FISH), fiber-FISH, array comparative genomic hybridization (aCGH) and paired-end mapping (PEM) (each of which have a read length of approximately 35 ~ 400 base pairs (bps)),[73, 75] nanopore allows for much more flexible read lengths (of a few bps to kbps). However, the average PHRED score of reads generated by MinION is still relatively low compared to other sequencers (e.g. Illumina, 454, Ion Torrent, PacBio, and others). At the moment, Illumina is the most popular DNA sequencer on the market. Still, depending on the equipment model and sample size, sequencing using Illumina can take

from 3-12 days to complete. Additionally, the current market price of Illumina ranges from $50,000 (MiniSeq) to over $6M (Illumina HiSeq X Five), costing significantly more than the nanopore-based sequencers.

2.2.2 Transcription factors. The second most well-known causal factor of cancer is aberrant activity of transcription factors (TFs), which are often members of multigene families with common structural domains.[3] TFs are the main regulators of gene expression and signaling pathways in all biological systems and bind to a specific sequence of DNA to promote or inhibit gene expression. In cells, a major portion of oncogenes and tumor-suppressor genes are encoded by TFs.[4, 76] Aberrant TF activity can occur due to changes in expression, protein stability, protein-protein interactions, post-translational modifications, and numerous other mechanisms.[77] In a healthy cell, upstream transcriptional regulators highly regulate all genes with similar functions. However, changes in TF activity leads to deregulation of genes involved in promoting cancer cell proliferation, survival, and inducing angiogenesis and metastasis of tumors.[4, 76] For example, nuclear TFs, the signal transducer and activator of transcription, has been linked to various human cancer cell lines and primary tumors, including leukemia and lymphoma, as well as breast, lung, pancreas, and prostate cancers.[63, 79-87]

Various direct and indirect techniques have been used to characterize TFs, along with other sequence-specific DNA binding proteins, including electrophoresis, electrophoretic mobility shift assay, nuclear magnetic resonance, X-ray crystallography, atomic force microscopy, optical tweezers, and direct fluorescent visualization, among others.[78, 88-92] However, most of these methods require some combination of chemical cross-linking between TFs and DNA, modification or tagging of the TF and DNA, and

amplification assays. Furthermore, due to the complicated requirements, these methods would lack the ability to resolve fine details of the TF and DNA complex (i.e. partial *versus* full binding of the TF domains to DNA).[78] The specific mechanism of TFs binding to DNA sequences is still under intensive study and is a major area of interest in molecular biology.[78, 93]

Squires et al. used solid-state nanopores as biosensors for the characterization of DNA, RNA, and proteins (Figure 3). With the use of an electric field, the researchers could guide the polymers through a nanopore and identify individual molecules. The current-blockage patterns generated during translocation of charged molecules provides an abundance of information about TF local properties, as well as TF-DNA interactions.[78] As previously noted, the regulation of TFs has not been well investigated, hence the use of solid-state nanopores could be a novel technique in describing these molecular interactions. As proof of technique, the Squires et al. has shown that their nanopores can distinguish between specific and nonspecific binding of TF, by analyzing the ion current of the canonical zinc-finger DNA-binding domain of Early Growth Response 1 (*zif268*). Characterization of the *zif268* was accomplished using the distinct blockage patterns of the current within the nanopore.[94] Through analyzing the data, the researchers found that there are three main types of blockages, existing mostly in five distinct patterns rather than randomly. These patterns have a direct correlation to preexisting data. Hence, the nanopore presents great potential in characterizing DNA complexes because of its ability to detect complex structures and protein conformations, with the possibility of removing TFs as needed. Squires et al. note that their nanopore sensor can identify small TFs in DNA as

well as distinguish between specific and nonspecific binding. This research technique allows information-gathering availability with respect to TF-DNA interactions.

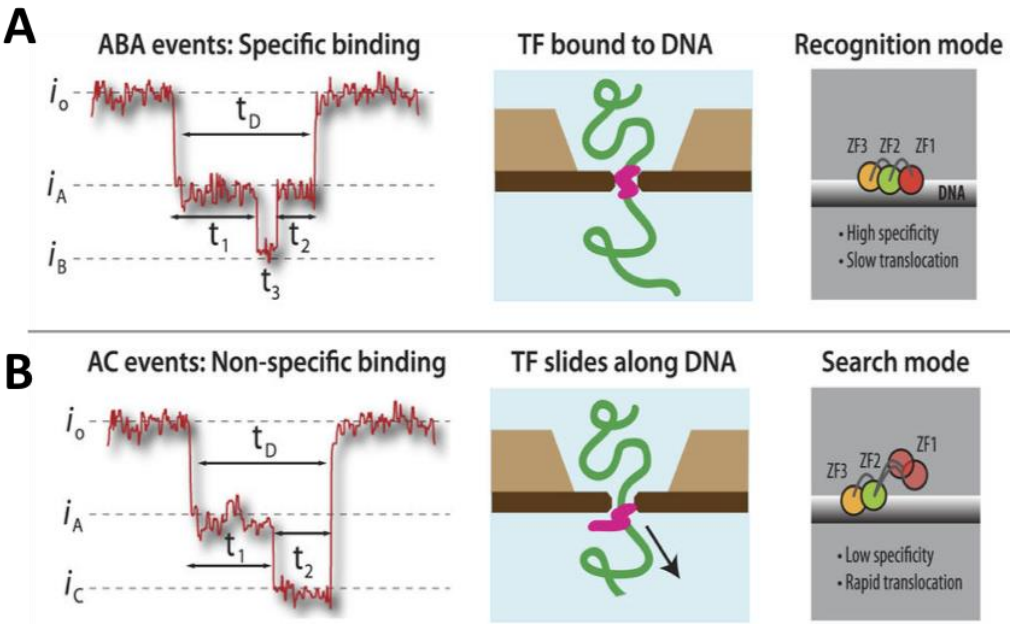


Figure 3. Distinguishing between specific and non-specific binding of TF—DNA with solid-state nanopore. Translocation event traces and proposed mechanisms for (A) specific binding, and (B) non-specific binding of TF to DNA.[78]

2.2.3 Telomeres. On average, telomeres are shortened by 19 bps per year due to aging, oxidation, stress, mitotic activity of tissues, and lifestyle.[95-98] When shortened to a critical length, telomeres lose their ability to protect the DNA chromosomes[99] and restrict the proliferation of normal somatic cells.[98] This leads to chromosomal fusion and degradation.[100-101] In contrast, approximately 85% of human cancer cells can achieve an "immortality" status by maintaining and elongating telomeres via the *de novo* synthesis of telomeric DNA.[100] Recently, a study was conducted on 47,102 individuals from the general population, where these individuals were followed for up to 20 years to find out the

relationship between telomere length and cancer. Although short telomere length is not an indication of cancer,[98] it was observed that cancer patients with shorter telomere length had increased risk of early death. This result was observed in patients with lung and esophagus cancer, malignant melanoma, and leukemia.[98]

Even though it has been years since the first research, the kinetics of telomeres in cancer cells remains elusive. At present, measuring the length of telomeres and observing the kinetics of folding are still challenging, as there is no gold-standard technique.[102] In order to fully understand the role of telomeres in cancer prediction or therapy, it is essential to understand the kinetics of telomere folding and other conformational changes as a response to different living and environmental conditions.

Work is currently underway to apply nanopore sensor in tracking the telomeric DNA G-quadruplex folding/unfolding (Figure 4). Several research groups have used biological nanopore to capture some or all four folded-structures of G-quadruplex, including hybrid (hybrid-1 and hybrid-2), basket, and propeller structures.[47, 103-105] Findings from these studies reported that even though the four G-quadruplex structures all folded from the same DNA sequence, they produced very different electrical signatures.[105] This was attributed to the overall shape and volume of each secondary structure. It was observed that both hybrid-1, -2, and basket forms had a diameter of 2.7 nm and 2.4 nm, respectively. Since the *cis* opening of the α-hemolysin pore has a diameter of 3.0 nm, these three folds can enter the large vestibule. However, the propeller fold, with a disk-shaped structure and diameter of 4.0 nm, exceeds the diameter of the nanopore *cis* opening and was unable to enter the vestibule.[106]

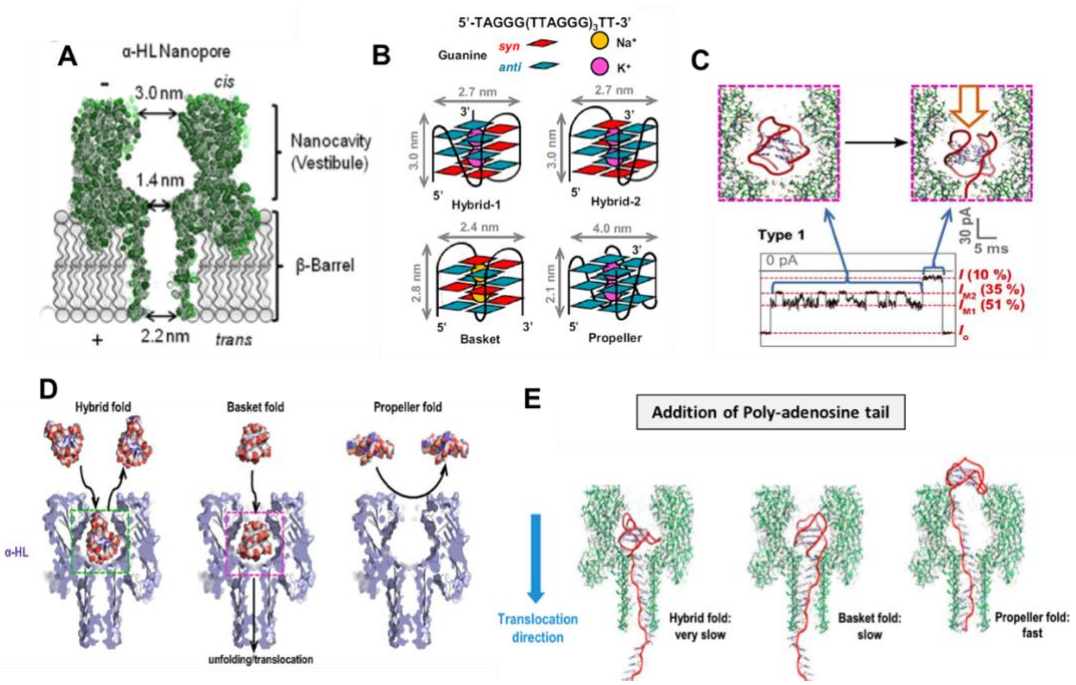


Figure 4. Capturing unfolding process of the four G-quadruplex structures with biological nanopore. (A) Schematic of the α-hemolysin nanopore, with the cis opening of 3.0 nm, constriction of 1.4 nm, and trans opening of 2.0 nm. (B) Folding structures and dimensions of G-quadruplex conformations: hybrid-1, hybrid-2, basket, and propeller. (C) G-quadruplex fold entered and unfold inside the nanocavity of α-hemolysin nanopore, causing two distinct levels of blockage. (D) Except the propeller fold, all other G-quadruplex can enter the cis opening of α-hemolysin nanopore without unfolding, but cannot pass through the pore constriction.114 (E) Models of the three conformations with the additional 5'-dA25 tail unraveling through α-hemolysin pore. Both hybrid and basket folds were able to enter the cis opening of the α-hemolysin pore, thus unraveled inside the pore nanocavity. On the other hand, propeller fold, because of its size, could not enter the nanopore. This conformation unraveled its structure outside of the pore, using the help of the 5'-dA25 additional tail.[105]

Another inventive solution to capture and unravel G-quadruplexes is to employ a 25-mer poly-2'deoxyadenosine tail (d25A-tail) on the 5' end of the telomeric DNA. Applying this method, the Burrows group reported the analysis of various folding motifs of the telomere sequence, with and without the 5'-d25A-tails.[105] Among the four loop topologies, only the basket fold was able to translocate through the nanopore without the addition of the homopolymer tail to the 5' end. For the G-quadruplex to move through the

nanopore, it needs to unravel to a singular strand which would be able to translocate through the narrow β-barrel, and the remaining G-triplex has to roll within the vestibule. This is likely a favorable process for the basket fold because of its nearly spherical shape.[107] Even though the volume of the vestibule is large enough to accommodate all four G-quadruplexes within its cavity, the narrow entrance of the vestibule prevented the propeller fold from entering the nanopore. However, with the addition of the 5' tail, the propeller fold was able to circumvent the problem of entering the cavity, and yet still had a very fast translocation signature. This is attributed to the fact that the propeller fold was able to roll outside of the vestibule while an electric force was applied to the dA25-tail as it threaded through the ion channel, without having any molecular interactions or steric hindrance that would have been experienced on the interior of the vestibule.[105]

In the light of those previous studies, for the first time, the unfolding kinetics of human i-motifs were studied using the α-hemolysin nanopore. Under acidic conditions, cytosine (C)-rich DNA sequences can adopt i-motif folds, since the hemi-protonation of C- rich strands allow C^+•C base pairs to form.[54] The Ding et al. conducted experiments on the human i-motif sequence at a constant ionic strength, but using various pH (5.0 – 7.2). Since the dimension of an i-motif (2.0 nm x 2.0 nm) is smaller than the *cis* opening (~3.0 nm) of the α-hemolysin pore, it can enter the pore without unfolding and be captured in the nanocavity.[54] Hence, a d25A tail was attached to the sequence, in order to increase the unfolding rate of i-motif. Upon the attachment of d25A, it was observed that at pH 5.0, the folded structure entered the α-hemolysin pore, yielding characteristic current patterns. However, when the pH were at 6.8 and 7.2 (higher than the transition pH 6.15), the percentage of strands still folded was 4% and 2%, respectively. Furthermore, the force

applied in this study was analogous to the forces exerted on genomic DNA by RNA polymerases II (5-20 pN) and DNA helicase (6-16 pN).[54] Hence, these studies strive to show the potential of α-hemolysin as part of biosensor development, aiding in our knowledge of the lifetimes of i-motifs of telomere sequences, and their biologically relevant structures, which can be used as drug delivery targets for cancer treatments.[108]

These findings are steps toward a better understanding of the folding and unfolding mechanisms of the telomere. When pre-detecting different cancer types, conventional methods, such as FISH, Southern blot, and quantitative-PCR, require complicated meta-analyses, chemical-crosslinking and intensive preparation. This can lead to inconsistent results.[98, 109-111] Whereas nanopore analysis, lacking all those complications, allows a better understanding of the kinetics and mechanisms, aiding in the analysis of how different oxidation, stress and factors affect the length of telomeres, as well as the correlation between cancer development and telomere immortality.

2.2.4 DNA methylations.

2.2.4.1 Hyper- and hypomethylation of CpGs. In humans, methylation of DNA is an epigenetic modification that transfers a methyl group from S-adenosyl-methionine to cytosine residues, forming 5-methylcytosine (5-mC). In mammalian cells, methylation of *CpGs* can directly or indirectly repress gene expression. For example, hypermethylation of *CpG* islands in the promoter region can directly lead to transcriptional silencing of tumor-suppressor genes. On the other hand, methylated *CpGs* can indirectly interfere with transcription to prevent the binding of basal transcriptional machinery or ubiquitous TFs. This process contributes to all of the typical hallmarks of a cancer cell originated from tumor-suppressor inactivation.[112] With aging, cell deregulation provides mutation

accumulation and epigenetic alterations (i.e. aberrant methylation in DNA) the chance to build up, causing proliferative advantages and genomic instability. Aberrant DNA methylation, including loss of methylation (hypomethylation) and gain of methylation (hypermethylation), has been classified as a common causal factor of many cancers.[113-117] For instance, hypermethylation is linked to various types of cancers, including lung, prostate, breast and colon cancers,[118-121] while hypomethylation of *CpGs* has been reported to be associated with kidney, liver, pancreas, lung, cervix, stomach and uterus cancers.[122-127] Hence, detecting aberrant DNA methylation can have an important role in cancer treatment and precancerous detection.[37, 56]

The overall level of 5-methylcystosine contained in the cell sample can be quantified using high-performance liquid chromatography (HPLC), high-performance capillary electrophoresis (HPCE), bisulfite sequencing, methylation-specific PCR, among many other methods.[128-134] However, these methods have certain drawbacks. For example, although HPLC and HPCE can accurately quantify the total amount of methylated *CpGs*, they have incomplete restriction enzyme cutting, offer limited region of study, require substantial amounts of high molecular weight DNA, and are labor intensive. Similarly, with PCR-based methods, only the methylation status of *CpG* sites that are complementary to the primers can be interrogated. Thus, the predominant methylation patterns in the sample may not necessarily reflect the actual results (false positive results).

With nanopore analysis, current methods used in the detection of aberrant *CpGs* methylation usually employ either a methylation specific labeler, or an electro-optical tagging.[56-57, 135] The first method, as proposed by Shim et al., employs an engineered methyl-*CpG*-binding domain protein (i.e. MBD1x or Kaiso Zinc Finger proteins) as a

selective labeler to detect and quantify hypermethylated *CpG* sites in double-stranded DNA (dsDNA)).[56, 135] As the DNA translocated through the nanopore, the presence of 5-mC·labeler complexes caused a signature current blockage, allowing the detection and coarse quantification of 5-mC sites on a single molecule.[135] Indeed, this method set an initial application in screening for the presence of hyper- and hypomethylated DNA. Moreover, Shim et al. pointed out that with the versatile binding affinity of KZF to various methylation patterns, the studied assay can allow various patterns to be screened.[56] Since nanopore analysis requires low volumes of DNA for testing, the technique will be more applicable and practical for clinical use. Without the need of DNA replication and amplification, detecting *CpG* methylation using nanopores requires much less labor in comparison to other conventional methods.

The second method, as previously discussed, uses an electro-optical solid-state nanopore to detect and quantify hypomethylation in DNA.[57] In this approach, enzyme DNA MTases was assisted by small molecular weight synthetic cofactors to catalyze a one-step enzymatic reaction. This enzyme-cofactor complex was directly conjugated onto fluorescent probes and attached to the unmethylated *CpG* sites. The Meller group was able to detect and differentiate between fully methylated, partially methylated and unmethylated dsDNA, using ultrasensitive electro-optical nanopore sensing as the tool for single-fluorophore multicolor quantification. Unlike MBPs, DNA MTase only labeled unmethylated *CpG* sites of the target DNA. This allowed the direct targeting of hypomethylated *CpG* sites in the genome (i.e. promoter regions of oncogenes). Furthermore, this electro-optical solid-state nanopore showed a high potential for employing multiple DNA MTases and other epigenetic biomarkers. With the aid of those

biomarkers, orthogonal labeling/sensing of 5-mC can be achieved in the future.[57] Further research must be done in order to develop a calibrated scale to count the number of unmethylated *CpGs* in the target sequence.

2.2.4.2 Other variants of CpGs methylations. The family of Ten-Eleven Translocation (TET) proteins have been shown to oxidize methylcytosine (mC) into hydroxymethylcytosine (hmC) and further oxidize hmC into formylcytosine (fC) and carboxylcytosine (caC) (Figure 5).[136-139] 5-hmC normally exists at a high level in self-renewing and pluripotent stem cells.[136, 140] Both mc and hmC influence mammalian embryonic stem cell maintenance,[141-142] angiogenesis,[143] and development.[144] Thus, hmC is a promising molecular biomarker with predictive and prognostic value.[145] As for fC and caC, there is still very little research being done. Because the topic has just recently been discovered, we currently lack a robust method to distinguish between these five chemical modifications of cytosine. Even distinguishing between mC and hmC is a challenge for available methods.[142, 146]

The presence of bulk 5-methylcytosine (5-mC) and 5-hydroxymethylcytosine (5-hmC) on ss- and dsDNA has been successfully detected and distinguished using both solid-state and biological nanopores.[147-149] For instance, the Drndic group proposed a method using solid-state nanopore to discriminate two different structures that translocated through the pore (5-mC and 5-hmC). Upon the addition of 3kbp dsDNA, a sequence of current blockage was generated, in which the magnitude of each spike was related to the excluded volume of biopolymer that occupies the pore. From the differences in ΔI_{max} values, Wanunu et al. was able to discriminate between 5-mC and 5-hmC. Shorter end-to-end distance of the more polar 5-hmC indicated an increased flexibility in 5-hmC comparedg to cytosine

and 5-mC. Moreover, it was shown that different proportions of 5-hmC in DNA fragment containing cytosine and 5-mC can be quantified using ionic current signal.[147] The second device used in the detection of *CpG* methylation variants employed both the wild-type *phi29* DNA polymerase (*phi29* DNAP) and MspA in the same assay.[148-149] With this unique approach, the Wescoe et al. reported a direction detection of all five cytosine variants (C, mC, hmC, fC and caC). In this single-molecule tool, a *phi29* DNA polymerase drew ssDNA through the pore in single-nucleotide steps and the ion current through the pore was recorded.[148] Overall, the single-pass call ranged from approximately 91.6% to 98.3% depending on neighboring nucleotides.[148-149] Because the knowledge of the five cytosine variants, especially fC and caC, is still very limited, the possibility of these variants having an impact on genome-wide demethylation or other modifications in cancer cells should not be eliminated.

These studies have shown nanopore analysis potential as a robust and efficient tool for the study of DNA methylation. The technique can directly detect *CpG* methylation without the need for DNA amplification or complicated preparation processes. Due to its special characteristics, methylation of *CpG* is usually erased during replication and amplification. Hence, nanopore analysis could be a more practical and reliable method to screen and detect aberrant DNA methylation in cancer patients.

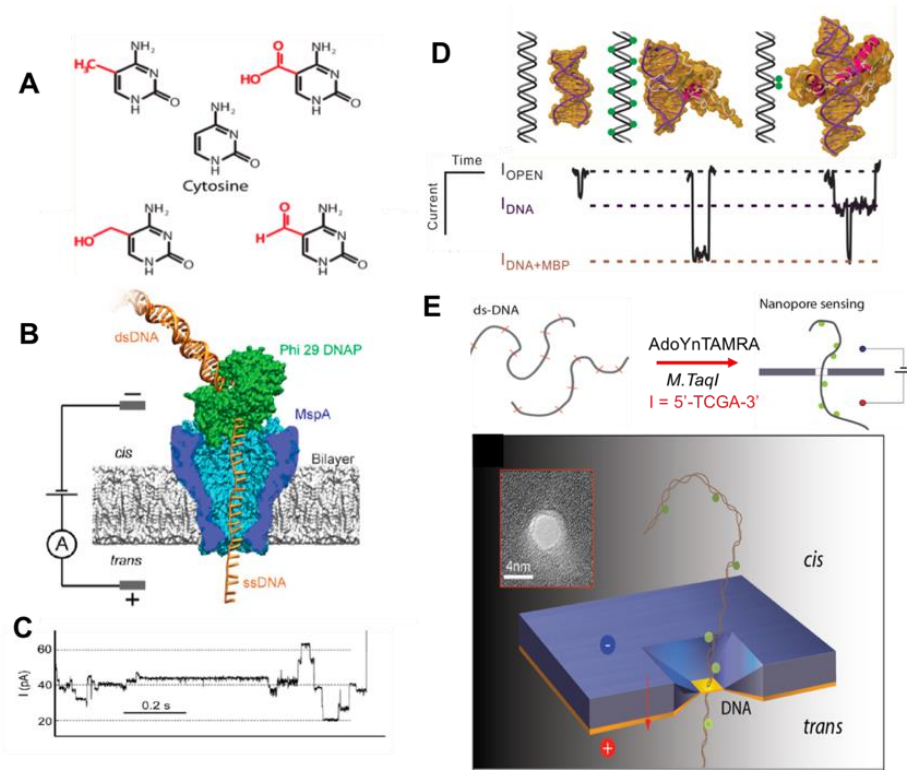


Figure 5. Distinguishing variants of cytosine with biological and solid state nanopores. (A) Chemical structures of cytosine and its variants. First row: mC (left) and fC (right). Second row: cytosine. Third row: hmC (left) and caC (right).123 . (B) Schematic of the Phi 29 DNAP-MspA complex. MspA pore constriction is shorter and narrower compared to α-hemolysin(as shown in the top), allowing short subtle structural changes to be distinguished. (C) A typical trace of DNA translocation through the Phi 29 DNAP – MspA complex.122 (D) Detection of DNA methylation with methyl binding proteins (MBP) using solid state nanopore. MBPs bind to methylated CpGs on DNA, allow the detection and differentiation between unmethylated, hypermethylated and locally methylated DNAs. (E) Detection of DNA methylation with optical-tagging using solid-state nanopore.

2.2.5 Histone-DNA modification. Aberrant DNA-methylations are also linked to the presence of aberrant modifications in histones,150-154 which are the gene activity's dynamic regulators. Histones go through several post-translational modifications, such as acetylation, methylation, phosphorylation, ubiquitylation and others. Specifically, methylation and acetylation of lysine residues on the nucleosomal core histones play an

essential role in gene expression and chromatin structure regulation.[155] In normal cells, histones in DNA sequences are hypoacetylated and hypermethylated. The two key regulators of histone modifications are histone deacetylases and histone methyltransferases (MTases), which are associated with methyl-binding proteins (MBPs) and MTases.[155-157]

Several studies have been conducted on the translocation or unravelling of a nucleosome and its subunit structure through nanopore.[158-160] Generally, it was found that DNA-histone complexes lead to higher applied voltage required and overall longer time periods to translocate through the nanopore, most likely due to either: (1) the bulky disk shape nucleosome experiencing a higher drag force comparing to a bare dsDNA, (2) the positively charged histone core lowering the total net charge density of nucleosomes, causing the translocation speed in electrophoresis to reduce, and (3) the unwinding process of histone-DNA complex.[159, 161]

As previously discussed, epigenetic modifications have been known to affect the structural integrity and stability of nucleosomes. Given this fact, it was hypothesized that methylation of *CpGs* on dsDNA would affect the way nucleosomes fold and/or unravel. To test this hypothesis, the Langecker et al. investigated the influence of DNA methylation on the stability of unlabeled mononucleosomes.[160] Similar to the results reported in other studies, under the electrophoretic force, the nucleosomal DNA tail entered the pore and gradually unraveled under increasing voltage, which was much higher in comparison with free DNA capture.[162] This experiment was repeated on nucleosomes with and without methylated DNA sequences, yielding that methylation of *CpGs* did not affect the nucleosome assembly, stability, or unraveling trajectories. This finding suggested that histone modifications (*e.g.* acetylation and phosphorylation) play a much more dominant

role in nucleosomal maintenance than DNA methylation. The confirmation of methylation-independent nucleosome stability indicated other possible mechanisms by which DNA methylation alters gene expression, for example, modulating the binding of transcription activators/repressor.[160]

The nanopore-based studies outlined herein lay the groundwork for understanding and predicting the influence of different histone core modifications on the nucleosome structure,[160] about which our knowledge is still quite limited. Unlike conventional methods (e.g. single-gene chromatin immunoprecipitation (ChIP), ChIP with a DNA array (ChIP-on-chip),[163-164] HPLC, HPCE, and many others), nanopore devices are more versatile, because they do not rely heavily on the quality of the polyclonal antibodies or antibodies that are available.[154] Although the study here indicated that DNA methylation does not affect the nucleosome assemble, further studies need to be done in order to confirm the role of DNA methylation in other processes (i.e. regulating transcription activators/repressors binding, or gene expressions), as well as the relationship between acetylation and phosphorylation on nucleosome assembly and chromatin stability.

2.2.6 MicroRNA. MiRNAs are small endogenous biomolecules that are in length of 18-22 bps. They play an important role in embryonic differentiation, hematopoiesis, cardiac hypertrophy and numerous cancer-related processes, including proliferation, apoptosis, differentiation, migration and metabolism.[165-166] Because a single miRNA can target up to hundreds of mRNAs,[167] an aberrant miRNA expression may affect several transcripts and cancer-related signaling pathways. In cancer cells, because of the genetic diversity of tumors and cancer cell lines, an individual miRNA can be up-regulated in one type of cancer and down-regulated in another.[167] Overall, miRNAs function depends on

their targets within the specific tissue.[3] Usually, the up-regulated miRNAs function as oncogenes by down-regulating tumor-suppressor genes, while the down-regulated miRNAs function as tumor-suppressor genes by down-regulating oncogenes.

Detection of miRNAs faces several challenges, mainly due to the short length of miRNAs. Some quantitative methods have been applied to miRNA detection with enhanced sensitivity and/or selectivity, including quantitative reverse transcription real-time polymerase chain reaction (qRT-PCR) assays, microarrays, colorimetry, bioluminescence, enzyme turnover, electrochemistry, molecular beacons, deep sequencing and single-molecule fluorescence.[168-171] Unfortunately, these techniques incur DNA amplification errors, unavailable internal controls, and cross-hybridization. Also, the short sequence of miRNAs makes the designing of probes and primers even more challenging.[168, 170]

MiRNAs have been investigated as potential molecular biomarkers, because their expression levels are associated with various diseases.[172] For instance, each year, lung cancer causes approximately 1.2 million deaths worldwide.[173] Since there is no effective screening procedure available, more than 70% of lung cancer patients were diagnosed with less than a 15% chance of a 5-year survival rate.[173] More than 100 types of miRNAs have been identified to deregulate lung cancer progression.[172] Noticeably, high levels of miR155 and low levels of *let-7a-2* have been associated with a significantly poor prognosis and shorter survival times in lung cancer patients.[174-175] Many research groups have used biological and solid-state nanopores for the detection of miRNAs in different tissues (Figure 6). For example, the solid-state nanopore was used for rapid detection of probe-specific miRNAs (*miRNA-122a* and *miRNA-153*).[176] Specifically, for every 1 fmol of

miRNA duplex per mL solution, the capture rate was 1 molecule per second. In this study, the *p19* protein from the Carnation Italian ringspot virus was used to enrich *miRNA-122a* and *miRNA-153*. Since miRNA concentrations were 1% relative to other cellular RNAs, to detect a specific miRNA using a nanopore sequence, an enrichment step was required.[176] *p19* binds 21-23bps dsRNA in a size-dependent, but sequence-independent manner. Additionally, the highly affinitive and selective viral p19 protein does not bind ssRNA, tRNA or rRNA. This eliminates the possibility of false results from mismatched binding.[177] Detection of 250 molecules in 4 minutes was sufficient to determine miRNA concentration with 93% confidence.[176]

A different approach from using viral proteins for probe-specific miRNAs detection is to employ an engineered-probe with a programmable sequence to differentiate single nucleotide differences in miRNA family members.[172] Wang et al. proposed a system that enabled sensitive, selective, and direct quantifications of cancer-associated miRNAs in the blood. In this study, the group constructed a robust protein nanopore-based sensor that utilized an oligonucleotide probe (P155) to detect aberrant expression of *miRNA-155* and *let-7a-2* from lung cancer patients.[172] The generated signature electrical signals provided a direct and label-free detection of the target miRNA in a fluctuating background, such as plasma RNA extract.[172] Probe (P155) has a programmable sequence and can be optimized to achieve high sensitivity and selectivity. Additionally, using chemical modifications, distinct probes can further be engineered with specific barcodes, allowing multiple miRNAs to be simultaneously detected. Furthermore, with the development of miRNA markers, manipulatable miRNA profile detection nanopore arrays can be constructed for a noninvasive screening and early diagnosis of cancer.[172]

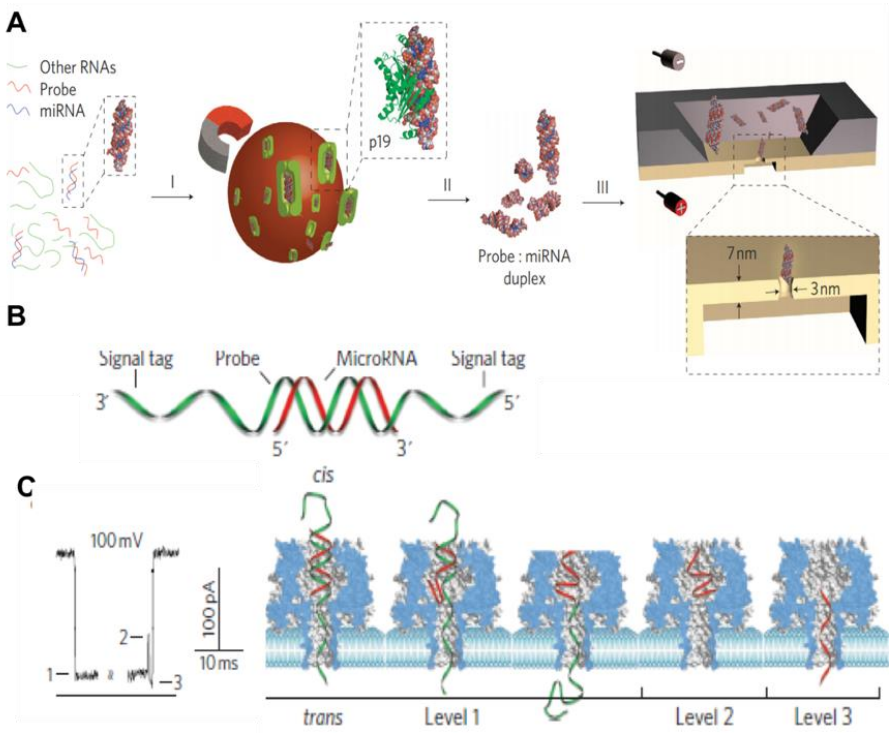


Figure 6. Detection of a miR-155 using using solid-state and biological α- Hemolysin nanopores. (A) Schematic of miRNA detection with viral proteins for probe-specific miRNA, using solid-state nanopore. Protein from Carnation Italian ringspot virus was used to enrich miRNA form background fluid. (B) Detection of probe-specific miRNA using alpha-hemolysin biological nanopore. MiRNA-155 (shown in red) was attached to a DNA P155 probe (shown in green). (C) At 8.0 pH and 100mV, translocation of the miRNA-155·P155 resulted in various current blockage patterns. (C) A typical current blockade with three characteristic blocking levels, representing the mechanism of miRNA-155·P155 complex dissociation and translocation through the pore (as shown in the right-hand side).[176]

Compared to qRT-PCR assays, microarrays, colorimetry, bioluminescence, and other current methods,[168-171] nanopore arrays is a simpler, faster method that can be utilized to detect miRNAs in cancer patients. This approach lacks all the complications that conventional methods have, such as DNA amplification errors, unavailable internal

controls, and cross-hybridization. Early detection is one of the most crucial contributors to a higher survival rate, especially lung cancer patients.[173]

2.3 Summary and Conclusion

In this chapter, we have concisely reviewed the main genetic and epigenetic causal factors of cancer, as well as summarized how nanopore have been used in the research of each factor (Table1). Thanks to several unique features of this emerging technology, the nanopore-based analysis offers four main benefits: First, nanopore analysis offers long reads of genomic DNA (>10 kB). Therefore, linkages between modified cytosines may be revealed that are biologically significant and otherwise difficult to discern. For example, it was shown that histone-DNA interaction is not affected by methylation of DNA. Also, for the first time, differences between caC, fC, and hmC from mC were successfully distinguished. Second, the genomic DNA is read directly as it transports through the nanopore. Thus, errors (false-positive results) caused by copying do not occur. Third, with biological nanopore membranes, the study of biomolecules' folding-unfolding kinetics and mechanisms are possible to accomplish. Furthermore, the DNA fragment can be retained in nanopore indefinitely, allowing rereads of a captured DNA fragment.[149] Finally, many conventional methods are still impractical for clinical testing, because these methods require highly trained experts, intensive labor, a high capital cost, and a large footprint. With nanopore technology, there are no such requirements, offering more flexibility and practicality for research labs and clinics.

Although the concepts of nanopore analysis in early cancer detection are exceptionally promising, several key technological challenges must be addressed before this method can be implemented in clinical uses. First and foremost, the biggest drawback

of nanopore-based methods is high mismatch and error rates. Because the nanopore membrane thickness, especially with respect to biological nanopores, is relatively large compared to a nucleotide, nanopore sensitivity is still low at the single-nucleotide level. Furthermore, even though different DNA conformations and foldings yield distinguishing characteristic current blockades, information about the molecular structure cannot be determined by nanopore membrane alone in order to confirm the exact structure that causes a signature blockades.

In utilizing nanopore technology, researchers need the aid of other equipment, such as circular dichroism (CD), FRET, FISH, among many others. This limits the use of nanopore membranes as an independent, stand-alone tool for molecular studies in general, and early cancer detection, specifically. Moreover, since one single biological molecule can quickly adopt multiple, complex conformations under different environments, many research groups choose to use short/simplified sequences in their nanopore studies. Hence, the complexity of cancer cells has not yet been demonstrated and/or fully investigated with nanopore membranes.

Table 1

Summary of nanopore applications in detecting cancer markers

Cancer Biomarker	Type of nanopore		Support Needed	Advantages	Limitations	Ref
	BL	SS				
Structural variants	✓		*	Does not require multiple processin steps and file output formats. Lov capital cost and short sequencing time	Low sensitivity and accuracy, high mismatch rate. Thus, PHRED score is not high enough for cancer detection yet.	1
Transcriptio-nal factors		✓	*	Label- and tether-free. Does not require chemical-crosslinking or tagging. Hence, allow direct detection and distinguishing between full versus partial, and specific versus nonspecific bindings.	Not able to predict the exact binding site TFs on an unknown DNA sequence.	2
Telomere	✓		*	Does not require complicated sample preparation and labeling, thus, allow the kinetics and folding/unfolding mechanism of the four different G-quadruplex structures to be studied.	The correlation between telomere folding/unfolding and shortening has not been clarified.	3-6 7,
	✓		Poly-adenosine tail			
Aberrant methylation of *CpGs*	✓	✓	Methyl specific labelers	Does not require PCR or complicated sample amplification. Hence, eliminate false-positive results and allow epigenetic changes to be captured. Real-time detection for all four variants of cytosine (mC, cacC, fC, hmC).	Only provide coarse quantification of methylation on *CpGs*. Thus, it is still challenging to finely quantify and map the methylation profile of *CpGs* at the promoter region (essential for precancerous and tumorous detection).	9-14
		✓	Optical-tagging			
Histone-DNA modification		✓	*	Show that methylation of DNA does not affect the ability of nucleosome to ravel and unravel. Moreover, NP allows nucleosomes to unravel with the double-stranded DNA remain unzipped.	Has not proved the ability to detect and/or map global changes of post-translational modifications (i.g. methylation, acetylation) on nucleosome.	15-18
Expression of miRNA	✓		Engineered probe	Does not require amplification and cross-hyberdization of samples.	Multiple miRNAs can be up- or downregulated at the same time in cancerous cells. Hence, several miRNAs have to be tested and clustered together. However, since the sensitivity and accuracy of NP still low, this will potentially lead to false results.	19, 20
		✓	Viral proteins	Multiple miRNAs can potentiallybe detected using engineered probe. Data were consistent with good statistical distribution.		

Chapter 3

Biophysical Properties of DNA - Investigation of Compacted DNA Structures Induced by Na^+ and K^+

3.1 Abstract

DNA compaction is a highly promising method for gene therapy and delivery. Experimental conditions to induce these compacted DNA structures are often limited to the use of common compacting agents, such as cationic surfactants, polymers, and multivalent cations. In this study, we showed that in highly concentrated buffers of 1M monovalent cation solutions at pH 7.2 and 10, biological nanopores allowed real-time sensing of individual compacted structures induced by K^+ and Na^+, the most abundant monovalent cations in human bodies. Since the binding affinities of each nucleotide to cations is different, the ability of a DNA strand to fold into compacted structure depends highly on the type of cations and nucleotides present. Our experimental results were favorably comparable to findings from previous molecular dynamics simulations for DNA compacting potential of K^+ and Na^+ monovalent cations. We estimated that the majority of single-stranded DNA molecules in our experiment were compacted. From the current traces of nanopores, the ratio of compacted DNA to linear DNA molecules was 30:1 and 15:1, at a pH7.2 and pH10, respectively. Our comparative studies revealed that Na^+ monovalent cations had a greater potential of compacting the 15C-ssDNA than a K^+ cation did. This chapter was adapted from our previously published paper (Vu, Trang, Shanna-Leigh Davidson, and Jiwook Shim. "Investigation of compacted DNA structures induced by Na^+ and K^+ monovalent cations using biological nanopores." *Analyst* 143.4 (2018): 906-913.)

3.2 Introduction

In an aqueous solution, DNA is a long, highly charged polymer, and is usually in an elongated conformation due to strong repulsion forces between the negatively charged phosphate groups on the DNA backbone.[178] Upon the addition of compacting agents, DNA can quickly go through a conformational change, mainly due to the neutralization of negatively charged DNA through cation binding.[178-179] Interaction of cations with the DNA phosphate backbone has been known to induce several DNA phase separation processes, including DNA compaction, condensation, aggregation, and precipitation, without a clear distinction between these processes.[180] Previous literatures have intensively studied and reviewed several *in vivo* and *in vitro* compacting agents, including multivalent cations,[181] polycations,[182] neutral polymers,[183] cationic nanoparticles,[184-185] and surfactants.[186-188] In addition, the presence of monovalent cations can also promote DNA compaction processes in both concentrated and diluted solutions of DNA.[180, 189-195] Several previous studies have investigated the role and mechanism of cations in the process of binding and compacting DNA.[184-185] However, the existence of the DNA compacted structure has not been discussed in the nanopore analysis setting yet, mainly due to the complex quantum-chemical mechanism of the DNA compaction process.

In this chapter, we used α-hemolysin NPs to capture the cations-induced compacted structure of a single-stranded DNA with a specific sequence containing only cytosine nucleotides (5'-$(CCC)_5$-3'). In order to investigate the ability of Na^+ and K^+ monovalent cations as compacting agents for ssDNA, studies were conducted with buffers containing high salt concentration at various pH values.

3.3 Methods

3.3.1 Chemicals. Chemicals and reagents used in the experiments were purchased from the following vendors: 1,2-diphytanoyl-sn-glycero-3-phosphocholine (Avanti Polar Lipids); pentane (Sigma-Aldrich); potassium chloride (99.5%, Sigma-Aldrich); sodium chloride (Sigma – Aldrich); Tris base (Promega); Ethylenediamine Tetraacetic acid (EDTA) disodium salt dehydrate electrophoresis crystalline powder (Fisher BioReagents); agarose (Bio-Rad); pentane (Fischer Chemical), hexadecane CH3(Ch2)14CH3 (Avantor); and hydrochlodric solution 1M (LabChem). Synthetic oligonucleotides (Integrated Technology) with the following sequences were used: 5'-CCC CCC CCC CCC CCC – 3'. The control sample is an 18-bases ssDNA with the sequence of 5'-TAATCATCGCGTACTAAT-3'. Salt solutions were buffered with 10mM Tris and titrated with HCl until the required pH of 5.0, 7.2, or 10.0. DNA solutions were suspended in a buffer solution containing 10mM Tris, 1mM EDTA, pH 8.0 at room temperature before usage.

3.3.2 Circular dichroism. CD spectra were collected using the Jasco CD spectropolarimeter, model J-810. All measurements were measured at 15$_o$C as maintained by the temperature-control units affiliated to the spectrometer. All samples used for CD measurements contained the same salt (1M) and ssDNA (1 μM) concentrations, as used in the single-molecular study. Spectra were captured in the 220-320 nm wavelength, 1.0 nm bandwidth, 50 nm/min scanning speed, and a standard sensitivity.

3.3.3 Experimental setup. The experiments herein were performed following previously described protocols.[45, 196-198] Briefly, a Teflon film divides the testing chamber into two separate compartments, named *cis* and *trans* chambers, to which negative and

positive electrodes are attached to, respectively. Artificial lipid bilayer, acting as an electric insulation, is formed over a premade orifice of 100μm in diameter in the Teflon film. α-hemolysin (α-HL), a toxin released by the *Staphylococcus aureus* bacterium, is well known for its ability to form a mushroom-shaped protein channel with a well-defined structure, as well as dimensions. Specifically, the α-HL consists of a *cis*-opening (2.6nm), a nano-constriction (1.4nm), and a *trans* opening (2.0nm).[52] When α-HL protein is introduced to the *cis* chamber, it can insert a mushroom-shaped channel into lipid bilayer, connecting the *cis* and the *trans* sides. DNA or any other charged biomolecules are then injected to the *cis* chamber. Under an applied current, DNA molecules are driven through the α-HL nanopore, causing a current-drop (blockage). A lipid bilayer, 1,2-diphytanoyl-sn-glycerophosphatidylcholine (Avanti Polar Lipids), was formed spanning the hole, creating a planar layer of insulation. The *cis* and *trans* chambers were both filled with symmetrical salt solutions of 1M concentration, containing either NaCl or KCl as per the testing conditions. The protein nanopore, *α-hemolysin*, was injected into the bilayer on the *cis* side, creating a single open channel joining the two chambers. A final concentration of 1 μM 15C-ssDNA was introduced to the *cis* chamber. Ag/AgCl electrodes (1.5% Agarose in 3M KCl) was attached to each of the *cis* and *trans* chambers, with the *cis* chamber being grounded such that a positive voltage would drive the negatively charged DNA molecules through the nanopore channel and into the *trans* chamber. Data acquisition included an electrophysiological setup, with picoampere (pA) current traces being recorded as DNA molecules moved through the single channel.

3.3.4 Single channel analysis. An Axopatch 200B patch-clamp amplifier was employed with Clampex 10.7 software (Molecular Device Inc.) to record and acquire data through a Digidata 1550B A/D converter (Molecular Device Inc.) at a sampling rate of 20 kHZ, and an analogue, eight-pole, low-pass Bessel filter of 5kHz. The data were analyzed using Clampfit 10.7 (Molecular Device Inc.), and Excel (Microsoft). Single-channel currents were determined using amplitude histograms by fitting the peaks to Gaussian functions. Dwell-time histograms fitted to an exponential distribution allowed us to determine the duration of short-lived blocks for DNA translocation. Molecular modeling and chemical structures were created using Maestro (version 2015-4, Schrödinge Suite). All graphical presentations were organized and cleaned up using Illustrator (Adobe).

3.4 Results and Discussion

We chose single-stranded 15-mer oligonucleotides consisting only of cytosine, annotated as 15C-ssDNA, as a proton-sensitive DNA strand to study the compacted DNA structure. To emphasize the unique characteristic of 15C-ssDNA to adopt compacted structures, control experiments were performed on an 18-bases ssDNA sequence (containing all four nucleotides: A, T, C, G). Previous studies have shown that C-rich ssDNA has the ability to form i-motif structures under slightly acidic environments (below the transitional pH 6.15), and sustain in linear form under neutral or basic environments (pH higher than 6.15).[199-202] Thus, the cytosine-only sequence chosen for the current study makes this DNA molecule highly improbable to form stem-loop intramolecular base pairings (i.e. compacted structures) at pH $\geq$ 6.15. Moreover, a previous study showed that using solid-state nanopore, the probability of DNA knots observed in long DNA molecules (~2kbps) is 1.83%.[203] Thus, the short length of the chosen sequence rules out the possibility

of DNA knots formation in this study. All experiments herein were conducted with 15C-ssDNA, under constant electrophoretic forces at 100, 120, or 140mV, and symmetrical salt concentration of either 1M NaCl or 1M KCl in both *cis* and *trans* chambers. Buffers with various pH levels at 5.0, 7.2, and 10 were used in the experiments in order to investigate the effect and compaction potential of monovalent cations on both folded and unfolded single-stranded DNA. All events are characterized by their current blockage amplitude ($\%I_b/I_o$, where I_b and I_o are current amplitudes of the blocked and the empty pore), translocation time through the nanopore (t_b), and frequency (f). All buffers are prepared with 10mM Tris and 1mM EDTA and, hereafter, all buffers are named according to the type of cation present (either K^+ or Na^+) and pH level of the buffer. For example, Na5.0 indicates 15C-ssDNA sample in 1M NaCl at pH 5.0. Current – voltage (I-V) curves of all experiments fit the characteristic shape of the α-hemolysin nanopore I-V curve (Figure 7).

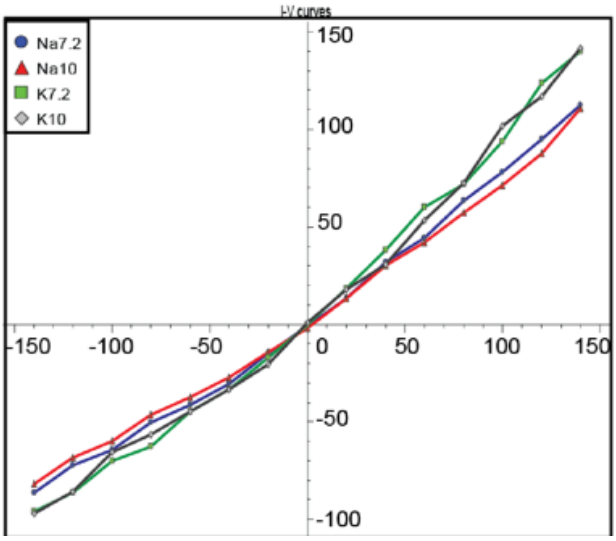


Figure 7. IV curves of the experimental conditions, containing 1M of either KCl or NaCl, at 7.2 and 10 pH. All I-V curves fit the characteristic I-V plot of an -hemolysin nanopore.

3.4.1 Effect of monovalent cations on the folded, i-motif DNA structure. To analyze the nanopore ionic current signal for the fully-folded i-motif structure, we first examined the translocation properties of 15C-ssDNA at pH 5.0 in both K5.0 and Na5.0 buffers. At pH 5.0, when the 15C-ssDNA was added to the *cis* chamber, and electrophoretically driven through the α-hemolysin NP at 100, 120, and 140mV, long-lived partial current blockage events were observed in both K5.0 and Na5.0 samples (Figure 8), indicating the presence of i-motif inside the nanopore. Capture and analysis of the fully folded i-motif structure in K^+ environment has been investigated using α-hemolysin nanopore before.[54] Under slightly acidic conditions, the cytosine-rich DNA sequence becomes hemi-protonated, allowing for the formation of the i-motif structure (Figure 8A). This i-motif structure is approximately 2.0 x 2.0nm in diameter, [54] thus smaller than the cis entrance to the mushroom-head shape nanocavity, but larger than the constriction and β-barrel of the nanopore. Thus, the folded structure of i-motif can enter the channel and reside in the nanocavity. Then the folded i-motif structure can either exit back to the cis opening or translocate through the nanopore to the trans side after unfolding in the nanocavity. The process can last up to several seconds or minutes, causing event blockades with different durations and amplitudes, as shown in Figure 8B (K5.0 and Na5.0), with no real statistical difference between dwell times of i-motif structure in K^+ and Na^+ environments at pH 5.0. Our data herein are fully consistent with previous studies, which demonstrated that half-life, $t_{f1/2}$, of the i-motif structure at pH 5.0 can vastly vary from seconds to minutes.[204-205] In these studies, $t_{f1/2}$, of a i-motif structure represents the amount of time it takes for half of the sample population to become folded.[205]

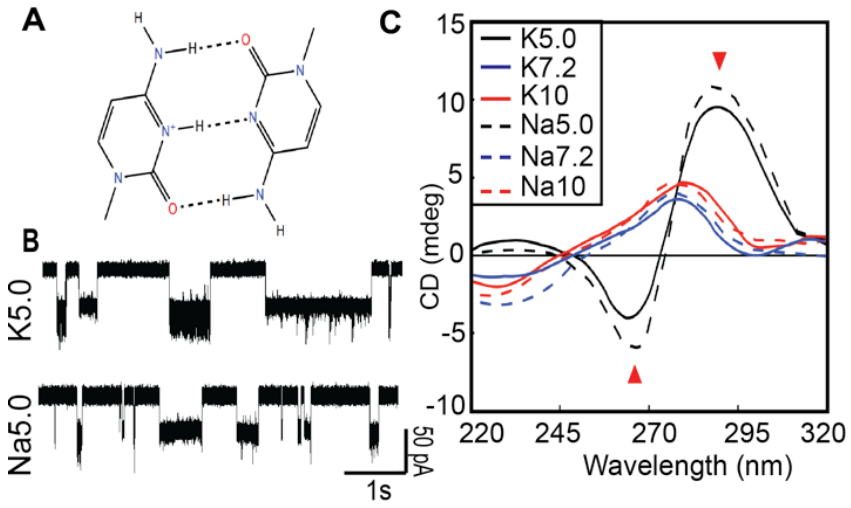


Figure 8. Formation of i-motif at pH 5.0. (A) Under slightly acidic conditions, the cytosine-rich DNA sequence is hemi-protonated, and can adopt an i-motif structure. (B) reflects the residence of i-motif structure inside the NP for both samples, containing either 1M KCl and 1M NaCl, at pH 5.0. The residence of i-motif structure inside the nanocavity causes current blockages that can last up to several seconds or minutes, with %I/Io ~ 50-60%. No event with %I/Io < 30% was observed in both samples (K5.0 and Na5.0). (C) Circular dichroism measurement confirmed the existence of i-motif at pH 5.0 in both K5.0 and Na5.0 samples, which show a typical positive peak at 285nm and a negative peak at 265nm. Other samples at higher pH do not exhibit the i-motif characteristic peaks. Overall, the main differences in CD spectrum are between samples with a different pH, but not between those containing different types of cations (K^+ versus Na^+) . Samples were labelled per the type of cation in the buffer (either K^+ or Na^+), followed by the solution pH value (i.e, Na5.0 means experiment was conducted in 1M NaCl buffer solution, pH 5.0).

Circular dichroism (CD) spectroscopy analysis in our laboratories confirmed that the DNA molecules K^+ and Na^+ buffer solutions reveal a typical i-motif conformation only at low pH of 5.0 (Na5.0 and K5.0 samples), but a random coil at neutral and high pH (Na7.2, K7.2, Na10 and K10 samples). Specifically, the CD spectrum of both Na5.0 and K5.0 samples showed a strong positive peak near 285nm, a negative peak near 262nm, and a cross over at 275nm, indicating the definite presence of i-motif structure in the solutions.[206] In contrast, at pH 7.2 and 10.0, all of the samples (Na7.2, Na10, K7.2, and K10) exhibited a dramatic decrease in the CD signal and completely lost the characteristic i-motif peaks. In general, the i-motif characteristic peaks were not present in samples with

a higher pH, thus confirming that differences in CD spectra are due to differences in pH, and not due to usage of different cations (K^+ versus Na^+) in samples. The results obtained from the event data traces of K5.0 and Na5.0 (Figure 8B) and CD spectrum indicate that under slightly acid conditions (pH 5.0), there is an absence of compacted DNA structure going through the α-HL NP. The binding (or lack thereof) of monovalent cations to ssDNA does not induce DNA compaction at pH level of 5.0. Thus, it is hypothesized that the C+-C base pairing has a significantly higher affinity and can last much longer than cation-DNA binding. Currently, it is unclear whether or not the binding of monovalent cations can slow down the translocation speed of a fully-folded i-motif structure through the nanopore.

3.4.2 DNA compaction induced by K^+ and Na^+ monovalent cations. In order to investigate the effect of monovalent cations on ssDNA compaction, single-channel nanopore analysis of the fully unfolded, linear 15C-ssDNA was conducted at pH 7.2 and 10 (Figure 9 and Figure 10). In an earlier single-channel study of C-rich DNA translocation, it was shown that as the pH level was raised to above 6.15, most i-motif structures unfolded into linear DNA, resulting in mostly short-lived blockage events.[54]. Indeed, we observe a similar characteristic trend of unfolded, single-stranded DNA translocation ($\%I_b/I_o \geq 80\%$) in all samples at both pH 7.2 and 10. However, in contrast to the previous finding with specific sequencing of i-motif, our experimental data with cytosine only sequencing DNA oligos exhibited two types of blockages: Type 1 is the translocation events of linear ssDNA ($\%I_b/I_o \geq 80\%$), and Type 2 events are caused by the encounter of DNA compacted structures with the cis opening of the nanopore ($\%I_b/I_o \sim 30\%$), as shown in Figure 9A-B and Figure 10A–B. A transmembrane voltage (+100mV) was applied from the trans side, with the cis side grounded

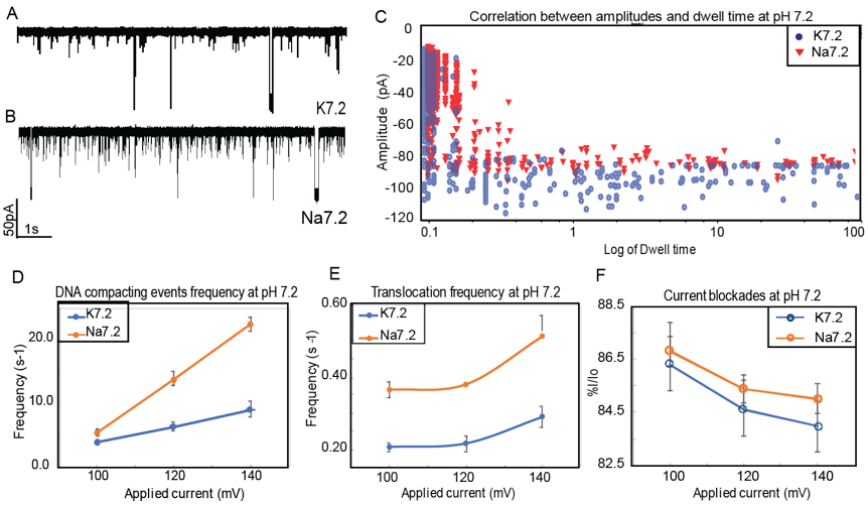


Figure 9. DNA translocation through nanopore at pH 7.2. Data traces of (A) K7.2 and (B) Na7.2 were collected at 120mV applied current. Both samples exhibit two main types of events: Type 1 are translocation events of DNA (%I/Io > 70%), and Type 2 are DNA compacting events (%I/Io < 70%), caused by the encounter of DNA compacted structure with the NP. (C) Correlation between amplitudes of current blockages and their dwell time K7.2 (blue) and Na7.2 (red). All DNA compacting events have very short dwell times, whereas deeper-current blockages can last several seconds in both type of cations (K^+ and Na^+). The number of events observed in Na7.2 is slightly higher than in K7.2. Similar results were obtained through the events frequency analysis for both (D) DNA compacting events and (E) translocation events. (F) Correlation between the %I/Io and the applied current

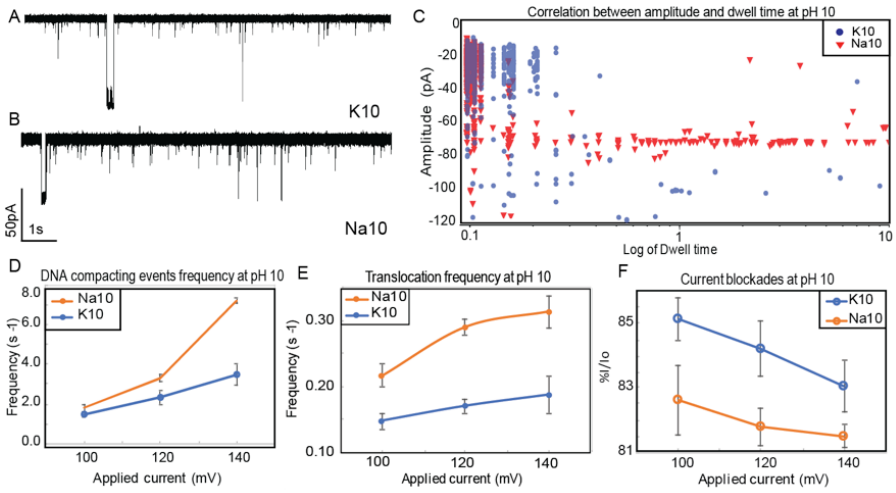


Figure 10. DNA translocation through nanopore at pH 10. Data traces of (A) K10 and (B) Na10 were collected at 120mV applied current. Both samples exhibit two main types of events: Type 1 are the translocation events of DNA (%I/Io > 80%), and Type 2 are DNA compacting events (%I/Io < 30%). (C) Correlation between amplitudes of current blockages and their dwell time K10 (blue) and Na10 (red). All Type 2 events exhibite short dwell times, whereas deeper-current blockages can last several seconds. Similar results were obtained through the events frequency analysis for both (D) DNA compacting events and (E) translocation events. (F) Correlation between the %I/Io and the applied current.

Hereafter, Type 2 events are also referred to as "DNA compacting events." All samples were captured at a sampling rate of 50 μs, and at any given voltage (100, 120, or 140mV), Type 2 events exhibit a significantly shorter dwell time (t_b^c ~ 100us) and smaller current-blockage amplitudes than Type 1 events, forming two distinct clusters of event on the amplitudes-dwell time scatterplots (Figure 9C and Figure 10C). Interestingly, as the voltage ramped from 100 to 140 mV, a decrease in $\%I_b/I_o$ is observed for Type 1 events (Figure 9F and Figure 10F). For the control experiment, Type 1 events show a significant increase in event occurrence, while Type 2 events cannot be observed (Figure 11 and Figure 12). This result indicates that the control sample (18-bases ssDNA) cannot adopt the compacted structure. Events with $\%I/I_o$ ~ 50% are suspected to be the stem-loop structures (hairpin) formed by nucleotides base-pairing (A-T and C-G) within the ssDNA.

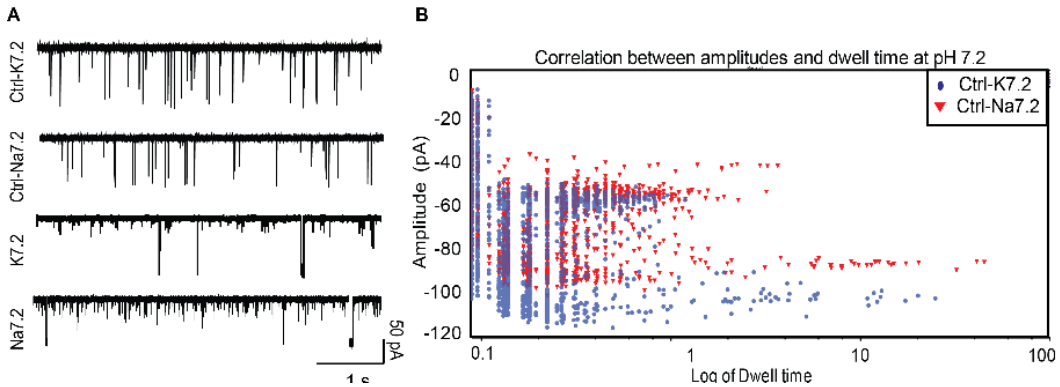


Figure 11. Negative control experiment at pH 7.2. (A) Event data traces of Control K7.2, Control Na7.2, K7.2, and Na7.2. The control sample is an 18 bases ssDNA containing all four nucleotide base types (A, T, C, G). The absence of shallow blockage event (%I/Io < 30) indicates the control sample's inability to form compacted structures. (B) Scatter plot represents the correlation between event blockage amplitude and dwell time of the Ctrl-K7.2 and Ctrl-Na7.2. The two event clusters indicate the presence of hairpin structure (%I/Io ~ 50%) and the translocation of a linear ssDNA (%I/I ~ 90%). All events were collected at 120 mV.

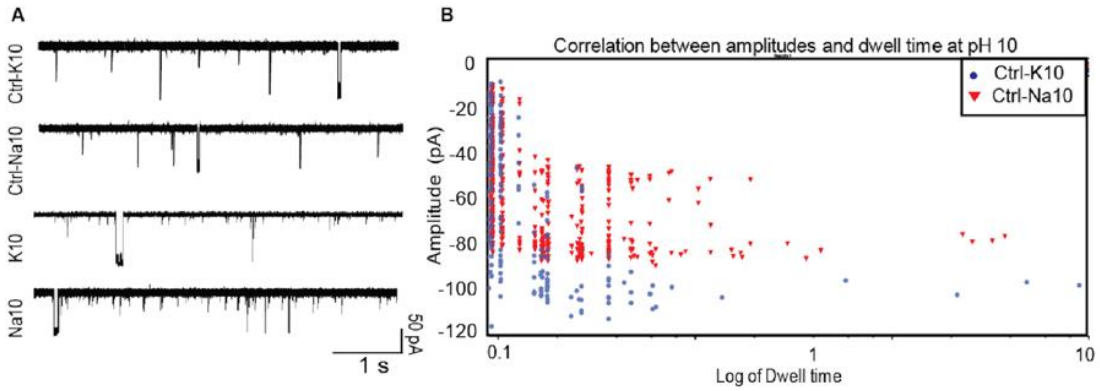


Figure 12. Negative control experiment at pH 10. (A) Event data traces of Control-K10, Control-Na10, K10, and Na10. The control sample is an 18 bases ssDNA containing all four nucleotide base types (A, T, C, G). The absence of shallow blockage event (%I/Io < 30) indicates the control sample's inability to form compacted structures. (B) Scatter plot represents the correlation between event blockage amplitude and dwell time of the Ctrl-K10 and Ctrl-Na10. The two event clusters indicate the presence of hairpin structure (%I/Io ~ 50%) and the translocation of a linear ssDNA (%I/I ~ 90%). All events were collected at 120 mV.

3.4.3 Comparing between translocation and DNA compacting events. We hypothesize that Type 2 events are produced by secondary structures, with dimensions larger than 2.6nm (*cis* opening of the nanopore), and are the result of the DNA compaction induced by monocations (K^+ and Na^+). Under electrophoretic force, these compacted DNA structures are drawn toward the *cis* opening of the pore, but cannot translocate through due to their large size, thus quickly bounce away. This process results in an abundant number of short-lived blockages, with ($\%I_b/I_o \sim 30\%$), and $t_b^c \sim 100\mu$s. At 120mV applied voltage, the ratios between the frequency of Type 1 events and Type 2 events are approximately 1:36 (Na7.2), 1:29 (K7.2), 1:12 (Na10), and 1:14 (K10) (Tables 1 and 2). With increased application of applied voltages, frequencies of both Types 1 and 2 also significantly increased due to the high electrophoretic force. Most notably, as the voltage ramps from 100mV to 140mV, frequencies of Type 1 and Type 2 events rise approximately 2-fold, and 5- to 10-fold, respectively. Thus, single-channel analysis revealed that Type 2 events are far more plentiful than Type 1 events in all four samples (Na7.2, K7.2, Na10, and K10), implying that the abundancy of 15C-ssDNA was compacted by K^+ or Na^+ at pH 7.2 and 10 (Figure 9D-E and Figure 10D-E).

Table 2
15mer DNA event summary (pH 7.2)

	Type 1 events			**Type 2 events**		
	I/Io (%)	**Δt (ms)**	f **(s^{-1})**	**I/Io (%)**	**Δt (ms)**	f **(s^{-1})**
Na7.2	85.4 ± 1.2	0.42 ± 0.02	0.38 ± 0.03	25.8 ± 6.7	0.1 ± 0.01	13.9 ± 8.48
K7.2	84. ± 1.8	0.42 ± 0.04	0.22 ± 0.02	29.5 ± 5.5	0.1 ± 0.01	6.34 ± 0.35

Note: All values were recorded at 120mV+ applied voltage.

Table 3
15mer DNA event summary (pH 10)

	Type 1 events			Type 2 events		
	I/Io (%)	**Δt (ms)**	f**(s-1)**	**I/Io (%)**	**Δt (ms)**	f**(s-1)**
Na10	81.8 ± 0.6	0.55 ± 0.09	0.29 ± 0.01	24.7 ± 5.8	0.09 ± 0.01	3.4 ± 0.65
K10	84.2 ± 1.7	0.37 ± 0.03	0.17 ± 0.01	23.7 ± 6.9	0.10 ± 0.01	2.4 ± 0.9

Note: All values were recorded at 120mV+ applied voltage.

3.4.4 Comparing DNA potential of K+ and Na+. Single-channel recordings of the 15C-ssDNA reveal that, at both pH values of 7.2 and 10, most Type 1 events (translocation of linear ssDNA) cause a current blockage, $\%I_b/I_o$, of ~ 80-86%. At 120mV applied voltage, the ratios between Type 2 event frequencies, f^c(s-1), of Na7.2:K7.2 and Na10:K10 are 2.188:1 and 1.409:1, respectively (Table 2 and Table 3). Thus, the results indicate that between the two types of monovalent cations, Na^+ has a higher compacting potential for ssDNA.

Overall, the binding affinity and compacting potential of a monovalent cation to DNA strands are characterized by several factors, including but not limited to: (1) its hydrated radius, and (2) the chemical structure of the available nucleotides.[180, 207] It has been shown that a smaller radius corresponds to stronger hydration, because the small size of the hydrated cation allows for effective interactions with DNA because of higher Coulomb electrostatic potential.[180, 208] Subsequently, cations with a smaller hydrated radius can bind to DNA stronger than those with a larger hydrated radius.[180] In contrast, as a result of partial dehydration of monocation in the compact state, higher binding affinity of cations correlates to lower DNA compacting potential.[180] Thus K^+ has higher binding affinity, but lower DNA compacting potential than Na^+.[180]

Moreover, previous studies have pointed out that due to the differences in their molecular structures, the affinity for cations of the five nucleobases (guanine, adenine, cytosine, thymine, and uracil) are also varied.[209-212] Specifically, it was determined that the affinity Na^+ has for nucleobases, measured in kcal.mol^{-1}, is 42.1 for guanine, 40.9 for cytosine, 31.3 for adenine, 33.0 for thymine, and 32.3 for uracil.[210-211] With the aid of exocyclic amino groups, guanine and cytosine can be involved in cation binding through resonance effects; hence, have a much stronger cation binding than other nucleobases.[211] In our present work, the chosen ssDNA sequence contains only cytosine, which is one of the two nucleobases with the strongest binding affinity to Na^+; thus, the effect of Na^+ on inducing DNA compaction is significantly enhanced. Thus, the difference in our results in comparison to those of the previous group[54] mainly arises from the characteristic affinity of cations for nucleobases. The differences in event data traces caused by compacted and non-compacted DNA can be seen in Figure 11 and Figure 12.

In the previous study, because the sample DNA strand also contained adenine and thymine nucleotides,[54] the effect of cation binding and thus, induced DNA compaction, became much weaker and could not be observed with α-hemolysin nanopore (no Type 2 events). Thus, our findings in the current study fully agree with those from other laboratories,[180, 207] proving that Na^+ monovalent cations have a greater compacting potential on the 15C-ssDNA than K^+ cation does.

3.5 Conclusions

In this chapter we have investigated the interaction of monovalent cations (K^+ and Na^+) with ssDNA using an α-hemolysin biological nanopore platform. Although many studies have been done on capturing secondary structures formed by cytosine- or guanine-

rich sequences, the focus of those studies was usually using nanopore technology to capture the folding and unfolding of G-quadruplexes or i-motif structures. Thus, current-blockages caused by compacted DNA structures in a solution can be easily overlooked or neglected. Here, we presented our findings on the compacted structure of ssDNA induced by monovalent cations using biological nanopore. The study herein lay the groundwork for understanding and predicting the compacting process of ssDNA induced by monovalent cations (K^+ and Na^+) in nanopore-based studies. Moreover, in the recent years, many studies have been conducted to develop a nanocarrier for antisense oligonucleotides and single-stranded siRNAs in drug delivery, as well as in cancer gene therapy.[213-216] We believe that the ability of monovalent cations to compact a short oligonucleotide, as discussed in this current study, has a promising potential in the development of drug delivery and gene therapy.

Chapter 4

Investigation of Ckit1 G-Quadruplex Stability

4.1 Abstract

In this chapter, we investigated the formation of G-quadruplex on *Ckit1* promoter sequence, which plays an important role in several human malignancies. Ckit1 is a guanine-rich sequence that can form stable G-quadruplex structures in the presence of cations, causing gene downregulation and leading to suppression of cancer cell proliferation. Our study evaluated Ckit1 G-quadruplex structural stability dependence on cations and CX-5461. Conventional methods often utilize fluorescence spectroscopy and circular dichroism to study CX-5461 efficiency on stabilizing G-quadruplex. However, these methods cannot provide real-time molecular dynamic sensing of a structure's folding/unfolding behavior. This study utilized the nanocavity of a biological nanopore as the main tool for single-molecule analysis of Ckit1 G-quadruplex. Specifically, Ckit1 G-quadruplex formation and stability with and without CX-5461 presence was analyzed using α-hemolysin nanopore, circular dichroism and thermal denaturation. Our results showed that Ckit1 G-quadruplex stability is cation-dependent ($K^+ > Na^+$), resulting in longer current-blockage events and significantly decreased event capture rate with K^+. Addition of CX-5461 drug enhanced Ckit1 G-quadruplex stability resulting in extended current-blockage events (seconds- to minutes-long) in nanopore study and a significantly increased melting temperature. Using machine learning models, we were able to predict the binding state of CX-5461 to Ckit1 G-quadruplex in an Na+ environment, with a 92.4% sensitivity, 70.3% specificity and 81.3% average accuracy, respectively.

4.2 Introduction

G-quadruplexes are nucleic acid higher-order structures, made of one, two or four guanine-rich DNA or RNA strands. G-quadruplex consists of stacked G-quartets, which are cyclic arrangements of four guanines held together through Hoogsteen hydrogen bonds and stabilized by a central cation (e.g. K^+). G-quadruplex structures are frequently found in the telomeres and gene promoter regions. To date, over 300,000 putative G-quadruplex-forming sequences have been found in the human genome.[217] These structures play a functional role in regulating gene expression, controlling chromatin structure, and destructing chromosomal stability.[218-221] With more than 40% of human protein encoding genes containing one or more putative G-quadruplex forming motifs in the promoter regions, stabilizing G-quadruplex structures is a promising strategy to control gene expression at transcriptional and translational levels.[222]

CKIT is a proto-oncogene encoding the Ckit transmembrane tyrosine kinase receptor, which participates in the expression of genes, regulates cell differentiation, proliferation, migration, and resists cell apoptosis.[223-225] Overexpression and mutations of the Ckit gene have been found in a number of cancers, including pancreatic, leukemia and melanoma.[226-228] Promoter region of the Ckit gene contains three guanine-rich regions that can fold into stable G-quadruplex structures, known as Ckit1, Ckit2, and kit*.

For the Ckit proto-oncogene specifically, the presence of G-quadruplex structures can prevent cancer cells from replicating. Stabilization of Ckit gene G-quadruplex has been linked with inhibition Ckit tyrosine kinase receptor transcription and expression. This leads to cell proliferation and cause DNA damage that would kill cells without nucleic acid repair pathways, such as homologous recombination pathways.[229] Thus, down-regulation

of Ckit expression through stabilizing kit1 G-quadruplex is an attractive strategy for cancer therapy.[230-232] Certain molecules, such as CX-5461 (Figure 13), can bind and stabilize Ckit1 G-quadruplex structures; thus, effectively blocking the replication of cancerous cells.[233] Upon binding to Ckit1 G-quadruplexes, CX-5461 reduces the binding affinity of SL1 pre-initiation complex and RNA Pol I complex to rDNA promoters, leading to apoptosis.[234] This molecule was proven to be most efficient in tumorous cells that lack DNA repair mechanisms such as the BRCA1 and BRCA2 tumor suppressor genes.[235]

In this study, we report the changes in formability and stability of Ckit1 G-quadruplex in the presence of K^+ and Na^+ cation. Through electrical current signatures, we showed that binding of the CX-5461 molecule has a stronger stabilizing efficiency and further alters the volume of Ckit1 G-quadruplex in an Na^+ environment as compared to a K^+ environment. Ultimately, we employed a simple machine learning algorithm to predict the CX-5461 bound/unbound G-quadruplex.

4.3 Methods

4.3.1 Chemicals and reagents. Chemicals and reagents used in the experiments were purchased from the following vendors: 1,2-diphytanoyl-sn-glycero-3-phosphocholine (Avanti Polar Lipids); pentane (Sigma-Aldrich); potassium chloride (Sigma-Aldrich); sodium chloride (Sigma – Aldrich); Tris base (Promega); Ethylenediamine Tetraacetic acid (EDTA); pentane (Fischer Chemical); and hexadecane CH3(Ch2)14CH3 (Avantor). CX-5461 (Sigma Aldrich) was purchased in powder form, and suspended in acetic acid (VWR). CX-5461 solution was then aliquoted and stored at -20˚C.

4.3.2 DNA samples. Synthetic single-stranded DNA oligos were designed and purchased from Integrated DNA Technology (IDT – Coralville, IA). DNA oligos utilized in this study had the following sequences:

Control (linear): 5' – AGG GAT TTC GCT TTT AGG AGG G-3'

Ckit1: 5' – AGG GAG GGC GCT GGG AGG AGG G-3'

Ckit1 2 tails: 5'– dT10 - AGG GAG GGC GCT GGG AGG AGG G - dT10 – 3'

Upon delivery, all DNAs were suspended in a standard DNA storage buffer containing 10mM Tris, 1mM EDTA and titrated to 8.0pH. All samples were aliquoted into small volumes and stored at -80˚C until usage, in order to prevent multiple thaw-freeze cycles that can degrade and affect DNA quality. For both nanopore and CD experiments, DNA final concentration of 1µM was used.

4.3.3 Experimental setup. We employed an electrophysiology setup and followed a protocol described in previous studies.[53, 236] Each experimental chamber was filled with electrolyte solutions containing 1M KCl or 1M NaCl, buffered with 10mM Tris and 1mM EDTA titrated to 7.2pH. α-hemolysin protein was inserted into the bilayer from *cis* side forming a single-channel penetrated across the lipid bilayer. DNAs were inserted to the *cis* side. For experiments with CX-5461, electrolyte solutions were titrated to 8.3pH, in order to compensate for the acidity of CX-5461 solution.

4.3.4 Single channel recording. Single-channel recordings were performed using Axopatch 200B (Molecular Device Inc.), filtered with a built-in 4-pole low-pass Bessel Filter at 5kHz. Data were acquired with Clampex 10.7 (Molecular Device Inc.) and Axon Digidata 1550B A/D converter (Molecular Device Inc.), at a sampling rate of 20 kHz. The

data were analyzed using Clampfit 10.7 (Molecular Device Inc.), Excel (Microsoft) and R software.

4.3.5 Data analysis and visualization. Events with %I/Io < 50% were considered as brief interaction of DNA and the pore entrance, rather than DNA translocation; thus, disregarded from data analysis.[237] For every experimental condition, approximately 2000 DNA translocation events were included for the analysis. Event blockage amplitudes were determined from amplitude histograms by fitting the peaks to Gaussian function. The duration and occurrence of current-blockage event for DNA translocation were obtained by fitting the dwell-time histogram to an exponential distribution. Data were given as the mean ± SD, based on at least three separate pores. Plots were generate using Excel (Microsoft) and ggplot2 package (R).

4.3.6 Circular dichroism and thermal denaturation. Jasco CD spectropolarimeter, model J-810, were utilized for both normal absorbance and thermal denaturation studies. For all CD measurements, 1µM DNA concentration, CX-5461: DNA ratio of 5:1 (when applicable) were used, in consistency with nanopore studies. Spectra were capture in the 220-300 nm wavelength, 1.0 nm bandwidth, 50nm/min scanning speed, and a standard sensitivity. All experimental solutions contained 1M KCl or NaCl (for normal CD) and 0.1M KCl or NaCl (for thermal denaturation studies). The thermal denaturation was performed utilizing the same J-810 that was used for the CD portion of this study. In addition, a Jasco PTC-423S temperature attachment was used to increase the temperature from 20C to 95C at a rate of 1C/min. CD was recorded at temperatures below 100C to prevent the sample from boiling over and affecting the data. All CV results were exported as .csv files for further analysis and plotting with Excel.

4.4 Results and Discussion

4.4.1 Formation of Ckit1 G-quadruplex in K+ and Na+. In our experiment, we first employed circular dichroism (CD) to confirm the ability of Ckit1 single-stranded DNA (ssDNA) to adopt G-quadruplex structure. In CD, as circularly polarized light pass by, optically active chiral molecules (e.g. DNA and protein structures) will absorb the light differently, allowing them to be measured and quantified. In our study, CD spectra of Ckit1 DNA in Na^+ or K^+ solutions both exhibited strong negative and positive peaks at 237 and 263 nm, respectively (Figure 13D), confirming the formation of G-quadruplexes with parallel topology. Unlike telomeric sequence, whose G-quadruplex folding topologies depend on different cations,[105] the folding topology of Ckit1 G-quadruplex does not depend on cation nature (Na^+ vs. K^+).

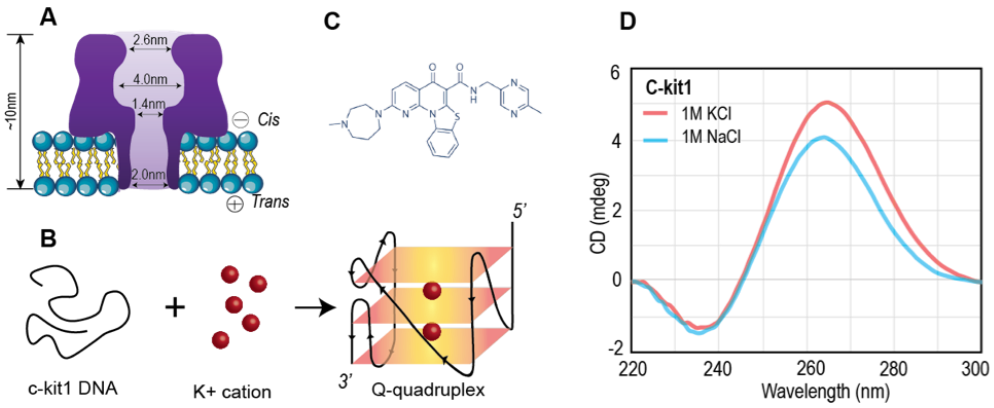


Figure 13. (A) Schematic cross-sectional structures and dimensions of an α-hemolysin nanopore embedded in lipid bilayers. (B) Spontaneous folding of a linear single-stranded Ckit1 DNA into G-quadruplex structure in the presence of K^+. The Ckit1 G-quadruplex composes of three G-quartets (each yellow layer) stacking on top of each other, and stabilized by K^+ cations. (C) Chemical structure of CX-5461 cancer drug, which has been shown to promote the formation and stabilize Ckit1 G-quadruplex structure. (D) CD spectra of Ckit1 DNA in the presence of K^+ or Na^+ consisted a positive peak at 263nm and a negative peak at 235nm. This result confirmed the presence a G-quarduplex structure with parallel folding topology

The transports events of Ctrl ssDNA and Ckit1 ssDNA through nanopores were recorded in 1M KCl and 1M NaCl, with and without the presence of CX-5641 molecules (Figure 14). Ctrl DNA has the same length with Ckit1 DNA with substitution of two opposite guanines of every G-tetrad with thymine (Figure 15). This substitution disrupts the formation of G-quartet, preventing the Ctrl sequence from folding into the G-quadruplex. Instead, Ctrl simply translocates through the α-hemolysin nanopore with a translocation time (Δt) of 2.68 ± 0.14 ms and capture rate (f) of 12.24 ± 0.13 s^{-1}. These values are in a similar range with those previously reported for linear, single-stranded DNA translocation through the nanopore.[238]

Compared to Ctrl, Ckit1 DNA yielded current blockage events with significantly longer Δt and reduced f (Table 1). Specifically, experiments on Ckit1 DNA yielded two major event types, named Type 1 and Type 2. Type 1 consists of independent short-lived blocks, whose open pore current blockage (%I/Io) was approximately 69.66 ± 1.81%. These spike-like events had similar amplitude and durations with those resulted by the linear single-stranded Ctrl DNA translocation through the nanopore. Whereas Type 2 events featured a deep current-blockage lasting milliseconds before returning to open-channel current level. The absence of Type 2 events in experiments with Ctrl DNA confirmed that these long-lasting events result from the translocation of Ckit1 G-quadruplex. Strong DNA secondary structure, such as G-quadruplex, require high energy to break up the stacked G-quadruplex structure, before it can enter the nanopore constriction site, causing a large temporal dispersion and prolonged translocation time.[53, 105, 239]

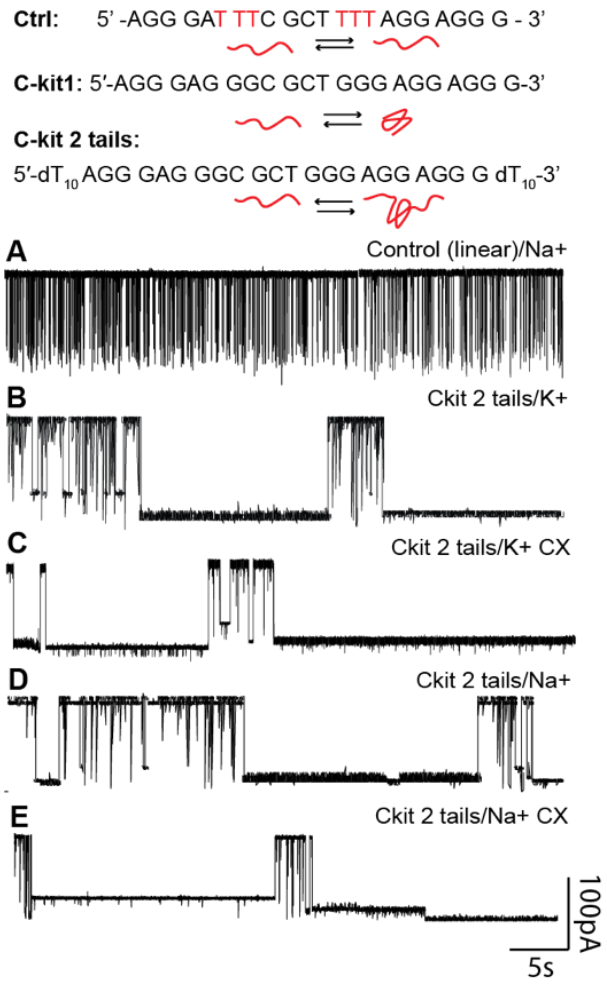


Figure 14. Data traces from a single α-hemolysin nanopore showing current-blockage events of: (A) Control DNA (Ctrl) in Na^+, and Ckit1 DNA in (B) K^+, (C) K^+/ CX-5461, (D) Na^+ and (E) Na^+/CX-5461. The control DNA is incapable of folding into secondary structure, causing spike-like events (top panel); whereas Ckit1 DNA spontaneously adopt the G-quadruplex structure in the presence of K^+ or Na^+, leading to long lasting current blockages. In all experiments, the concentration of both DNAs was 1 µM. All traces were recorded at +160mV in a 1M salt solution buffered with 10mM Tris, 1mM EDTA and calibrated to pH 7.2.

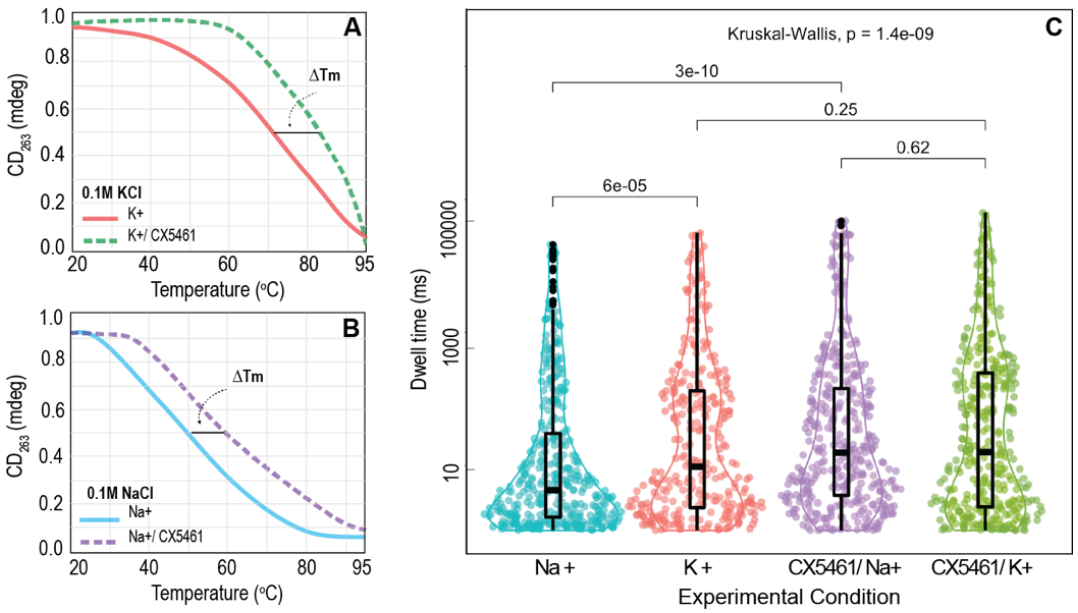


Figure 15. Stability of G-quadruplexes as determined by thermal denaturation analysis in (A) 0.1M KCl and (B) 0.1M NaCl with and without the presence of CX-5461 molecules. The secondary structure melting temperature, Tm, is a temperature, where 50% of the DNA are folded. (C) Violin plots represent the distribution of current-blockage event dwell time, Δt. Plots were arranged in the order of increasing mean Δt (Na_+ < K_+ < Na_+ with CX-5461 < K_+ with CX5661).

As Ckit1 DNA was instantaneously folded into G-quadruplex structure in the presence of monovalent cations (Na_+/ K_+), the remaining unfolded Ckit1 linear structure (Type 1 event) was significantly less than that of Ctrl DNA. Specifically, the occurrence rate, f, of Ckit1 Type 1 in K_+ was approximately 5.5 s-1, almost 2.25 times lower than the 12.24 s-1 for the Ctrl Type 1 event in the same solution. In order to determine the cation and stabilizing molecule dependency, we calculated the equilibrium formation constant for G-quadruplex in all experimental conditions. As shown in previous studies, equilibrium formation constant for G-quadruplex, K_f, is the concentration ratio between folded DNA and linear DNA in the mixture.

From the correlation between DNA concentration and event capture rate, K_f can also be calculated using the equation: $K_f = \frac{f_{ctrl}}{f_{ckit1,L} - 1}$.[53] In which, f_{ctrl} is the capture rate of the linear, single-stranded Ctrl DNA and $f_{ckit1,L}$ is the capture rate of linear Ckit1 DNA. The calculated K_f for Ckit1 G-quadruplex was 3.83 in K^+, 1.84 in Na^+, 4.65 in K^+/CX-5461, and 4.23 in Na^+/CX-5461.The equilibrium formation constants suggest the sequence of most favored experimental conditions for Ckit1 G-quadruplex formation is: K^+/CX-5461 > Na^+/CX-5461 > K^+ > Na^+ (as shown in Table 4).

Table 4
Ckit1 current blockage event summary

		Δt (ms)	$f_{Ckit1,linear}$ (s-1)	K_f	Ku	k_f
Ct	Na+	2.68 ± 0.14	12.24 ± 0.13	-	-	-
Ckit1	K+	3.38 ± 0.09	4.19 ± 0.19	3..83	0.30	1.13
	Na+	3.15 ± 0.10	7.64 ± 0.20	1.64	0.32	0.58
	K+/CX-5461	6.06 ± 0.39	3.63 ± 0.14	4.65	0.17	0.77
	Na+/CX-5461	6.03 ± 0.21	3.89 ± 0.11	4.24	0.17	0.70

4.4.2 Stability of Ckit1 G-quadruplex. A thermal denaturation experiment was performed to determine the stability of the G-quadruplex structures by comparing its melting temperature in Na^+ and K^+ solutions, as well as investigating stabilization effect of CX-5461. With the experimental solution being water-based, thermal denaturation studies

must be halted under 100˚C, in order to prevent evaporation of sample. As Tm of Ckit1 G-quadruplex is approximately 88˚C in 1M NaCl, we decreased the solution ionic strength (while keeping DNA concentration at 1µM), which allowed us to record the melting temperature of the structures at a range sufficiently far away from the evaporation point as to not affect the data (Figure 16). We decided that 0.1M would be the ideal concentration for further testing. The final thermal denaturation study was performed in 10mM Tris-HCl buffer (pH 7.2) containing 100mM cation (Na^+ or K^+). When applicable, CX-5461 concentration was 5 µM, which corresponds to a 5:1 ratio of ligand to DNA. Reduction of ionic strength resulted in a lowering of Tm of unbound Ckit1 G-quadruplex to 70˚C and 49˚C in 0.1M KCl and 0.1M NaCl solutions, respectively (Figure 15A – B). Addition of CX-5461 increases Tm of Ckit1 G-quadruplex structure, resulting in a ΔTm of 13˚C and 11˚C in K^+ and Na^+ (Figure 15A – B). Higher melting temperature in K^+ than in Na^+ proves that the structure is more stable in K^+.

In order to visualize the distribution of G-quadruplex translocation time through the pore, we constructed a violin plot for dwell time of events in each condition (Figure 15C). It is important to note that only events lasting longer than 1ms were included in the violin plots, as those with shorter dwell time were considered as translocation of unfolded linear ssDNA (Type 1 events). The constructed violin plots for the dwell time, Δt, of Ckit1 type 2 events exhibited a bimodal distribution with two populations of events lasting from milliseconds to several seconds (Figure 14C). Using paired-wise testing, we found a statistically significant difference (95% confidence interval) in dwell time when Type 2 events in Na^+ compared to in K^+ and Na^+/CX-5461. Despite an overall increase in event

dwell time in the presence of CX-5461, there was no statistical difference between event dwell time in K^+ and K^+/CX-5461.

Previous studies showed that the unfolding rate constant (ku) of G-quadruplex is an indicator of structural stability and can be estimated using Δt, through the equation: $k_u = \frac{1}{\Delta t}$.[53] Moreover, the kinetic folding rate constant, k_f, is the product of the equilibrium formation constant and unfolding rate constant, or: $k_f = K_f k_u$.[53] Using these equations, we calculated the folding and unfolding rates for Ckit1 G-quadruplex, as shown in Table 4. In the absence of CX-5461, Ckit1 G-quadruplex folded much more rapidly in a K^+ environment than in an Na^+ environment (with a k_f of 1.13 and 0.59, respectively), while unfolding more slowly. On the other hand, in the presence of CX-5461, k_u and k_f of Ckit1 G-quadruplex were similar between Na^+ and K^+ environments. (see Table 4).

There are multiple factors that contribute to the stabilization of G-quadruplexes and their topology, including stacking interactions, hydrogen bonding, solvation, and cation bonding.[240-241] To stabilize the G-tetrad stacks, the binding cation needs a sufficient ionic strength, in order to compensate electrostatic repulsion between the phosphate oxygens of four Guanines, instead of two in regular DNA duplexes.[241] In general, K^+ induced G-quadruplex is much more stable and has longer lifetime than Na^+ induced G-quadruplex, owing to the higher ionic strength and lower hydration energy of K^+ ions.[241-244]

Furthermore, the smaller Na^+ cations can be coordinated within the plane of tetrads, and occupy a range of positions with lower steric constraints; thus, reducing the electrostatic repulsions.[241] While Na^+ can alternatively occupy coordination sites and become closer to the core of the tetrads, Na^+ G-quadruplex confers a higher plasticity that allows a better fit with a binding molecule.[241] On the other hand, K^+ is coordinated at the

cavity between two G-tetrads and bridges together eight Guanines O6 atoms, thereby becoming more stable and rigid.[241]

Previous studies involving thrombin binding aptamer (TBA) found that ΔTm between K^+ and Na^+ coordinating TBA decreases from 29˚C (free) to 12˚C (bound).[241] This can partially explain what we observed in our study: compared to Na^+, K^+ cation creates a much more stable G-quadruplex structure. Presence of CX-5461 in K^+ solution produced extremely stable structures that could not be included in data analysis, as the event lasted several hundred seconds to minutes. Therefore, the enhanced stability induced by CX-5461 was better observed in Na^+ experimental buffers. This was reflected by significant increase in event dwell time in Na^+ (with CX-5461 versus without), but not in K^+ buffer.

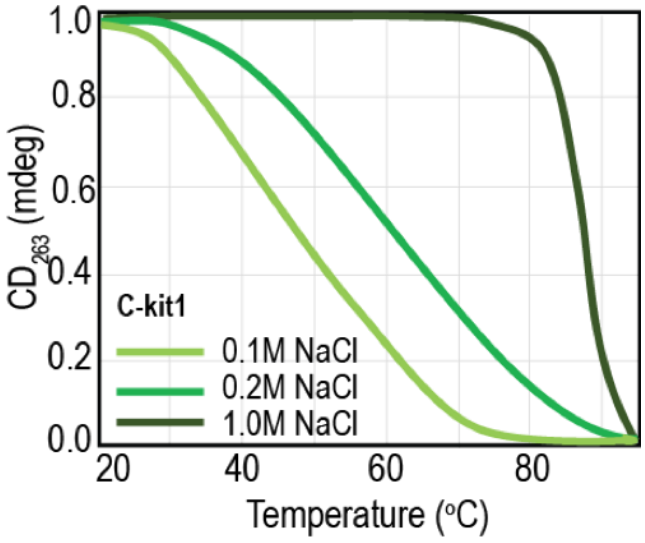


Figure 16. Thermal denaturation of Ckit G-quadruplex in buffers containing: 1M, 0.2M or 0.1M NaCl. Melting temperature, ΔTm, is determined when there are 50% structure unfolded. ΔTm of Ckit1 G-quadruplex reduces approximately 30˚C as salt concentration decrease 10-fold. DNA concentration was held constant at 1μM for every experiment.

4.4.3 Changes in volume of Ckit1 G-quadruplex. As single-channel data traces of Type 2 (i.e. G-quadruplex translocation) exhibited multiple events with well-defined step-wise structure that varied in length (Figure 17A–D), we decided to examine the hidden subgroups in Type 2 events of Ckit1 DNA. For this, we utilized an unsupervised machine learning technique, which can identify similarities and patterns in datasets, to cluster Type 2 events using their amplitude (%I/Io) and dwell time (Δt). Clustering is one of the most common data analysis techniques used to explore the structure of the data. Specifically, we used the kmeans function in R to classify the Type 2 translocation events in each experimental condition. K-means is a randomly-initialized iterative clustering method that partitions the dataset into k predefined distinct subgroups and the positions of each group's centroid.

Traditionally, dwell time is typically used as the main indicator to classify distinct translocation events. However, the amplitude of current-blockage events is known to reflect the size and interaction of the targeted molecule with the nanopore. Thus, we selected both %I/Io and Δt as features for unsupervised machine learning. The two selected features (%I/Io and Δt) have greatly different measurement scales of 0 – 100% and 1 – 1000s ms, respectively. To avoid bias introduced by one group being dominant in size, we first converted each feature to a 0 to 1 scale. After clustering and calculation of centroids, data were converted back to their original values for better visualization and comparison between conditions. To determine the optimal parameter k for all data subset, we constructed an elbow plot, which suggested that the optimal number of clusters in nanopore Type 2 events was k = 2 (Figure 18). Figure 17E–H represents the scatter plots of the ionic traces with respect to %I/Io and Δt.

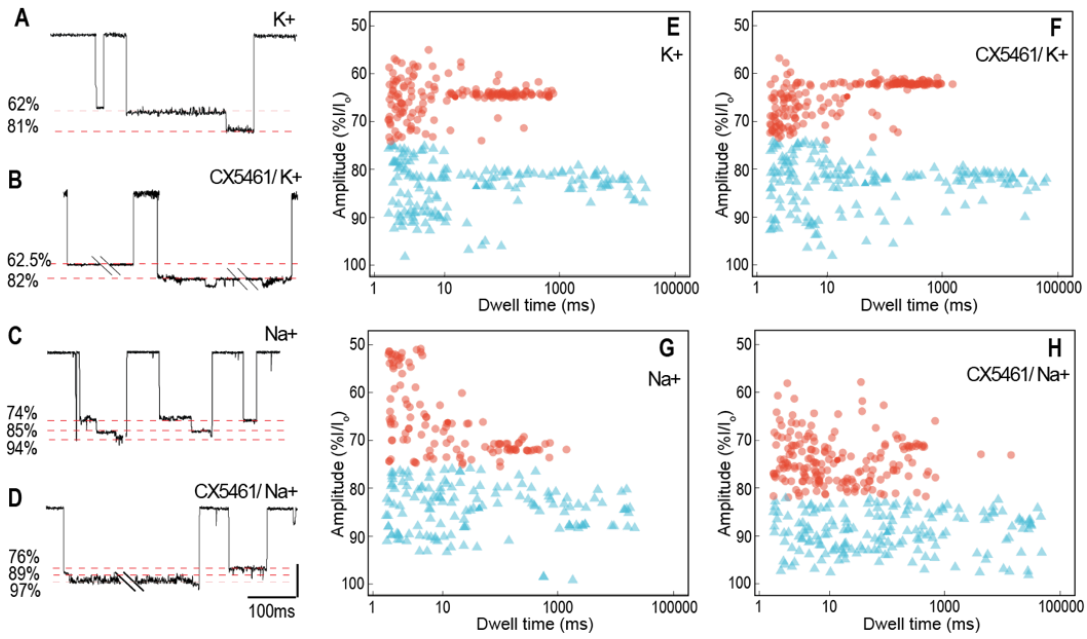


Figure 17. Clustering analysis for all samples. Examples of Ckit1 G-quadruplex data traces in buffers containing: (A) 1M KCl, (B) 1M KCl with CX-5461, (C) 1M NaCl or (D) 1M NaCl with CX-5461. Current-blockage events exhibit step-wise changes in amplitude, representing different stages of G-quadruplex threading and translocating through the α-hemolysin pore. (E-H) Clustering analysis of all current-blockage events in all experimental conditions using k-mean method. In each solution, there are two main clusters of events with shallow (red) or deep current blockage amplitudes (blue), reflecting the step-wise current blockage amplitude changes. At least 1000 events were included in each scatterplot. Events with $\Delta t \leq 1$ms were excluded in this scatter plotting since those events were linear DNA translocation.

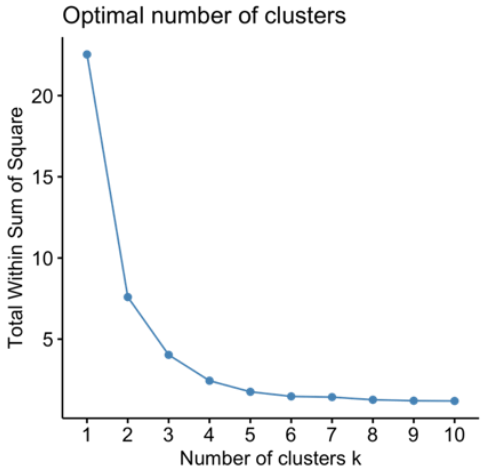


Figure 18. Elbow plot for determining optimal number of clusters for amplitude-dwell time scatterplot.

In each plot, there are two event clusters (showed in red and blue), named Type 2a and 2b, respectively. Although the event translocation time is largely dispersed (as discussed previously), event amplitude seemed to be the main deciding factor in Type 2 subgroups classification, as shown in Figure 17E–H. These level changes (Type 2a and 2b) reflected the two stages of G-quadruplex translocation through the nanopore, which were: (a) capturing and (b) unfolding of the DNA secondary structure (Figure 19B). The average amplitude of Type 2a and 2b events were 67.4±3.9% and 85.1±2.9%, respectively.

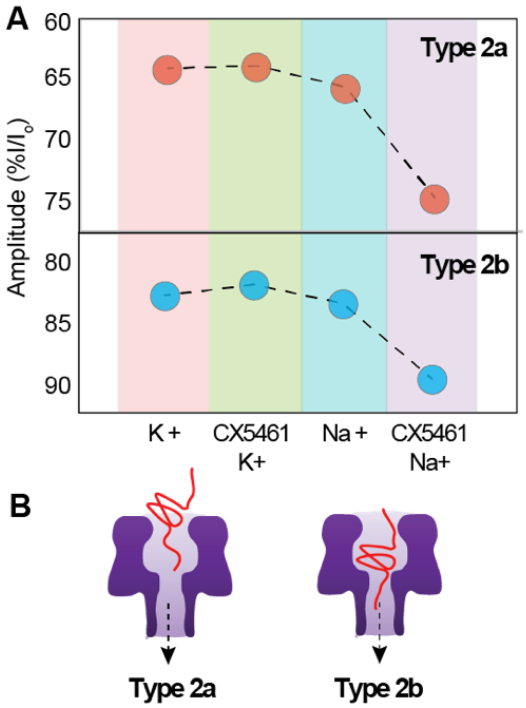


Figure 19. Ckit1 G-quadruplexes volume. (A) Comparing changes in Ckit1 current-blockage amplitudes of level 1 and 2 in buffer solutions contains K^+ or Na^+ (with and without CX-5461). The presence of CX-5461 exhibits minimal effect on Level 1 and level 2 %I/Io in K^+, but significantly increases %I/Io of both levels in Na^+ solutions. (B) Schematic explaining different stages of Ckit1 G-quadruplex translocation through α-hemolysin nanopore, corresponding to current-blockage level 1 and 2.

Previous studies involving TBA G-quadruplex and beta-Cyclodextrin (β-CD) showed that the amplitude of current-blockage events reflected the change in volume of the G-quadruplexes formed by different cations.[53, 245] Thus, the current-blockage amplitude, %I/Io, of event Type 2a and 2b reflect an overall trend that volume of K^+ G-quadruplex is smaller than Na^+ G-quadruplex (Figure 19A). This has been explained in previous studies on G-quadruplexes. Briefly, as the Na^+ ion is smaller (d = 0.95 Å), it can fit in the plane of a quartet; while the larger K^+ ion (1.33Å) is coordinated in the center of the cavity between two planes. Position of cations within the G-quadruplex, in turn, changes the cation-carbonyl distance (d), explaining the reason why volume of K^+ G-quadruplex is smaller than Na^+ G-quadruplex, as observed in previous studies and our results herein.[53] Specifically, $d_{K\text{-}O}$ is shorter than $d_{Na\text{-}O}$, giving rise to stronger attraction between cation and the carbonyl, and a more stable G-quadruplex in K^+ presence. Interestingly, the binding of CX-5461 changed the volume of Ckit1 G-quadruplex differently. As shown in Figure 17E–H and Figure 19A, while the volume of K^+ G-quadruplex remains mostly unchanged, volume of Na^+ G-quadruplex significantly increased when bound by CX-5461, reflecting through a 10 and 5 pA increase in the amplitude of event Type2a and 2b, respectively.

4.4.4 Prediction of CX-5461 binding status. Having determined the changes in formability and stability of Ckit1 G-quadruplex in correlation with cations and stabilizing molecules, we sought to demonstrate the application of α-hemolysin nanopore as a molecular screener for G-quadruplexes. To do this, we used a supervised machine-learning to train classifiers for pattern recognition and discriminate Ckit1 G-quadruplex formed in the different experimental conditions discussed thus far. In each experimental condition,

we randomly selected 2000 current-blockage events, then divided the datasets into 80% and 20% for model training and validation, respectively. Two features, Δt and average %I/Io, collected from the current blockages were used as inputs to a k-nearest neighbor (KNN) classifier. We first trained the model using all event captured by the nanopore. This first run yielded an average accuracy and sensitivity of 69.6 ± 2% and 72.2 ± 5%, respectively. The model's prediction ability was quite low, due to the fact that Ckit1 DNA could instantaneously fold and unfold into G-quadruplex structure. Thus, the events captured by nanopore comprised of both linear and folded DNA (Type 1 and 2 events), as previously discussed. Thus, we retrained the model using only events lasted longer than 1ms, in order to exclude Type 1 events (linear DNA). With the filtered datasets, the model performance was enhanced significantly, yielding a new average accuracy and detection sensitivity of 81.3 ± 4 % and 92.4 ± 2 %. Table 5 shows the classification accuracy, sensitivity, and specificity the model for different experimental condition pairs.

Table 5
KNN classification results for G-quadruplexes formed in different conditions

Distinguish Ckit1 G-quadruplex formed in		**Sensitivity (%)**	**Specificity (%)**	**Balanced Accuracy (%)**
Condition 1	Condition 2			
K+	Na+	93.55	77.78	85.66
CX-5461/K+	KCl	89.66	58.33	73.99
CX-5461/Na+	Na+	95.45	70.00	82.73
CX-5461/Na+	CX-5461/K+	90.91	75.00	82.95

**Note: Samples in Group 1 were considered as the positive result for classification.*

Results showed that the model's ability to distinguish one type of G-quadruplex from another is in the order of: (Na^+/CX – Na^+) > (K^+ – Na^+) > (Na^+/CX – K^+/CX > K^+/CX – K^+. As we discussed earlier, the most pronounced changes in the structure's volume and stability occurred when CX-5461 binds to Na^+ G-quadruplex; whereas no statistically significant changes were found in for Ckit1 G-quadruplex in K^+ solution with and without CX-5461. Thus, the order of the model's performance could be explained. A limitation in our prediction model comes from having nonhomogeneous control samples. When working with dynamic experimental conditions, such as those investigated here, the process of ssDNA folding/unfolding and of CX-5461 binding/unbinding to the G-quadruplex structure constantly occurs. This lowers the model performance, as the training dataset are nonhomogeneous.

4.5 Conclusions

In summary, we have presented the ability of α-hemolysin nanopore as a real-time molecular dynamic sensing for Ckit1 G-quadruplex formation and stability in K^+ and Na^+ environments. While G-quadruplex structures can be extremely stable, their topology and stability depend on many factors, including the length and sequence composition of the total G-quadruplex motif, the size of the loops between the guanines, strand stoichiometry and alignment, and the nature of the binding cations.[240] Our study showed that Ckit1 G-quadruplex has a parallel folding topology in both environments, with a more tightly packed structure in the presence of K^+ cations. Binding of CX-5461 molecules significantly enhanced the stability and increased the volume of Ckit1 G-quadruplex in Na^+, but not in K^+. In the absence of CX-5461, Na^+ G-quadruplex and K^+ G-quadruplex featured greatly different equilibrium properties. Specifically, Na^+ G-quadruplex folds much slower and

unfolds rapidly, compared to K^+ G-quadruplex. On the other hand, presence of CX-5461 significantly enhanced the folding rate of linear Ckit1 DNA and slows down the unfolding rate of G-quadruplex, leading to a similar equilibrium folding kinetic constant in both K^+ and Na^+.

Apart from determining the structural formation and stability, we further employed clustering analysis to determine the hidden pattern of the nanopore current blockage amplitudes, which have distinct levels reflecting the two stages of (i) capturing and (ii) G-quadruplex translocation through the pore. The volume of G-quadruplex is correlated with the cation species and presence of CX-5461. These changes in the structure volume and stability made it feasible to employ machine learning algorithms to classify G-quadruplexes formed in different experimental conditions. With the ability to computationally predict the binding stage of cation and CX-5641 molecule to G-quadruplex, we believe α-hemolysin nanopore could be an interesting tool for real-time monitoring the binding reference of G-quadruplex to cations and stabilizing molecule in mixture solutions.

Chapter 5

Mutations on Ckit1 G-Quadruplex Reduce CX-5461 Efficacy

5.1 Abstract

In this article, we employed α-hemolysin biological nanopore to the effect of point mutation on Ckit1 G-quadruplex stability and drug efficacy. Our study employed six DNA samples including Ckit1 and its mutated sequences (containing 1-6 mutations). The signature blocks in the nanopore revealed that number and position of mutation have an impact causing destabilization of the G-quadruplex structure. Specifically, sequences with the same number of mutations on the same G-quartet layer exhibited similar stability. Furthermore, CX-5461 significantly increases G-quadruplex structural stability only in sequences with no mutation on the top G-quartet. Moreover, using a random forest classifier on nanopore data, we were able to distinguish mutated from unmutated Ckit1 sequences with an overall sensitivity and specificity of 81.9% and 83.4%, respectively. Understanding the effects of mutation on G-quadruplex stability is beneficial for constructing methods to predict treatment response in personalized medicine approaches.

5.2 Introduction

In this study, we investigated the efficacy of CX-5461 on mutated C-kit1 promoter sequences to elucidate the importance of each component G-tetrad in the overall stability of the G-quadruplex structure. Activating mutations on C-kit1 have been observed in several types of malignancies, notably leukemia, melanoma, and gastrointestinal tumors.[246-248] Our study employed a biological nanopore sensor and circular dichroism spectroscopy to compare the stability of Ckit1 G-quadruplex and its mutated sequences in the presence of sodium cations (Na^+). Our earlier work revealed that while the overall

structural stability is lower in Na^+ than K^+ environment, the effect of CX-5461 on stabilizing the Ckit1 G-quadruplex was more pronounced in Na^+ (see Chapter 4). We found that CX-5461 affects the stability of the mutated Ckit1 G-quadruplex to varying degrees, depending on the mutation position and number. Having more than one mutation on the top G-quartet might inhibit binding of CX-5461 to the structure, leading to no significant change in G-quadruplex stability. In addition, using nanopore data and machine learning algorithms, we were able to distinguish mutated sequences from unmutated Ckit1 with 81.9% sensitivity and 83.4% specificity.

5.3 Methods

5.3.1 Chemicals and reagents. The chemicals and reagents used in this study along with the associated vendors are as follows: Sodium chloride (Sigma-Aldrich); 1,2-diphytanoyl-sn-glycero-3-phosphocholine (Avanti Polar Lipids); pentane (Sigma-Aldrich); Tris base (Promega); Ethylenediaminetetraacetic acid (EDTA) (Sigma-Aldrich); pentane (Fischer Chemical); hexadecane CH3(Ch2)14CH3 (Avantor); CX-5461 lyophilized powder (Sigma-Aldrich); and acetic acid (VWR). The CX-5461 was then aliquoted into 1.5 mL microcentrifuge tubes which were wrapped in aluminum foil and stored in a freezer at -20˚C.

5.3.2 DNA samples. All DNA used in this study were designed and purchased from Integrated DNA Technology (IDT - Coralville, IA) and had the following sequences:

Ckit1: AGG GAG GGC GCT GGG AGG AGG G (0 mutation)

L1M1: AGG GA**T** GGC GCT GGG AGG AGG G (1 mutation)

L1M2: AGG GA**T** GGC GCT **T**GG AGG AGG G (2 mutations)

L2M1: AGG GAG **T**GC GCT GGG AGG AGG G (1 mutation)

L2M2: AGG GAG **T**GC GCT G**T**G AGG AGG G (2 mutations)

M2C: AGG GA**T** **TT**C GCT **TTT** AGG AGG G (6 mutations)

Upon delivery, DNA samples were resuspended in a standard storage buffer (containing 10mM Tris, 1mM EDTA and titrated to 8.0 pH). All samples were then aliquoted and stored at -80˚C until experimentation. This was done to prevent the degradation of the DNA quality caused by multiple thaw-freeze cycles. In both the nanopore and circular dichroism experiments, the final concentration of the DNA was 1 µM.

5.3.3 Experimental setup. Detailed protocol and description of the α-hemolysin nanopore experimental setup has been discussed in previous studies.[47, 50] A 1 µM aliquot of DNA sample was inserted into the *cis* chamber. The *cis* chamber was grounded, so that when applying a positive voltage, the negatively charged DNA would be driven through the nanopore to the *trans* side. Electrolyte solutions in the *cis* chamber were titrated to pH 8.15 in experiments using CX-5461, in order to compensate for the acidity of CX-5461 solution.

5.3.4 Data analysis and visualization. Event blockage amplitudes were determined from amplitude histograms by fitting the peaks to Gaussian functions (single-sample and mixture models). The duration and occurrence of short-lived blockages for DNA translocation were obtained by fitting the dwell-time histogram to an exponential distribution. Only current-blockage events with %I/Io ≥ 60 % and Δt > 1ms were selected for data analysis.[73] Those with %I/Io < 60% or Δt < 1ms are considered as brief interaction or translocation of linear ssDNA, respectively. Data were given as the mean ± SD, based

on at least three separate experiments. RStudio was used to run ANOVA and Tukey's HSD tests to determine the significance of dwell-time data.

5.3.5 Circular dichroism and thermal denaturation. In the studies for both absorbance and thermal denaturation, a model J-810 Jasco CD spectropolarimeter was utilized. When measuring CD, a CX-5461:DNA (1 μM DNA concentration) ratio of 5:1, when applicable, was used, which is consistent with pre-existing nanopore studies. Spectra were captured within a 220-300 nm wavelength range, 1.0 nm bandwidth, 50nm/min scanning speed, and a standard sensitivity. 1M NaCl experimental solutions were used in normal CD studies and 0.1M NaCl experimental solutions were used in thermal denaturation studies. J-810 is utilized in both normal CD and thermal denaturation portions of this study. In order to increase the temperature from 20˚C to 95˚C, a Jasco PTC-423S temperature attachment with a heating rate of 1˚C/min was utilized. To prevent the sample from boiling and affecting the data, CD was recorded only at temperatures below 100˚C. The results of the CD were exported as .csv files to be analyzed further and plotted in Excel.

5.4 Results and Discussion

5.4.1 Experimental design. This study aims to explore the effects of point mutations on Ckit1 G-quadruplex formation and stability, as well as the effects of cancer drug CX-5461 on mutated C-kit1 G-quadruplexes versus native C-kit1 G-quadruplexes. In this study, we employed variations of the Ckit1 DNA, each containing different mutations of guanine. Specifically, six DNA samples included: Ckit1 (positive control), L1M1, L1M2, L2M1, L2M2, and M2C (negative control). Ckit1 is the unmutated sequence, while M2C contains six mutated guanine bases in total, with two at each G-quartet layer. The L/M scheme in sample naming convention was in reference to the position of G-quartet

layer ("L") and number of mutation ("M"), respectively. For example, "L1M2" refers to C-kit1 DNA that has two-point mutations on the part of the sequence that would form the first layer of the G-quadruplex. In Figure 20, the first column shows schematics of the Ckit1 G-quadruplex, with the three layers of G-quartets (yellow layers), its central Na^+ cations (red dots), and location of the mutated guanine (purple circles). As the G-quadruplex is symmetrical, we only ran experiments with mutations on the first and second G-quartet, assuming CX-5461 effects on G-quadruplex stability if bound to the first or third layer would be similar. In order to increase efficiency of the experimental design, we ran experiments with C-kit1 DNA with up to two mutations. The G-quadruplex requires four guanine bases in order to form. Therefore, two mutations on the C-kit1 G-quadruplex sequence would have a sufficient effect on destabilizing of the G-quartet.

5.4.2 Changes in current-blockage events among conditions. First, we compare raw data traces obtained from nanopore experiments of all samples in the same experimental conditions (1M NaCl, +160mV, with and without presence of CX-5461). As showed in Figure 20, all samples exhibited two types of events, which are spike-like (Type 1 events) and long-lasting blockages (Type 2 events). In Ckit1 DNA, addition of CX-5461 significantly increases the dwell time and frequency of Type 2 events (as discussed in Chapter 4). However, compared to Ckit1, raw nanopore data traces of mutated sequences exhibited a higher amount of spike-like events (Type 1), indicating that the mutated sequences were less likely to form stable G-quadruplexes. Moreover, increased incidence of Type 1 events also reflects higher amounts of linear DNA being captured. This result is much more pronounced in the L2M2 and M2C samples, compared to other mutated sequences. L1M1 appears to have the most similar pattern to Ckit1 DNA, compared to

other samples. Similar to Ckit1 and L1M1, L1M2 and L2M1 data traces contain both Type 1 and 2 events; however, the increased frequency of Type 1 events indicated less G-quadruplex formation.[113] Interestingly, addition of CX-5461 to L2M2 has shown slight increases in the signal event length when compared to L2M2 without the addition of CX-5461 (Figure 20).

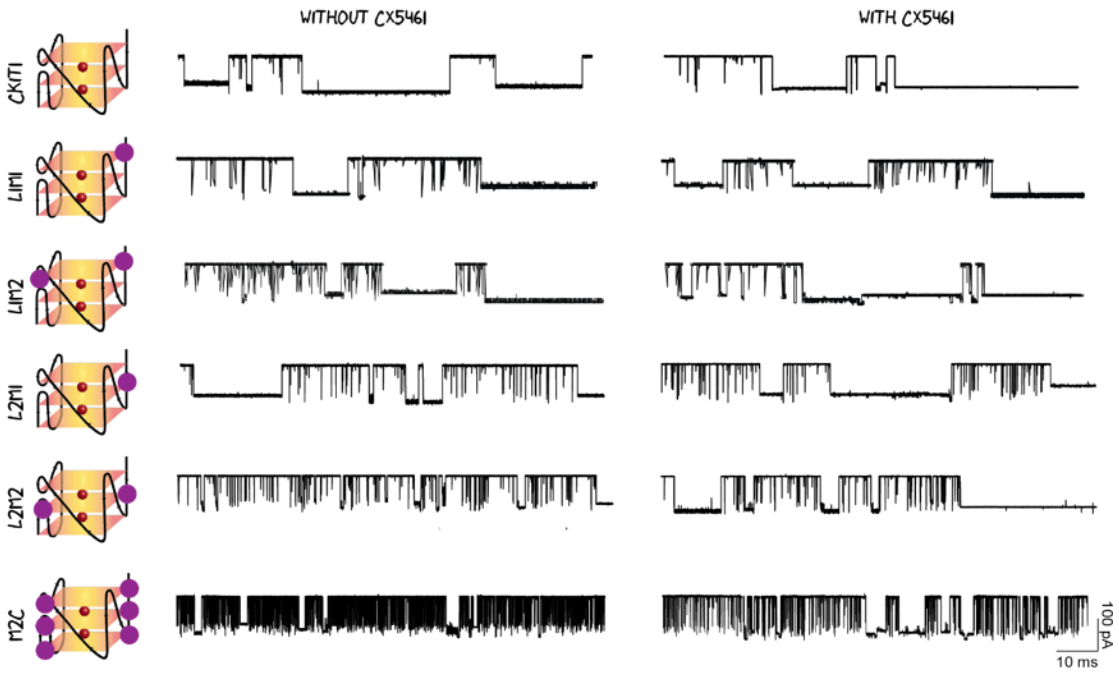


Figure 20. Nanopore raw data traces. (Left) Schematics represent six variations of the Ckit1 sequence, containing 0 – 6 mutation points (purple circles) and their location on the G-quadruplex structure. Experiments were conducted in duplicate with and without CX-5461 for all DNA sequences, including: positive control DNA (Ckit1), mutated sequences (L1M1, L1M2, L2M1, L2M2 and M2C) (top to bottom). (Middle) The positive control (C-kit1) DNA spontaneously adopts the G-quadruplex, leading to long lasting current blockages (top panel); whereas M2C (containing 6 point mutations) is incapable of folding into a unimolecular G-quadruplex, causing a significant increase in frequency of spike-like events (bottom panel). (Right) Addition of CX-5461 did not show a clear visible effect on mutated sequences. In all experiments, the concentration of both DNAs was 1 µM. All traces were recorded at +160mV in a 1M salt solution buffered with 10mM Tris, 1mM EDTA and calibrated to pH 7.2.

Using circular dichroism (CD), we measured the optical absorbance of all the sequences investigated in this study (Figure 21). Both Ckit1 and Ckit1_CX formed a parallel G-quadruplex structure that was characterized by a strong positive peak at 263nm and a negative peak at 237nm (Figure 21A). Compared to Ckit1, the CD absorbance spectra of mutated sequences exhibited both positive and negative peaks between 220 – 300nm, as shown in Figure 21B – F.

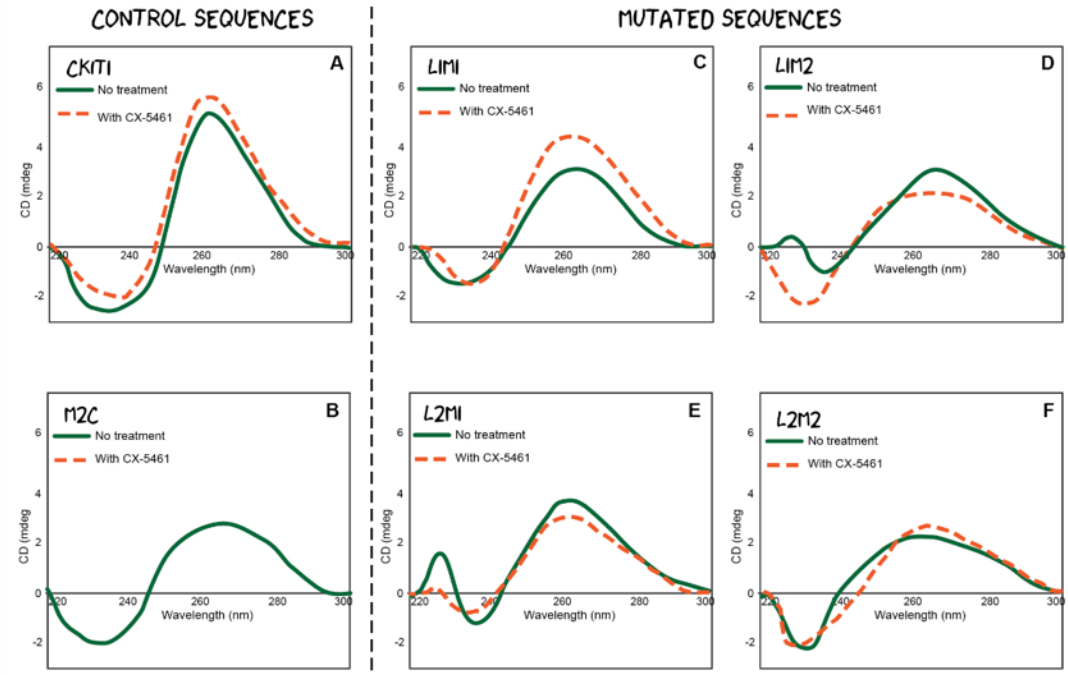


Figure 21. Circular Dichroism spectra of all tested sequences were measured between 220 – 320nm, at room temperature. Samples were measured without (black line) and with the presence of CX415461 drug (dotted orange line). CD spectra of C-kit1 DNA had a positive peak at 263nm and a negative peak at 235nm, reflecting the presence a G-quadruplex structure with parallel folding topology. For all other sequences, these two peaks (at 235nm and 263nm) were broadened and shifted, indicating a change in the DNA structural topology.

L1M1 and L1M1_CX spectra were similar to that of Ckit1 and Ckit1_CX, respectively (Figure 21C). However, the peaks were significantly broadened and/or shifted in L1M2, L2M1, L2M2 and M2C samples. Diminishing of 263nm and 235nm peaks in mutated sequences indicate a reduction in parallel G-quadruplex formation and an increase in topological changes. This effect varies with the number and position of mutations. In the following sections, we will take a closer look at the structural volume and stability of these structures.

5.4.3 Structural volume of Ckit1 and mutated sequences. We would like to understand if the volume of Ckit1 G-quadruplex is affected by presences of mutations and CX-5461. Previous work has shown that the potential volume of the secondary structure of the DNA could be determined by the current blockage amplitudes.[104] For each sample, we selected at least 1000 current-blockage events that lasted longer than 1 ms for the analysis, as those with shorter dwell time were considered as translocation of unfolded linear ssDNA (Type 1 events). In accordance with the nanopore conductance with (I) and without (Io) sample translocation, we calculated the current-blockage amplitude (%I/Io). Without CX-5461, all samples folded into structures with a similar volume (for summary statistics, see Table 6). Specifically, the mean current blockage amplitude of all samples was approximately 69.7 ± 2.3% (Figure 22). Addition of CX-5461 greatly increased the current-blockage amplitude of Ckit1 (similarly observed in Chapter 4) and M2C by over 5%, but exhibited little to no effect on other mutated structures (Figure 22).

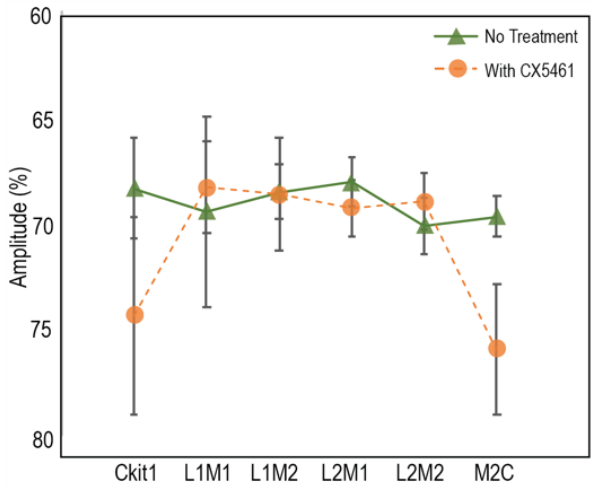


Figure 22. Average current blockage amplitude at +160mV of all samples, including: Ckit1, L1M1, L1M2, L2M1, L2M2 and M2C without treatment (green) and with the addition of CX-5461 (orange). All samples exhibited a similar average current-blockage amplitude of 69 ± 2%. Addition of CX-5461 showed minimal effect on L1M1, L1M2, L2M1 and L2M2, but increased %I/Io of both Ckit1 and M2C by approximately 5%. The average %I/Io of each sample was found by selecting at least 1000 events to construct a density plot, following by fitting with Gaussian normal distribution.

Table 6
Summary statistics of event amplitude (%I/Io)

Condition	With CX	Mean (%)	Max (%)	Min (%)	Median (%)	Std (%)
Ckit1		71.85	101.87	50.15	71.21	10.94
Ckit1	x	72.89	100.76	50.08	74.27	10.08
L1M1		68.68	86.26	51.22	67.64	9.34
L1M1	x	67.35	101.63	50.14	66.20	8.34
L1M2		70.78	98.71	50.48	69.77	8.58
L1M2	x	71.87	101.13	51.02	69.93	9.81
L2M1		72.03	101.61	50.41	69.13	9.05
L2M1	x	72.78	99.04	50.76	70.89	8.14
L2M2		72.75	99.04	50.76	70.88	8.14
L2M2	x	73.79	94.00	51.84	71.07	8.09
M2C		68.81	98.34	53.75	67.47	5.42
M2C	x	77.23	99.91	52.67	76.09	6.75

5.4.4 Effect of CX-5461 on mutated Ckit1 G-quadruplex stability. Stability of the G-quadruplex structure is measured in relation to length of the current blockage. First, we constructed a box-and-whiskers plot for all the experimental conditions (Figure 23A). Upon calculating the summary statistics of these samples, we found that the majority of the event dwell times were between 1-100ms (with the exception of Ckit1_CX), with median Δt decreasing in the sequence of: Ckit1 > L2M1 > L1M1 > M2C > L2M2 > L1M2. Addition of CX-5461 increased the samples' Δt to varying degrees, except for L2M1 and M2C, whose Δt significantly decreased (Table 7). Specifically, the median Δt of all samples in the presence of CX-5461 can be listed in decreasing order as follows: Ckit1 >> L1M1 > L2M2 > L1M2 > M2C > L2M1 (Table 7). In all experimental samples, there were multiple outliers, which were events with extremely long Δt (several seconds to minutes). This was a result of metastable state of DNA—a phenomenon that was well described in previous studies.[249-251] Thermal denaturation experiments were performed to assess the stability of the Ckit1 G-quadruplex and its mutated structures in each experimental condition (Figure 22B – E). The melting curves were collected at 263 nm (which was the positive peak in the circular dichroism experiments), with temperatures ranging from 20 - 95^{o}C. The melting temperature is determined when 50% of the DNA is unfolded. As discussed in Chapter 4, while the nanopore experiments were conducted in buffer containing 1M NaCl, thermal denaturation experiments were carried out in 0.1M NaCl environment, in order to keep the melting temperature of DNA secondary structure below 100˚C.

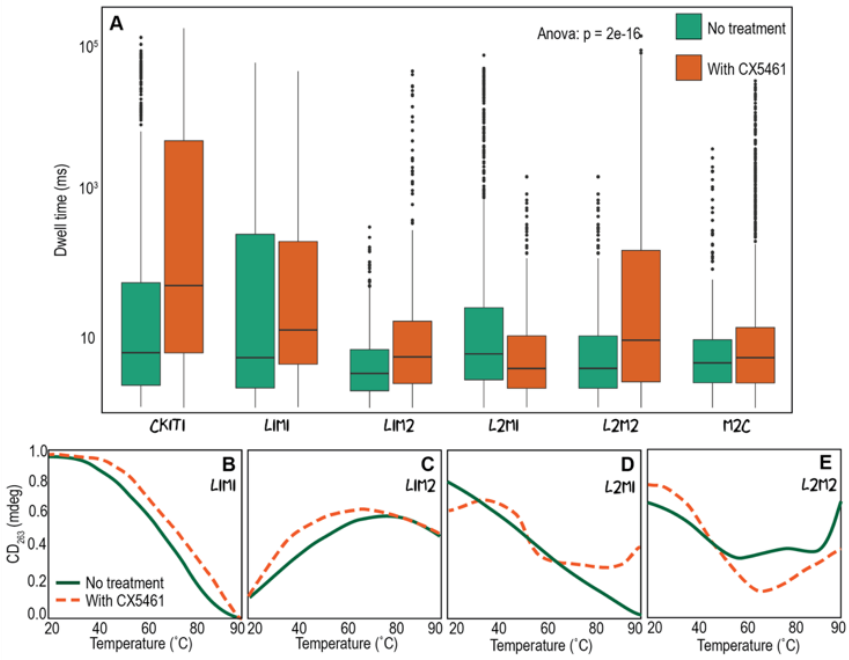


Figure 23. Stability of mutated Ckit1 G-quadruplex. (A) Grouped boxplot representations of the distribution of event dwell times for all samples, including: Ckit1, L1M1, L1M2, L2M1, L2M2 and M2C. All experiments were conducted in duplicate with (orange, right side) and without (green, left side) the addition of CX-5461. Ckit1 DNA is relatively more stable than its mutated sequences, as indicated by a higher median dwell time. Presence of CX-5461 increases stability of most sequences to varying degrees, except for L2M1, whose stability decreases. (B) Thermal denaturation studies were performed to determine the stability of G-quadruplex structures formed by mutated sequences with (orange dotted line) and without (solid green line) the presence of CX-5461. The secondary structure melting temperature, Tm, is the temperature, where 50% of the DNA are folded. Circular dichroism spectra were collected at 263nm, with temperature ramping from 20 – 95˚C.

Among the samples, The L1M1 condition exhibited a melting curve that was remarkably similar to the positive control Ckit1 DNA, with the addition of CX-5461 slightly shifting the melting curve to the right. Specifically, the observed melting temperature of L1M1 and L1M1_CX were approximately 65˚C and 70˚C, respectively (Figure 23B). This indicates excellent stability of the G-quadruplex both with and without CX-5461. Notably, Tm of L1M1 was slightly higher than that of Ckit1 in the same experimental buffer. This increase in structural stability was also reflected through a slight

increase in nanopore event dwell time of L1M1 sequence in comparison to Ckit1 DNA. However, this increase in dwell times was not statistically significant. On the other hand, L2M1 exhibited a straight down trending line and no inflection point in the melting curve (Figure 23D). The addition of CX-5461 to L2M1 somewhat rescued the sigmoidal shape that was characteristic of the melting curve of the positive control, but offbeat degrees of ellipticity persisted, indicating poor stability of the G-quadruplex. In other samples (e.g. L1M2, L2M2, and M2C), the behavior bore little to no resemblance to the positive control Ckit1 DNA with relatively low G-quadruplex stability at its peak degree of ellipticity. Moreover, the addition of CX-5461 did not rescue this aberrant behavior.

Table 7
Summary statistics of event dwell time (ms)

Condition	With CX	Mean (ms)	Max (ms)	Min (ms)	Median (ms)	Std
Ckit1		4432.63	114411.35	1.05	5.75	13919.90
Ckit1	x	11587.29	152850.00	1.02	47.36	26625.04
L1M1		4283.26	51802.34	1.02	4.94	10577.08
L1M1	X	1887.54	39545.14	1.04	11.73	6195.79
L1M2		9.43	296.95	1.02	3.00	24.21
L1M2	X	629.38	39771.09	1.02	5.04	3552.46
L2M1		2153.12	65396.62	1.02	5.53	7404.17
L2M1	x	31.49	1447.89	1.03	3.50	135.46
L2M2		31.42	1447.89	1.03	3.51	135.28
L2M2	X	6175.41	119939.50	1.02	8.54	16886.62
M2C		15.27	3449.96	1.02	4.17	123.22
M2C	X	314.18	29303.41	1.02	4.94	1869.56

5.4.4.1 Having more than one mutation destabilizes Ckit1 G-quadruplex. All three of the examined sequences with at least two mutations within the same level had significantly different dwell times than the positive control Ckit1 DNA ($p < 0.001$ for

L1M2, L2M2, and M2C), with all exhibiting decreased dwell times (as shown in Table 8). This indicates that having multiple mutations within the same level of Ckit1 compromises the overall stability of the G-quadruplex, while having just one mutation did not. Furthermore, there were no significant differences between the dwell times of L1M1 and L2M1 or L1M2 and L2M2, but there were significant differences between L1M2 - L1M1 and L2M2-L2M1, suggesting that the number of mutations has a much more relevant impact on G-quadruplex stability than the position (G-quartet level) of the mutation (Table 9).

5.4.4.2 The outer G-quartet layer is crucial to the binding of CX-5461. Addition of CX-5461 created significant changes in the stability of Ckit1, L2M1 and L2M2 with opposite effects, suggesting different interactions of CX-5461 to the mutated Ckit1 structures (Table 10). As expected, the addition of CX-5461 to the positive control Ckit1 DNA proved to significantly increase Ckit1 G-quadruplex stability, as discussed in Chapter 4. This was also the case with L2M2 ($p < 0.001$), which makes sense given that addition of CX-5461 somewhat rescued the shape of the denaturation curve. CX-5461 did not appear to have significant effects on L1M1, L1M2, or M2C ($p = 0.091$, $p = 0.976$, $p = 0.967$, respectively), although L1M1 did exhibit a notable, but not significant, increase in dwell time with the addition of Cx-5461. Because CX-5461 stabilizes G-quadruplex by binding to the DNA sequence,[252] this may implicate the outer G-quartet (L1) is critical for an efficiency binding of CX-5461), because its mutation results in abrogation of the stabilizing effect. Having more than one mutated guanine on the first layer inhibits the formation of layer 1 G-quartet, thus interfering with CX-5461 binding to the structure. This is supported by the ability of CX-5461 to rescue stabilization of the G-quadruplex in the

Ckit1 and L2M2 conditions, both of which lacked mutations in the first layer of G-quartet (L1). However, addition of CX-5461 to L2M1 produced a significant difference that resulted in a decrease in dwell time, and thus a decrease in structural stability. This is indicative of a destabilizing interaction of unknown cause that L2M1 has with CX-5461. This follows the trend that was illustrated in the thermal denaturation curve produced by L2M1, where the addition of CX-5461 resulted in a drastically offbeat degrees of ellipticity from the expected melting curve.

Table 8
Statistical comparison between dwell times of mutated sequences and Ckit1.

Groups	Adj. p-value	
L1M1-Ckit1	1.0000	ns
L1M2-Ckit1	0.0000	***
L2M1-Ckit1	0.0003	***
L2M2-Ckit1	0.0000	***
M2C-Ckit1	0.0000	***

**Note: adjusted p-values was obtained from using Anova on all samples, followed by Tukey's post-hoc test.*

Table 9
Statistical comparison between dwell times of mutated sequences.

Groups	Adj. p-value	
L1M2-L1M1	0.0000	***
L2M2-L2M1	0.0003	***
L2M1-L1M1	0.0950	ns
L2M2-L1M2	1.0000	ns
L2M2 - M2C	1.0000	ns
L1M2 - M2C	1.0000	ns
L1M1 - M2C	0.0000	***
L2M1- M2C	0.0000	***

**Note: adjusted p-values was obtained from using Anova on all samples, followed by Tukey's post-hoc test.*

Table 10
Effect of CX-5461 on Ckit1 DNA and mutated sequences.

Groups	Adj. p-value	
Ckit1_CX-Ckit1	0.0000	***
L1M1_CX-L1M1	0.0907	ns
L1M2_CX-L1M2	0.9761	ns
L2M1_CX-L2M1	0.0003	***
L2M2_CX-L2M2	0.0000	***
M2C_CX-M2C	0.9666	ns

**Note: adjusted p-values was obtained from using Anova on all samples, followed by Tukey's post-hoc test.*

5.4.5 Distinguishing unmutated Ckit1 sequences from mutated ones. We explored the feasibility of detecting mutated Ckit1 sequences by employing machine learning classification models on data obtained from nanopore single–molecular analysis. In particular, there were five models, including: K-nearest neighbor (KNN), Random Forest (RF), Logistic Regression (LG), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA), employed for the binary classification of mutated and unmutated Ckit1 DNA.

To minimize the potential issue of overfitting one group, we sampled a balanced dataset, in which the number of mutated and unmutated sequences are represented eqrrrrrually (i.e. approximately 1000 events each). The dataset was comprised of two distinct groups. The first group, called "Ckit1", was comprised of the current-blockage events of Ckit1 DNA. The second group, called "Mut", consisted of an equal mixture of all the mutated sequences (e.g. L1M1, L1M2, L2M1, L2M2, and M2C). The dataset was then split into 80% and 20% portions for model training and validation purposes, respectively. Current – blockage amplitude (%I/Io) and dwell time (Δt) were selected as input features for the model. To evaluate the model's performance, we compared their

AUROC values. We ran the same five algorithms in duplicate on a dataset containing (1) all current-blockage events and (2) only those with %I/Io > 75%. The first one (1) yielded extremely poor results, with all algorithms returning an AUROC value of approximately 0.5 (data not shown). Results obtained from the latter dataset (2) were similar, except for RF, whose median AUROC value was approximately 0.89 ± 0.04 (as shown in Figure 24A).

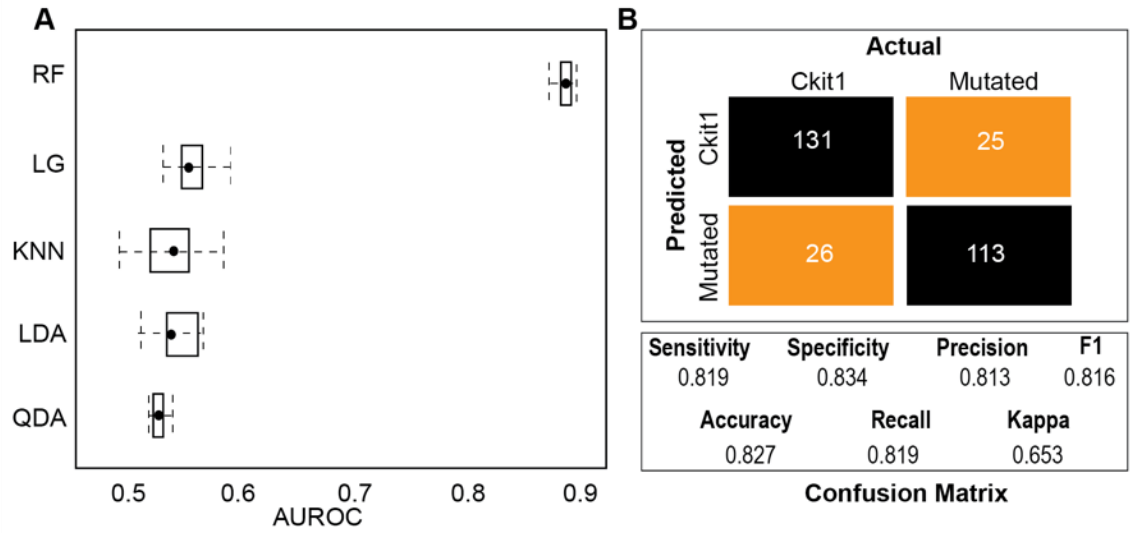


Figure 24. Machine learning models for classification of mutated C-kit1 sequence from unmutated one. (A) Evaluation of model performance based on the distribution of their AUROC value: models were ranked based on the median AUROC score and arranged from best to worst performance (top to bottom). (B) Confusion matrix shows results of random forest classification for mutated sequence from unmutated one using nanopore events. Unmutated sequences contained Ckit1 DNA and mutated group consists of all other sequences (i.e. L1M1, L1M2, L2M1 and L2M2). The random Forest classifier model yielded an overall sensitivity of 81.9%, specificity of 83.4% and balanced accuracy of 82.7%.

To visualize the true positive and false positive rates, we constructed a confusion matrix for the RF model (Figure 24B). This matrix shows the true positive and true negative prediction corners (black boxes) at the top left and bottom right, respectively. On the other

hand, the false negative and false positive predictions are placed in the lower left and upper right corners (orange boxes). Thus, for mutated and unmutated Ckit1 sequences, the RF classifier yielded an 82.7% accuracy, 81.9% sensitivity and 83.4% specificity. As determined before, point mutations have varying effects on the secondary structural volume and stability. It was surprising to see that most classifiers yielded a poor performance since point mutations can increase or decrease the mutated structure's volume and stability concurrently (depending on the number and position of mutation). This hindered the model performance. Applying the same classifiers on datasets containing Ckit1 and only one mutated sequence will enhance the accuracy of classifiers. However, this would not be practical in a real-world scenario, where the number and position of point mutation is unknown.

5.5 Conclusions

Through this nanopore single-molecule study, we determined the effect of point mutations on stability of Ckit1 G-quadruplex and CX-5461 stabilizing efficiency using the event dwell time. With a maximum width of approximately 2.1 nm,[253] the G-quadruplex is able to fit within the 2.6 nm wide nanocavity entrance. However, the β-barrel entrance (constriction site) of the nanopore is just 1.4 nm wide, making it impossible for the G-quadruplex structure to translocate without unfolding. Therefore, electrophoretic force drives the analyte into the nanocavity where it becomes trapped and either (1) unravels and translocates through the β-barrel and into the *trans* side of the test chamber, or (2) regurgitates backward into the *cis* solution.[47] Previous studies have demonstrated that translocation events with G-quadruplex have significantly longer dwell times on the order of several seconds due to the extra time needed for the structure to unravel while it is

trapped in the nanocavity.[47, 105] By contrast, simple linear DNA produces very short, transient current blockages.[47] Taken together, this indicates that dwell time provides a valuable assessment of the structural stability of the G-quadruplex, with higher dwell times indicating a more stable structure.

Results obtained from our study suggest that the number and position of mutations have an impact on G-quadruplex structural stability. Furthermore, addition of CX-5461 to samples does not have a significant effect on the stability of samples with a mutation on the top G-quartet (layer 1), implicating it as a crucial factor in the binding of CX-5461 to the DNA structure. Lastly, addition of CX-5461 to conditions with an intact level 1 (Ckit1 positive control, L2M1, and L2M2) resulted in an increase in G-quadruplex stability, with the exception of L2M1 which oddly exhibited a decrease in stability. Using Random Forest algorithm on nanopore data, we were able to distinguish mutated Ckit1 sequences from unmutated ones with an overall sensitivity of 81.9%, specificity of 83.4% and balanced accuracy of 82.7%. Excessive activation of the C-kit1 gene is implicated in several human malignancies and its expression is downregulated by the stabilization of its G-quadruplex structure.[254] Knowing the effect of mutation on drug efficiency can have implications in the implementation of cancer therapies, particularly the drug CX-5461. As mutations on G-quadruplex sequences have been observed in several cancers, this research may be helpful in constructing methods to predict therapeutic responses and personalizing cancer treatment.

Chapter 6

Employing LiCl Salt Gradient to Slow Down DNA Translocation for Label-Free Detection of Cytosine Methylation

6.1 Abstract

In this chapter, we demonstrate a label-free detection, biological nanopore-based method to distinguish methylated cytosine (mC) from naked cytosine (C) in sample mixtures containing both C and mC at prolonged translocation duration. Using 15-fold increase in LiCl salt concentration going from *cis* to *trans* chamber, we increased the translocation dwell time of ssDNA by over 5-fold and the event capture rate by 6-fold in comparison with symmetric concentration of 1.0M KCl (control). Moreover, salt gradients can create a large electric field that will funnel ions and polymers towards the pore, increasing the capture rate and translocation dwell time of DNA. As a result, in 0.2M – 3.0M LiCl solution, ssDNA achieved a prolonged dwell time of 52 μs/nucleotide and a capture rate of 60 ssDNA per second. Importantly, lowering the translocation speed of ssDNA enhances the resulting resolution, allowing 5'-mC to be distinguished from C without using methyl-specific labels. We successfully distinguished 5'-mC from C when mixed together at ratios at 1:1, 3:7 and 7:3. Distribution of current blockade amplitudes of all mixtures adopted bimodal shapes, with peak-to-peak ratios coarsely corresponding to the mixture composition (e.g. the density and distribution of events shifted in correspondence with changes in 18b-0mC and 18-2mC concentration ratios in the mixture).This chapter was adopted from our published article (Vu, Trang, et al. "Employing LiCl salt gradient in the wild-type α-hemolysin nanopore to slow down DNA translocation and detect methylated cytosine." *Nanoscale* 11.21 (2019): 10536-10545.)

6.2 Introduction

Cytosine methylation is the most well-established and understood epigenetic modification on DNA, affecting the activity and stability of genomic regions.[118, 255-256] Specifically, methylation of cytosine is the gain or loss of a methyl group on cytosine in the 5-position,[56, 257-258] which is frequently found at *CpG* dinucleotides.[256] These *CpG* nucleotides often occur in bunches widely known as "islands" which are linked to promoter regions of genes. Aberrant methylations of these regions can lead to gene inactivation or loss-of-function mutations.[259] Beside DNA promoter regions, cytosine methylation of various types of RNA (e.g. tRNA, mRNA, and non-coding RNAs) has gained increasing attention in recent years.[260-263] For instance, it was shown that specific signature methylation pattern of microRNAs (miRNA1, miRNA9, miRNA124 and miRNA137) could be used to identify ulcerative colitis patients with elevated risk for colorectal neoplasia.[264] Unlike genetic mutations, which can occur anywhere in a gene, cancer-specific cytosine methylation mostly locates at defined regions (e.g. the promoter of genes or on specific sequences), making it easier to devise targeted probes for molecular alterations.

Currently, bisulfite-conversion is the "gold-standard" to detect and profile DNA methylated cytosine.[265] However, bisulfite-conversion cannot directly identify 5-methylated cytosine (5-mC) from native DNA. Unmethylated cytosine must first be converted to uracil, consequently, increasing the complexity of library preparation and the potentials for artifacts and biases. Due to extensive fragmentations in the bisulfite-conversion process, any breaks at inserts can prevent downstream amplification of DNA.[265] Recently, several studies using both solid-state and biological nanopores have

demonstrated different ways to detect and quantify the presence of 5'-mC on both ss- and dsDNA by long-read sequencing,[266-267] employing methyl-CpG-binding domain protein (MBD),[56, 268] Kaiso-Zinc Finger (KZF),[56] optical-tagging,[57] among many others.[145-148, 275] Comparing to PacBio and bisulfate conversion, methylation detection with nanopore technology has yet to match the performance, but shows advantages on maintaining the DNA sequence complexity and imprinting.[267]

In this chapter, we investigated the ability of wild-type α-hemolysin nanopore in label-free detection of methylation on cytosine, with a focus on slowing down translocation velocity to enhance the result readout and accuracy. Through research and analytics, several methods have been proposed to impede translocation velocity via decreasing experimental temperature, altering bulk density, changing applied potential, or using an alternating electric field.[269] However, compared to those physical conditions, the effect of cation nature, electrophoretic, electroosmotic, and diffusioosmotic flows are far more significant on the translocation of DNA.[269-271] In our study, electrophoretic, electroosmotic, and diffusioosmotic flows have been adopted to slow down the translocation velocity of ssDNA through biological nanopores, enabling label-free detection of 5-mC from C in ssDNA mixtures.

6.3 Methods

6.3.1 DNA samples. Synthetic 18-mer, single-stranded DNA purchased from Integrated DNA Technology (Coralville, IA) have the following sequences:

Sample 1 (18b-0mC): 5'-TAA TCA TCG CGT ACT AAT-3'

Sample 2 (18b-2mC): 5'-TAA TCA T**M**G **M**GT ACT AAT-3' (with **M** represents 5'-mC)

DNAs were suspended in a buffer solution containing Tris (10mM) and EDTA (1.0mM), titrated to 8.0 pH, then aliquoted and stored at -80$_{o}$C until usage. For all experiments, DNA final concentration of 1 μM in the *cis* chamber was used.

6.3.2 Experimental setup. Each chamber was filled with electrolyte solutions with either LiCl or KCl solution, with a salt concentration of 0.2M, 1.0M or 3.0M, buffered with 10mM Tris and titrated to pH 7.2. The salt concentrations in each chamber varied based on the experimental conditions being tested.

6.3.3 Single-channel recording. Single-channel recordings were performed using Axopatch 200B (Molecular Device Inc.), filtered with a built-in 4-pole low-pass Bessel Filter at 5kHz. Data were acquired with Clampex 10.7 (Molecular Device Inc.) and Axon Digidata 1550B A/D converter (Molecular Device Inc.), at a sampling rate of 20 kHz. The data were analyzed using Clampfit 10.7 (Molecular Device Inc.), Excel (MicroSoft) and MatLab (2017a) software. Event blockage amplitudes were determined from amplitude histograms by fitting the peaks to Gaussian functions (single-sample and mixture models). The duration and occurrence of short-lived blocks for DNA translocation were obtained by fitting the dwell-time histogram to an exponential distribution. Only current-blockage events with %I/Io ≥ 70% were selected for data analysis.[35] Those with %I/Io < 70% are considered as brief interaction of DNA and the pore entrance (e.g. interaction of hairpin structure with vestibule), rather than DNA translocation, thus, were disregarded.[237] Data were given as the mean ± SD, based on at least three separate experiments.

6.3.4 Mixture analysis. Current blockage events were calculated as the percent blockage of the open pore current (%I/I_o). To predict the distribution of the two samples in mixture solutions, %I/I_o was first calculated, then fitted with Gaussian Mixture Model (two components) in MatLab (2017a).

6.4 Results and Discussions

We employed the α-hemolysin nanopore as a single channel sensing element to detect ssDNA at single-molecule level. Experiments were performed on two single-stranded DNA samples with identical sequences, containing either cytosine, or 5-methylcytosines at the center, hereafter, called C-DNA and mC-DNA. C-DNA has a sequence of 5'-TAA TCA TCG CGT ACT AAT-3', and mC-DNA has a sequence of: 5'-TAA TCA T**M**G **M**GT ACT AAT-3' (with **M** standing for 5'-mC). Most of the studies have largely investigated long nucleotide sequences (hundreds to thousands bases), thus, in this study, we selected 18-mer sequences to investigate the effect of electrophoretic, electroosmotic and methylation hydrophobicity on shorter strands, such as tRNA and miRNA. All experiments were reproduced multiple times, yielding comparable results. The open pore conductance varies upon the change of solute ionic concentration and gradient. Let C_{trans}/C_{cis} be the ionic concentration ratio between the *trans* and *cis* chambers. With $C_{trans}/C_{cis} > 1$ (higher salt concentration in the *trans* side), there is a constant diffusion of cation from the *trans* to *cis* chamber, creating a negative current even if there is no external voltage applied.

The α-hemolysin nanopore is water-filled, thus, the charge distribution inside the pore is relatively consistent, reflected by the approximately linear I-V curves of α-

hemolysin nanopore (Figure 25). The 5kHz low-pass Bessel filter in Axopatch 200B and 20,000 s^{-1} sampling events were selected for all measurements in this study.

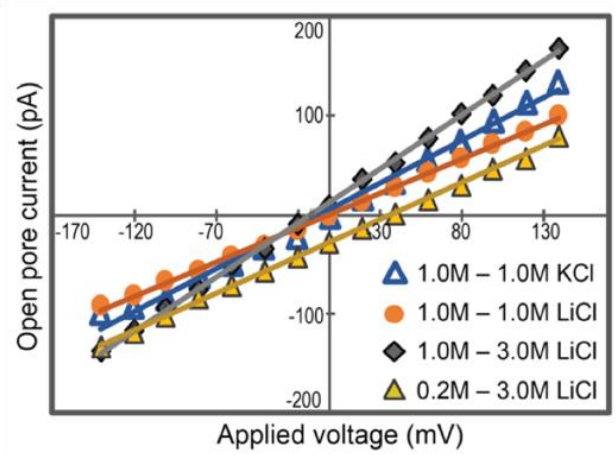


Figure 25. Current -voltage (I-V) curves of the α-hemolysin nanopore at various experimental conditions.

6.4.1 Li+ and salt gradient increase translocation time. Cations, depending on their ionic radii and concentration, can affect the electrophoretic mobility of DNA differently.[272-274] To examine this, we conducted two sets of experiments on both samples, 18b-0mC and 18b-2mC DNAs, at various conditions of: (1) symmetric and (2) asymmetric salt concentration in the *cis* and *trans* chambers. For symmetric 1.0M – 1.0M solutions, we found that the ratio for the translocation time, Δt, of ssDNA is KCl : LiCl ~ 1 : 2.3 (Figure 26A – B). When comparing within the same type of sample (either 18b-0mC or 18b-2mC) translocated through the pore at +120mV, as the ionic concentration in the *trans* side increases from 1.0M LiCl to 3.0M LiCl, the translocation time, Δt, increased approximately 1.3 times (Figure 26 C – D). More notably, the addition of a salt gradient is shown to further

intensify the magnitude of ssDNA translocation time through the nanopore. With a 15-fold increase in ionic concentrations going from the *cis* to the *trans* chambers (0.2M – 3.0M LiCl), the average translocation times of 18b-0mC and 18b-2mC DNA are 0.945 ms and 0.765 ms, respectively, at biased voltage of 120 mV. In other words, the addition of a salt gradient (0.2M – 3.0M LiCl in *cis/trans*) increases the translocation times of both samples by 2-fold and 5-fold, when compared to the results obtained from 1.0M – 1.0M LiCl and 1.0M – 1.0M KCl solutions of the same sample type (either 18b-0mC or 18b-2mC, as shown in Figure 26A – D). Note that the open pore current depends on the electrolyte's conductivity and ionic concentration. Thus, amplitudes of open pore current decreases with decreasing ionic concentration in the *cis* side, reflected by current blockages with various amplitudes (Figure 26E – F).

Overall, an increase in salt concentration gradient going from the *cis* to the *trans* chamber results in a proportional elongation of individual DNA translocation time. Notably, in any given experimented solutes and applied voltages, mC-DNA translocate through the nanopore faster than C-DNA does. This result agrees with findings from a previous study, showing that addition of a methyl group on Cytosine (mC) increases the local hydrophobicity and rigidity of G-mC bps, thus leading to a smoother and faster translocation of mC-DNA through the nanopores.[275]

To understand the changes in Δt observed in this experiment, we scrutinize the effect of Li^+ ions on ssDNA and its surrounding environment. Li^+ ions can effectively reduce the mobility of ssDNA through the α-hemolysin nanopore, because of its counter-ion effect and physical-chemistry properties. In theory, electrolyte solution containing monovalent counter-ions can alter the electrophoretic mobility of DNA in different ways,

including changing water's bulk properties,[276] modifying DNA conformation, and affecting the preferential binding of counter-ions to DNA.[272, 277-279] More specifically, binding of counter-ions partially neutralizes the negatively charged phosphate backbone; which in turn decreases the overall effective net charge and mobility of DNA molecule.[272, 280]

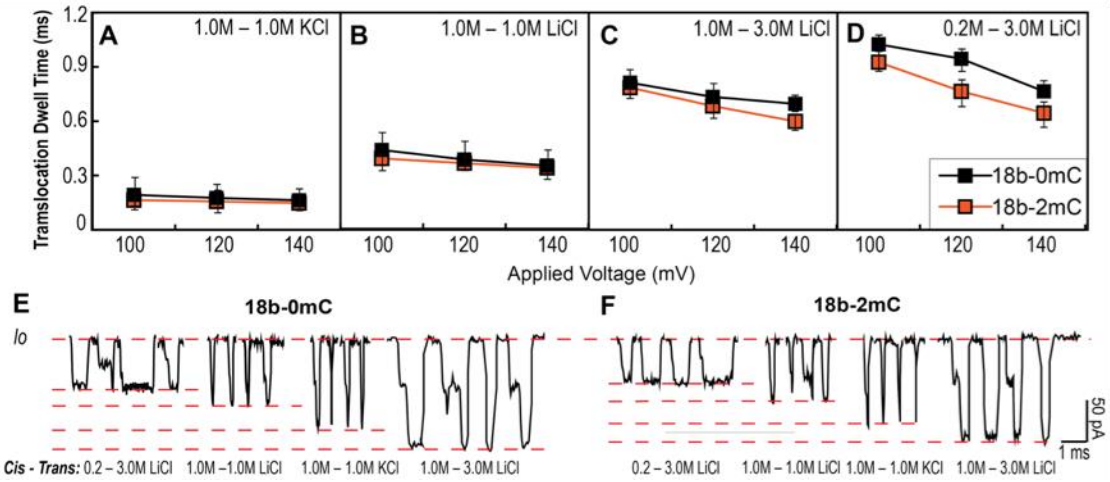


Figure 26. Comparison of cation effects on DNA translocation time. (A – D) 18b-0mC and 18b-2mC DNA translocation time, Δt, in: (A) 1.0M – 1.0M KCl (control), (B) 1.0M – 1.0M LiCl, (C) 1.0M – 3.0M LiCl, (D) 0.2M – 3.0M LiCl (optimal condition). The recorded Δt were plotted as a function of applied voltage at 100mV, 120mV, and 140mV. All points are the value of the fit with standard error. Each data is overlaid with over n = 2000 separated translocation events recorded per data point. The 18b-2mC DNA, with an increased in hydrophobicity due to the addition of methyl groups, can translocate through the nanopore faster than 18b-0mC DNA does at any given experimental condition. At 0.2M – 3.0M LiCl (cis/trans), event translocation times were prolonged approximated by 6-fold for both DNA samples, when compared to 1.0M KCl (control). (E-F) Characteristic current blockade signatures representing the 18b-0mC DNA (left) and 18b-2mC DNA (right) through the α-hemolysin nanopore are shown. 18b-2mC DNA generates shallower current blockages than 18b-0mC does for all applied voltages. Overall, the current blockage amplitude, ΔI, is a dependence of cation and its concentration in the experimental solution.

Previous studies showed that binding affinities of cations to DNA are inversely proportional to the radius of the specific binding ion. Thus, binding affinity of cations to DNA increases with decreasing ion radius, explaining the increased bond strength of binding Li^+ as opposed to K^+. The measured radii of Li^+, Na^+ and K^+ has been reported as follows: 0.69, 1.02 and 1.38 Å, respectively.[276, 281] Lithium shows the strongest binding affinity, as well as the longest binding time, resulting in increased translocation times of ssDNA in solid-state,[271] and in biological nanopores, as shown here.

However, the physical-chemistry properties of Li^+ in fact are a greater contributor in decreasing the mobility of DNA. Studies show that the effects of these binding counter-ions are due to changes in viscosity of the electrolyte solution, and perturbed hydrogen, leading to a decrease in the mobility of both ss- and dsDNA.[282] Specifically, lithium is considered a water making ion, since it has a greater number of hydrogen bonds in solution compared to neat water (ΔN_{HB}). In lithium, the ΔN_{HB} is 0.28, whereas potassium, a water breaking ion, has a ΔN_{HB} of -0.58.[272] A positive ΔN_{HB} indicates an increase in the friction between water molecules due to its effect on hydrogen bonding.[272] The rotational correlation times of water molecules may indicate DNA mobility since they are related to the number of hydrogen bonds between the water molecules.[279] Overall, the viscosity coefficient of lithium is higher than the viscosity coefficient of potassium, further indicating that DNA is less mobile in lithium ion solution than potassium ion solution.[272] Higher viscosity leads to a decrease in DNA mobility; and therefore, its increased translocation time.[283]

6.4.2 High ionic concentration decreases DNA mobility. Increase in solute bulk conductivity at high ionic concentration can prolong the translocation time of DNA.[284] An approximate 2-fold increase in Δt is observed as the ionic concentration switches from symmetric 1.0M to 3.0M of KCl or LiCl (Table 11 and Table 12). This can be explained by the interionic effects at high concentration, which as a result, increases the resistivity of the ionic solution.[284] Interionic effects occur when ions are submerged in an ionic space with a net charge opposite that of the ion's.[284-286] Thus, this ionic solution diminishes the mobility of the ions within through a drag force. This interionic effect is related to the concentration of the ionic solution in that a weak electrolyte solution promotes weak effects whereas a strong electrolyte solution promotes strong effects. Furthermore, in highly concentrated solute (e.g. 3.0M LiCl), DNA molecules are saturated with counter-ions that can decrease effective charges on the DNA and therefore, prolong the translocation time. Notably, lowering the ionic concentration in the *cis* chamber, while keeping the *trans* chamber with the same concentration (1.0M – 3.0M LiCl and 0.2M– 3.0M LiCl) can further increase the translocation time of ssDNA (Table 13). This phenomena were explained in a previous study.[287] Briefly, longer translocation time of DNA under asymmetric salt concentration is an effect of increasing electroosmotic flow of counter-ions along the DNA. With the salt concentration lowers in the *cis* than in the *trans* chamber, cations move from *trans* to *cis*, down the direction of both electrical and chemical potential gradients. Thus, electroosmotic flow of cation is significantly increased.[287] This further provides a drag force opposed to DNA motion in the pore, hence, hindering the translocation speed of DNA. Through lowering the ionic concentration of electrolyte

solution in the *cis* side, while keeping *trans* side constant, we achieved the highest electroosmotic flow of cation (trans to cis) at 0.2M – 3.0M LiCl.

Table 11
Summary of event analysis for 18b-0mC sample in KCl solutions

	Amplitude (pA)		**Dwell time (ms)**		**Frequency (s^{-1})**	
	Mu	Sigma	Tau	S.E.	Tau	S.E
1M-3M KCl	-179.05	3.82	0.32	0.04	10.08	0.38
1M-1M KCl	-101.79	6.67	0.17	0.003	6.72	0.43
3M-3M KCl	-282.99	4.84	0.47	0.015	1.30	0.06
3M-1M KCl	-169.38	4.56	0.13	0.011	0.98	0.08
3M-0.2M KCl	-109.72	2.34	0.14	0.067	0.24	0.04

Note: All data were collected at 120mV, pH 7.2. Experiments with asymmetric concentration are denoted in the order of cis – trans concentration.

Table 12
Summary of event analysis for 18b-0mC sample in LiCl solutions

	Conc.	**Amplitude (pA)**		**Dwell time (ms)**		**Frequency (s^{-1})**	
Condition	Trans/ Cis	Mu	Sigma	Tau	S.E.	Tau	S.E.
0.2M-3M LiCl	15	-79.63	2.49	0.945	0.08	55.62	3.32
0.3M-3M LiCl	10	-88.35	0.47	0.88	0.05	52.03	4.70
0.5M-3M LiCl	6	-87.14	0.21	0.81	0.06	47.62	2.08
1M-3M LiCl	3	-137.4	4.39	0.67	0.02	34.53	1.72
2M-3M LiCl	1.5	-152.3	3.65	0.57	0.07	2.44	0.16
3M-3M LiCl	1	-166.4	3.39	0.52	0.02	0.68	0.09
3M-1M LiCl	0.33	-168.1	4.31	0.19	0.02	1.47	0.13
3M-0.2M LiCl	0.067	-149.2	0.61	0.18	0.01	0.94	0.13
1M-1M LiCl	1	-79.27	3.54	0.38	0.02	4.8	0.31
1M-1M KCl	1	-101.8	6.67	0.17	0.01	6.72	0.43

Note: All data were collected at 120mV, pH 7.2. Experiments with asymmetric concentration are denoted in the order of cis – trans concentration

Table 13
Effect of salt concentration on the translocation events

	18b-0mC			18b-2mC		
	I (pA)	**Δt (ms)**	***f* (s^{-1})**	**I (pA)**	**Δt (ms)**	***f* (s^{-1})**
1M-1M KCl	-101.79 ± 6.67	0.17 ± 0.01	6.72 ± 0.43	101.06 ± 4.38	0.16 ± 0.01	10.29 ± 1.08
1M-1M LiCl	79.27 ± 3.54	0.38 ± 0.02	4.80 ± 0.31	77.50 ± 5.25	0.36 ± 0.02	7.04 ± 0.77
1M-3M LiCl	137.40 ± 4.39	0.67 ± 0.02	34.54 ± 1.72	131.21 ± 4.73	0.63 ± 0.02	35.31 ± 4.35
0.2M-3M LiCl	79.63 ± 2.50	0.95 ± 0.08	55.62 ± 3.33	74.03 ± 2.56	0.76 ± 0.08	59.84 ± 1.82

(Note: All data were collected at 120mV, pH 7.2. Experiments with asymmetric concentration are denoted in the order of cis – trans concentration)

6.4.3 Effect of salt gradient on the event occurrence. Salt concentration asymmetry (0.2M – 3.0M LiCl) results in a significant increase in the event capture rate. A series of experiments were conducted by varying the ionic concentration between the *cis* and the *trans* sides revealing significant changes in the amplitudes, as well as dwell time, and occurrence of current blockage events. Both of the latter properties are inversely proportional to C_{trans}/C_{cis} (the ionic concentration ratio going from *trans* to *cis* chamber). In other words, Δt and f_s significantly increases when $C_{trans}/C_{cis} > 1$ and decreases rapidly with $C_{trans}/C_{cis} < 1$ (event data traces are shown in Figure 27 and Figure 28). Interestingly, although the event capture rate is controlled by multiple factors (e.g., ionic concentration, voltage amplitude, and cation's nature), ionic concentration gradient between the *cis* and *trans* sides exhibits the most abundant effect. Figure 29A–D shows sample traces for typical events of 18bp ss-DNA translocations in 1.0M – 1.0M KCl, 1.0M – 1.0M LiCl, 1.0M – 3.0M LiCl, and 0.2M – 3.0M LiCl salt solutions (from top to bottom). We found that for 1.0M solutions, the ratios for the experimental event occurrence (or event

occurrence, f_s) of ssDNA are KCl/LiCl ~ 1.40 and 1.46 for C-DNA and mC-DNA samples, respectively. Upon changing the salt concentration between *cis* and *trans* chambers from 1.0M – 1.0M LiCl to 1.0M –3.0M LiCl and 0.2M – 3.0M LiCl, occurrence increase significantly. As shown in Figure 29E – F, while the translocation times are found to gradually increase, the event occurrences, f_s, quickly reached maximum value as the trans/cis concentration ratio exceed 2-fold. Specifically, C-DNA sample increase approximately 12 and 82 times upon changing the solute from symmetric 1.0M – 1.0M LiCl to 0.2M – 3.0M LiCl, and from symmetric 3.0M – 3.0M LiCl to 0.2M– 3.0M LiCl, respectively (Table 13). Interesting, both of the plots obtained from Figure 29E – F have a logarithmic trendline ($R_2 \geq 0.92$), different from a linear one obtained previously in solid-state nanopore system.[271, 287]

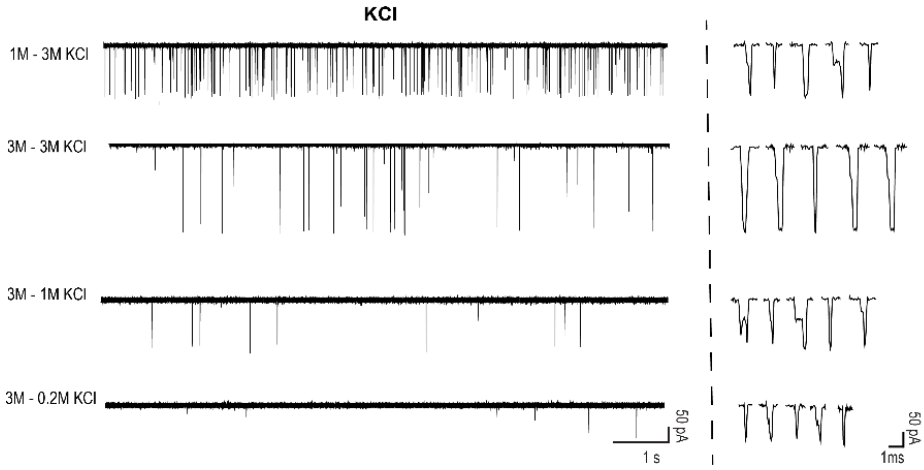


Figure 27. Sample data traces and current blockage events of 18b-0mC in KCl electrolyte buffers with varied concentrations. Different ionic concentrations between the cis- and the trans side lead to significant changes in events dwell time and occurrence. All data were collected at 120mV, pH of 7.2.

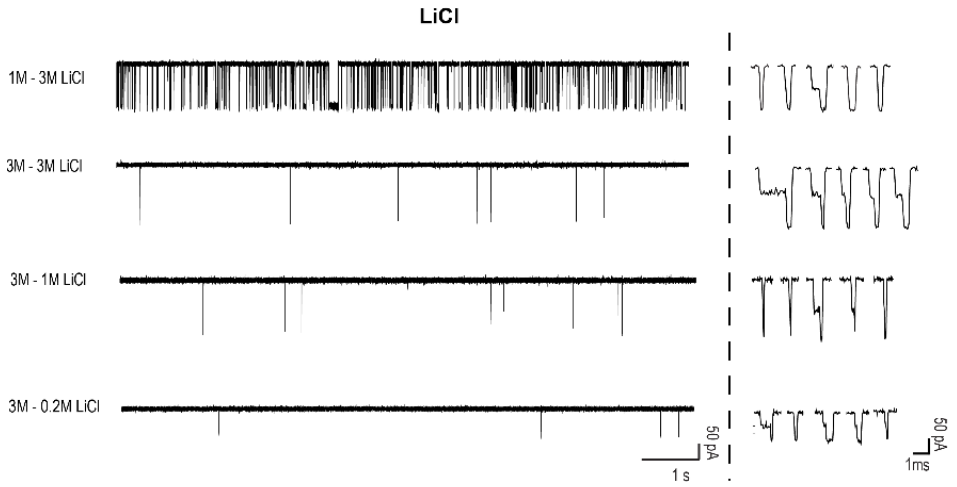


Figure 28. Sample data traces and current blockage events of 18b-0mC in LiCl electrolyte buffers with varied concentrations. Different ionic concentrations between the cis- and the trans side lead to significant changes in events dwell time and occurrence. All data were collected at 120mV, pH of 7.2.

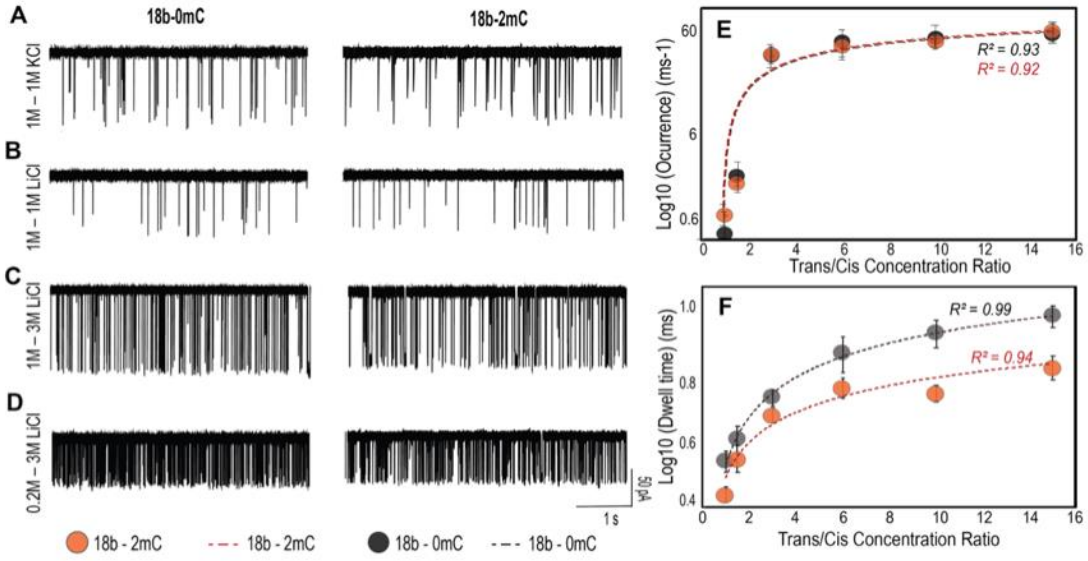


Figure 29. Effect of salt concentration on event occurrence. (A – D) Raw data traces of 18b-0mC and 18b-2mC DNA translocate through the nanopore at: (A) symmetric 1.0M KCl, (B) symmetric 1.0M LiCl, (C) asymmetric 1.0M – 3.0M LiCl, (D) asymmetric 0.2M – 3.0M LiCl. (E) Correlation between trans/cis concentration ratio and event occurrence of both samples, and (F) Correlation between trans/cis concentration ratio and dwell time of both samples. For all experiments, ionic concentration of the trans side was fixed at 3.0M LiCl, while cis concentration decreases from 3.0M to 0.2M LiCl. At any given experimental condition, the event occurrence of both samples was comparable, while the translocation time of 18b2mC were shorter than 18b0mC DNA. All data presented were collected at 120mV applied voltage.

In the presence of salt concentration asymmetry, the drastic increase in attraction of ssDNA toward the nanopore is a consequential result of two main factors: electrophoretic effects and diffusioosmotic flow.[288] Specifically, as the constant current of ions flowing through the nanopore, it creates a long range electric field funneling ions and DNA toward the pore,[287-291] increasing the DNA electrophoretic motion.[288] Moreover, because of the asymmetric salt concentration between the *cis* and the *trans* sides, there is a constant diffusion of ions between the two chambers, creating an additional electric field. For concentrations lower on the *cis* than the trans side of the pore, the created electric field is larger than if the concentrations were reversed.[292] The stronger the electric field, the better the pore is able to pull cations through the membrane.[292] Since the asymmetric salt concentration is increasing the electric field within the pore, it is also increasing the effective voltage in the *cis* side without reducing translocation times.[287] This effect is mostly due to the conductance of the pore which is dependent on both the polarity and magnitude of the voltage applied.[281] Thus, using asymmetrical salt concentration can greatly enhanced capture rate, proven in previous studies by successful detection of extremely low concentration samples (at pico- and nanomolar range).[293-294]

6.4.4 Label-free detection of cytosine methylation. Following the analysis of Li^+ and salt concentration asymmetry effect on ssDNA, we further investigate the ability of the α-hemolysin nanopore to detect and discriminate methylated cytosine in both pure and mixed samples with unmethylated cytosine. Using various methods, previous works have shown that both solid-state and biological nanopores can be used to distinguish 5'-mC from C.[56-57, 135, 283, 295-296] Because of the principle of electrical current spectroscopy, the α-

hemolysin nanopore can discern subtle structural changes on a single molecule through electrical current signatures.

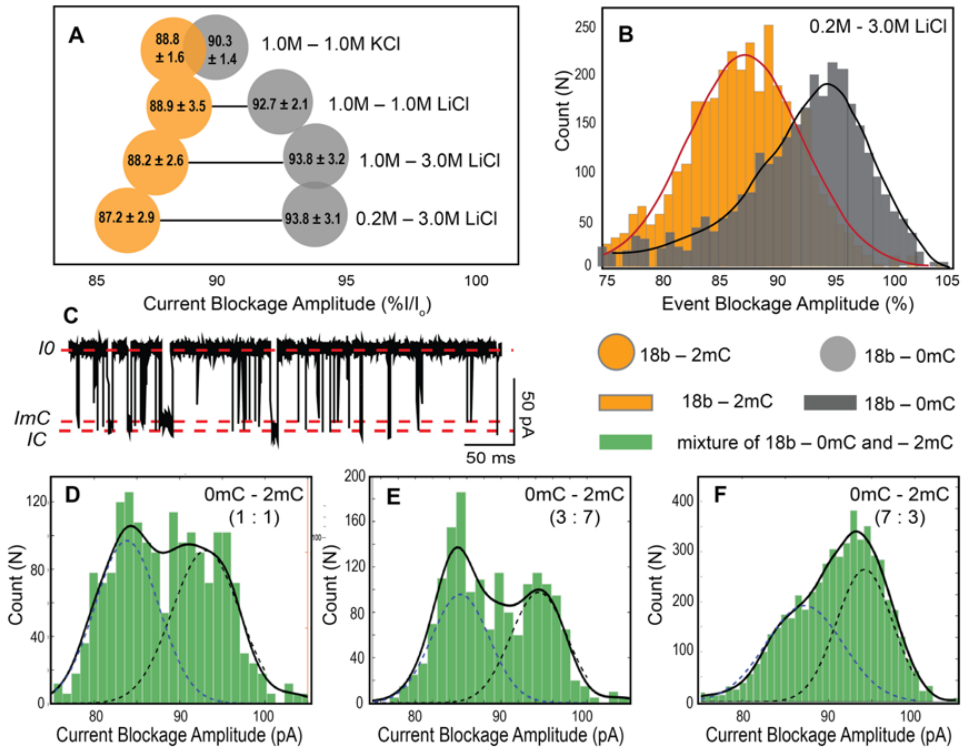


Figure 30. Detect and distinguishing methylated cytosine. (A) Peak values of gaussian fitting for current blockage, Δ*I* (pA), of pure samples containing only 18b-0mC DNA (orange) or 18b-2mC (yellow) at: 1.0M – 1.0M KCl (control), 1.0M – 1.0M LiCl, 1.0M – 3.0M LiCl, and 0.2M– 3.0M LiCl (optimal condition). Peak-to-peak distance increases as a function of the difference in salt concentrations between the cis and trans chambers, with the highest peak discrepancy achieved at 0.2M – 3.0M LiCl (cis/trans). (B) Gaussian-fittings for the distribution of current blockage amplitude for both samples at 0.2M – 3.0M LiCl. (C) Raw data trace of the 1C – 1MC mixture contains current blockage events with two different levels, denoted as ImC for mC-DNA events and IC for C-DNA events, respectively. (D-F) Distribution of current blockages of 18b-0mC and 18b-2mC mixtures of different ratios: (D) 1:1, (E) 3:7, and (F) 7:3. The distribution of event blockages exhibits bimodal shapes for all mixtures. All data were collected at 120mV applied voltage.

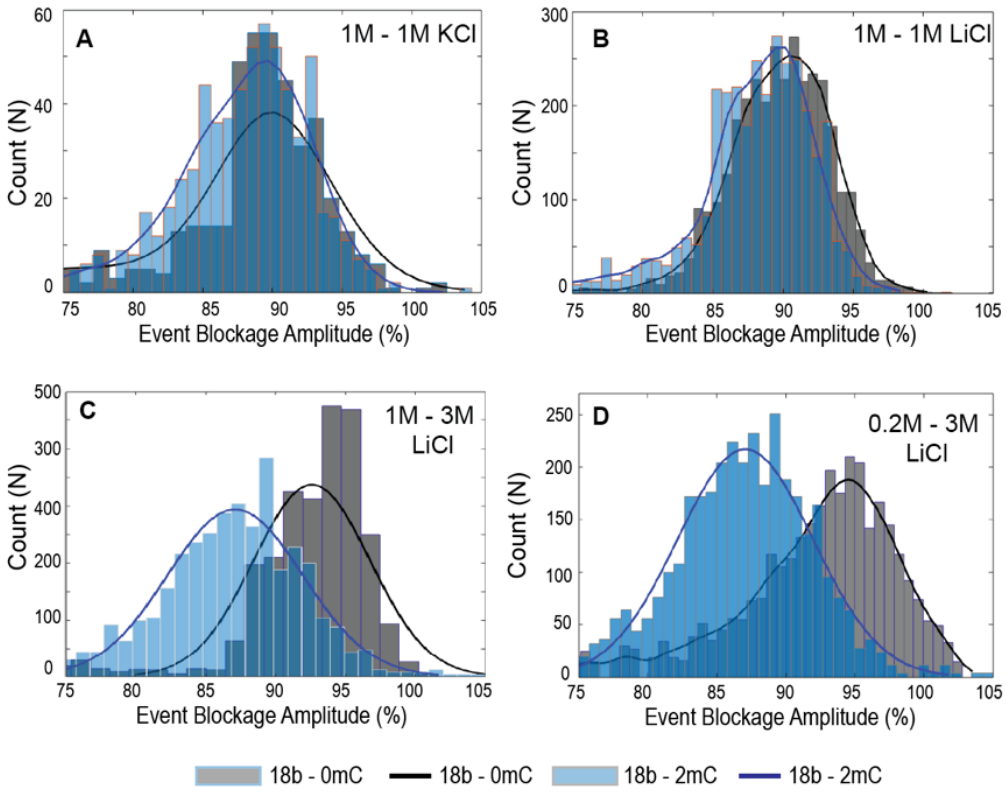


Figure 31. Gaussian fittings of 18b-0mC and 18b-2mC samples' event blockages at: 1M-1M KCl, 1M-1M LiCl, 1M-3M LiCl, and 0.2M-3M LiCl. As the concentration gradient between cis- and trans chamber increases, the peak-to-peak distance between two samples also increases. All data were collected at 120mV at pH 7.2 solutions.

The two samples employed in this study, C- and mC-DNAs, have an identical sequence, but differ in 5'-mC content. Gaussian-fittings of their current blockade information were then overlaid, revealing the changes in current blockages with regards to the electrolyte (see Figure 30A). Under symmetrical salt concentration of 1.0M KCl on both *cis* and *trans* sides, 5'-mC and C show the least discrepancy between one another (Figure 30A and Figure 31). Whereas in LiCl electrolyte solutions, the two Gaussian peaks are clearly separated from one another, with a peak-to-peak distance increase in direct relationship with the asymmetric concentration difference of LiCl between the *cis* and *trans* sides (Figure 31). The difference in peak position is most pronounced for experiments

conducted at 0.2M – 3.0M LiCl (Figure 30A – B). This finding agrees with our hypothesis that reduced translocation speeds of DNA through the nanopore can increase the resolution of readout results, allowing 5'-mC to be distinguished from C. Notably, C-DNA causes a slightly deeper current blockade than mC-DNA does, at any experimental conditions.

To further examine the ability of α-hemolysin nanopore in distinguishing 5'-mC from C within the same solution, we performed experiments on the mixtures of C- and mC-DNA samples at different ratios (C : mC): 1:1, 3:7, and 7:3 (see Figure 30D – F). As discussed in previous sections of this study, the 0.2M – 3.0M LiCl experimental condition has the most pronouncing effects on the event occurrence and dwell time of ssDNA; thus, was chosen for this second set of experiments. A sample current trace of the C-DNA : mC-DNA mixture (1:1 ratio), with two level event blockades, is depicted in Figure 30C. Herein, we developed a MatLab code employing a Gaussian-fitting for a mixture with n components (n = 2). Histograms of current blockades of all mixtures adopted a bimodal shape, contributing to the presence of 5'-mC and C. The predicted Gaussian-fittings yield peak-to-peak ratios that are coarsely corresponded to the mixture composition, specifically, with the density and distribution of events shifted in correspondence with changes in 18b-0mC and 18-2mC concentration ratios in the mixture (Figure 30E – F).

Different methods to detect and quantify 5'-mC on both ss- and dsDNA using nanopores have been demonstrated in previous studies.[56-57, 135, 295-296] In theory, the current blockade amplitude is directly corresponded to the size of the molecule translocating through the pore. Although the addition of a methyl group slightly increases the size of the 5'-mC base, whether 5'-mC DNA cause a smaller or larger current blockage than C DNA is dependent on the near sequence of the DNA strand. The magnitude of the current

blockade produced by mC, compared to C, translocating through the α-hemolysin nanopore is found to be varied between ours and previous studies,[295-296] this inconsistency in current blockage amplitude indicates that the size of a methylated DNA strand is sequence-dependent. It is well-known that cytosine methylation can alter the local DNA structure.[297-299] Cytosine methylation has influence on DNA curvature and can induce local distortion in the structure of DNA.[297-298] Moreover, the presence of methyl groups can exert an effect on A-tracts position up to three base pairs away from the site of methylation.[299] Furthermore, the presence of hydrophobic methyl group can subtly widen the major groove, and in turn, narrow the minor grove, leading to steric hindrance.[300]

Moreover, presence of a cytosine methylation can increase the angle of DNA roll by 6º or decrease its propeller twist by 5˚, depending on the local sequence.[301] Thus, we suggest that in order to achieve successful label-free detection and profiling methylated cytosine using α-hemolysin nanopore, the interaction between epigenetic modification and neighboring nucleotides should be carefully considered. Recently, the Rohs group developed a high-throughput method, named methyl-DNAshape,[301] which could be used to predict the effect of cytosine methylation on DNA. This should be a promising tool to couple with nanopore assay for cytosine methylation detection.

6.5 Conclusions

In this study, we have investigated the effect of salt gradients and the nature of monovalent cations (Li^+ and K^+) on ssDNA translocation dwell time through the α-HL nanopore. Firstly, the attraction of cations to DNA is shown to be an important factor in affecting translocation speeds. Cations reduce the overall charge of DNA molecules and the interactions between DNA and counter-ions cause significant changes in the physical

properties of DNA.[292] The counter-ions binding to DNA effectively lower the overall charge on DNA, lessening the electrophoretic drive of the system, and lowering translocation speeds. Ion binding to DNA is independent of the ion type, however, bond strengths are not equal; Li^+ bonds last longer on average in comparison to Na^+, which last longer than K^+ bonds. The difference in bond strengths lead to different net forces acting on the DNA, and thus velocity differs depending on the type of electrolyte solution.[302] Notably, the stronger electroosmotic flow in the presence of LiCl than KCl can further aid in the prolonged translocation time of ssDNA through the α-hemolysin nanopore.

Secondly, previous experiments show that DNA is saturated with counter-ions, thus, increasing the salt gradient will decrease the charge on the DNA.[269] Additionally, the continuum flow theory based on the Navier-Stokes equations assumes the state variables (such as density) do not significantly change over intermolecular distances. However, observations showed fluctuations in fluid density close to the surface in molecular dynamics simulations and experiments, hence, shear viscosity near the nanopore wall dramatically increases, leading to slower translocation times.[284] Salt gradients can create a large electric field that will funnel ions and polymers towards the pore,[309] increasing the capture rate and mean translocation time of DNA.[288] For concentrations lower on the *cis* than the *trans* side of the pore, the created electric field is larger than if the concentrations were reversed.[287] The stronger the electric field, the better the pore is able to pull cations through.[292]

Through a series of comparative studies between C-DNA and mC-DNA, we found that in a pure sample (containing only either 18b0mC or 18b2mC) with any of the salt gradient conditions listed in this study: (1) mC-DNA has a shorter translocation time, and

(2) a smaller current blockage amplitude than C-DNA does. In the presence of cytosine methylation, the local structure of DNA strands is altered in a sequence-specific manner, thus, increasing or decreasing the current blockage amplitude of methylated DNA comparing to an unmethylated strand. Interestingly, in mixtures containing both cytosine and methylated cytosine, mC-DNA exhibits a lower pore selectivity than C-DNA does. Using the LiCl salt concentration asymmetry, we were able to detect and distinguish mC-DNA from C-DNA without the use of labels, or chemical modifications.

Chapter 7

Enhancing Cytosine Methylation Detection with MBD2 Protein

7.1 Abstract

In this chapter, we explored the feasibility of enhancing the nanopore signal for methylated cytosine using engineered methyl-binding domain 2 (MBD2). Dissociation and translocation of double-stranded DNA through a biological nanopore causes a current blockage event with distinct levels. Our results showed that cytosine methylation increases DNA duplex stability, thus, significantly increasing the event translocation time. Moreover, our results showed that the presence of MBD2 protein significantly increased the event amplitude and strand dissociation time. Finally, we explored the feasibility of distinguishing unmethylated from methylated DNA using supervised machine learning models. Addition of MBD2 protein significantly increased the models' performance, yielding the optimal classification model with 88% sensitivity and 86% specificity.

7.2 Introduction

In Chapter 6, we investigated the possibility of label-free detection of DNA methylation using the α-hemolysin nanopore. With the aid of an ionic gradient between the *cis* and the *trans* chambers, we were able to distinguish between ssDNA containing methylated and unmethylated cytosine in mixtures with various composition. In theory, the current blockade amplitude directly corresponds to the size of the molecule translocating through the pore. However, label-free detection of 5-methyl-cytosine (5mC) using the α-hemolysin nanopore remains challenging because current blockade amplitude of 5mC is dependent on its interaction with the neighboring nucleotides, which can induce possible local conformational changes.

Methyl-CpG-binding proteins (MBPs) are proteins that are able to identify methylated cytosine paired with guanine dinucleotides (CpGs), to recruit co-repressor molecules and modify the surrounding chromatin to block transcription.[303] MBPs play an important role in mediating the silencing of gene expression, and are of particular interest because they serve as great identifiers for DNA methylation.

In this chapter, we explore the feasibility of using MBP, specifically methyl-binding domain 2 (MBD2 protein) as a tag for detection of methylated DNA. Specifically, we designed synthetic dsDNA fragments containing the *GREM_1* promoter region sequence, with single-stranded overhangs. MBD2 and methylated DNA conjugate to form MBD2-DNA complexes, which can amplify the signals of current-blockages caused by methylated cytosine on the dsDNA.[304-306] Additionally, the single-stranded overhangs can translocate through the constriction part of the pore and reside in the long β-barrel stem, allowing DNA duplex to unzip with a voltage pulling force. Using ionic spectroscopy, we showed that binding of MBD2 significantly enhanced the current amplitude difference in methylated and unmethylated DNA fragments. We also evaluated the effect of adding MBD2 protein on the overall performance of five supervised machine learning algorithms, to classify unmethylated from methylated DNA.

7.3 Methods

7.3.1 DNA samples. Synthetic double-stranded DNA with single-stranded overhangs were designed and purchased from Integrated DNA Technology (IDT – Coralville, IA). DNA oligonucleotides utilized in this study had the following sequences (underlined = single-stranded region, **bolded** = *GREM_1* promoter region, **M** = methylated cytosine):

u100 DNA (unmethylated DNA):

Forward strand: 5' – CTT CTG ACT GGG GAG AGG CGC GGC CCC GG**G CCG CGC TCT GCC GCCGC ATT TAA ACG GGA GAC GGC GCG ATG CCT GGC ACT CGG TGC GC**C TAT CTA TGT G– 3'

Reverse strand: 5'- **GCG CAC CGA GTG CCA GGC ATC GCG CCG TCT CCC GTT TAA ATG CC GCG GCA GAG CGC GGC** CCG GGG CCG CGC CTC TCC -3'

m100 DNA (Methylated DNA):

Forward strand: 5' – CTT CTG ACT GGG GAG AGG CGC GGC CCC GG**G CCG CGC TCT GCC GCMGC ATT TAA ACG GGA GAC GGC GCG ATG CCT GGC ACT CGG TGC GC**C TAT CTA TGT G– 3'

Reverse strand: 5'- **GCG CAC CGA GTG CCA GGC ATC GCG CCG TCT CCC GTT TAA ATG CM GCG GCA GAG CGC GGC** CCG GGG CCG CGC CTC TCC -3'

m100•MBD2: methylated DNA (m100) sequence conjugated with MBD2 protein.

Forward strand: 5' – CTT CTG ACT GGG GAG AGG CGC GGC CCC GG**G CCG CGC TCT GCC GCMGC ATT TAA ACG GGA GAC GGC GCG ATG CCT GGC ACT CGG TGC GC**C TAT CTA TGT G– 3'

Reverse strand: 5'- **GCG CAC CGA GTG CCA GGC ATC GCG CCG TCT CCC GTT TAA ATG CM GCG GCA GAG CGC GGC** CCG GGG CCG CGC CTC TCC -3'

Upon delivery, all DNAs were suspended in a standard DNA storage buffer containing 10mM Tris, 1mM EDTA and titrated to 8.0pH. All samples were then aliquoted

into small volumes and stored at -80˚C until usage, in order to prevent multiple thaw-freeze cycles that can degrade and affect DNA quality. For both nanopore and circular dichroism (CD) experiments, a final DNA concentration of 100 nM was used.

7.3.2 Methyl-binding domain 2 (MBD2) protein. Recombinant binding-domain of MBD2 protein was kindly provided by LumiMac Inc. (South Korea). The purified MBD2 protein monomer (MW = 10.7kDa) was stored at 1.3mg/ml of buffer containing 25mM Tris (pH7.5), 500mM NaCl, 0.1% Tween 20, 0.1mM EDTA, 0.5 mM DTT and 50% glycerol. The MBD2 protein samples were shipped in dry ice and immediately stored at -20˚C upon arrival. Prior to experimentation, MDB2 protein was reactivated in a solution containing 10mM Tris (pH7.2), 1mM MgCl2, 1mM DTT, 5% glycerol for 15 minutes at room temperature. This step was critical for assuring optimal MBD2 binding activity. Then, the reactivated MBD2 protein was incubated with methylated DNA at a 2:1 concentration ratio for 15 minutes at room temperature, before being added to the experimental buffer.

7.3.3 Single-channel recording. Single-channel recordings were performed using Axopatch 200B (Molecular Device Inc.), filtered with a built-in 4-pole low-pass Bessel Filter at 5kHz. Data were acquired at 180mV with Clampex 10.7 (Molecular Device Inc.) and Axon Digidata 1550B A/D converter (Molecular Device Inc.), at a sampling rate of 20 kHz. We used Clampfit 10.7 (Molecular Device Inc.), Excel (Microsoft) and R software for data analysis and processing.

7.3.4 Data analysis and visualization. Events with translocation time < 2ms or %I/Io < 50% were considered as either translocation of single-stranded DNA or brief interaction of DNA and the pore entrance; thus, they were disregarded from data

analysis.[237] We used ggplot2 package (R) to generate boxplots for the overall event dwell time and histogram of current blockage amplitude.

7.4 Results and Discussion

In this study, we employed three DNA samples with identical sequences but differences in methylation status: (1) unmethylated DNA (u100), (2) methylated DNA (m100), and (3) MBD2 bound methylated DNA (m100•MBD2). All DNA samples contain a double-stranded section (78 bps) and two overhang single-stranded ends (11 bps each). The specific length of the overhangs was designed to ensure that a sufficient number of nucleotides resided within the pore β-barrel at all times, to initiate the unzipping of DNA duplex upon electrical pulling force.[307-311]

7.4.1 Mechanism of DNA unzipping through the nanopore. As shown in Figure 32A, when dsDNA unzips in a nanopore it undergoes two distinct steps. The first step is when the ssDNA tail is caught in the nanopore and begins to translocate,[319] while the ssDNA tail is simultaneously pulled through the bulk solution to the nanopore because of the DNA's negative charge.[307] As the ssDNA tail gets caught in the nanocavity, the dsDNA region cannot translocate through the nanopore constriction section without melting into single-stranded DNAs (e.g. strand dissociation).[60] After the dsDNA region is unzipped, the two strands are physically separated (Step 2).[60]

Translocation of MBD2-dsDNA complex through the nanopore is similar to that of unbound dsDNA, with an extra step of protein dissociation (Figure 32B). As the distance between base pairs of DNA is approximately 0.34nm, the total length of the DNA sequences would be approximately 34nm total, with single-stranded overhangs of 3.74nm each.

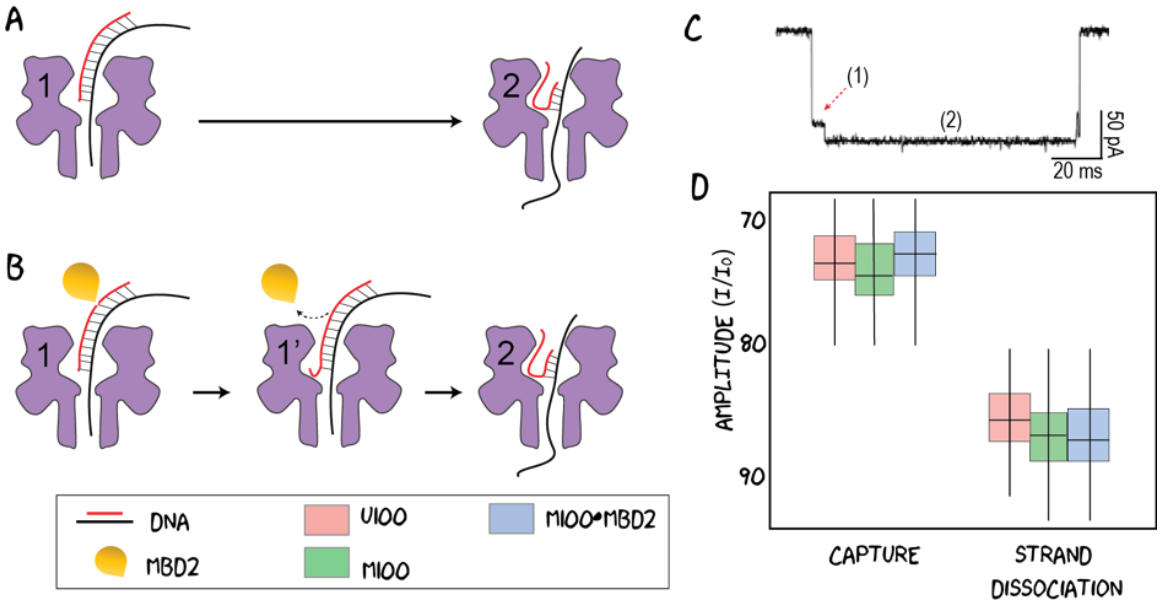


Figure 32. Mechanism of DNA translocation through the nanopore. (A) Translocation of and unbinding dsDNA occurs in two steps: (1) single-stranded tail gets captured by the pore and (2) complimentary strand dissociates under constantly pulling force of the electrophoretic field. (B) Mechanism of m100•MBD2 DNA translocation through the nanopore is similar to that of u100 and m100; however, MBD2 must detach prior to duplex full dissociation. (C) Typical step-wise function of a translocation event, reflecting the two stages: (1) DNA capture and (2) strand dissociation. (D) Box-plot of the event amplitude at each translocation step.

Since the methylated *CpG* is located at the 44th base from the 5' end, the double-stranded DNA dissociation (Step 1) must happen prior to MBD2 protein dissociation (Step 1). Specifically, as the MBD2-methDNA complex threads through the α-hemolysin nanopore, the MBD2 protein gets stuck at the pore entrance for a short period, then transiently dissociates from the DNA strand, causing a prolonged current blockage event. This initial ssDNA blockage results in a Level 1 current blockage (i.e. event capture), which characterizes the beginning of an event. When the first strand that has the initial ssDNA strand is translocated, the second strand coils in the vestibule of the nanopore.[307] This coiling results in a sharp and short upshot of the current blockage after the Level 1

blockage.[6] The current drops again as the second strand begins to translocate.[314] If the second strand exits the nanopore, then the signal will continue to return to the baseline after the initial spike (Figure 32C).[307]

Overall, the mechanism of dsDNA strand dissociation is governed by the DNA sequence (i.e. length, GC content) and by the environment that the DNA experiences (i.e. voltage, ionic strength, temperature).[309] Since the dsDNA region cannot enter the neck, interactions between the dsDNA and the nanopore contribute to the unzipping.[307] The length of the dsDNA region also affects the ease at which a strand will unzip.[312] Longer dsDNA region associates with a more stable the DNA complex.[312] This increase in stability changes the energy interactions of the energy field and the pore.[312] The force of the energy field causes the dsDNA region to experience an upward shear force.[313] As the dsDNA region unzips, the complex also experiences a traction force as the strands are pulled apart.[313]

7.4.2 Binding of MBD2 does not change event amplitude. We examined the overall average current-blockage event amplitude (%I/Io) of u100, m100, and m100•MBD2 DNAs (as shown in Figure 33). As mentioned previously, the samples employed in this study had an identical sequence, but differ in 5'-mC content. Gaussian-fitting of their current blockage information exhibited a prominent peak of %I/Io ~ 86.5 ± 1% (Figure 33), which were slightly shallower than that of single-stranded DNA translocation through the nanopore (as concluded in Chapter 6).[238]

Comparing between three samples, the mean current-blockage event amplitude (%I/Io) increased in the order of u100 < m100 < m100•MBD2 (Table 14). Specifically, the mean %I/Io were approximately: 85%, 85.6%, and 86.5% for u100, m100 and

m100•MBD2 DNAs, respectively. Moreover, we performed an unpaired t-test and %I/Io of all samples was statistically significant from each other at 99% confident interval (adjusted p << 0.01), as shown in Figure 33. The slight difference in %I/Io of u100 and m100 DNAs observed herein was resulted from interaction of methylated cytosine with adjacent bases, as previously discussed (see Chapter 6; Conclusions).[238] Addition of MBD2 protein resulted in a 0.9% increase in the event average %I/Io (from 86.5% in m100 DNA to 86.5% in m100•MBD2 DNA). This change was visible during the strand dissociation step but not during event capture (Figure 32D). In other words, after the m100•MBD2 complex was captured and started unzipping, the MBD2 protein got pulled toward the pore entrance, partially blocked the nanopore *cis* opening, resulting in a slightly deeper %I/Io.

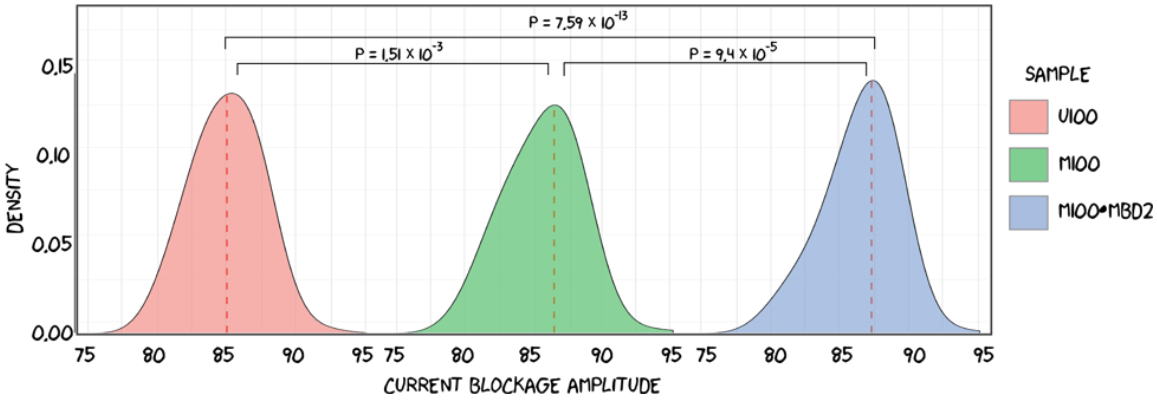


Figure 33. Changes in current blockage amplitude. Distribution of the overall current blockage amplitude (%I/Io) of three DNA samples, including: u100 (pink), m100 (green) and m100•MBD2 (blue). All samples exhibited a normal distribution curve, with a prominent peak at approximately 85 – 87%. Student's [STUDENTS'?] test showed significant change in all samples.

Table 14
Summary of current blockage event amplitude (%)

Condition	Mean (%)	Max (%)	Min (%)	Median (%)	Std.
u100	84.06	92.37	75.29	84.33	3.25
m100	84.71	94.95	75.17	85.60	3.55
m100•MBD2	85.40	98.13	75.02	86.39	3.93

While each current blockage event was a step-wise function, the duration of each amplitude level was greatly different. In the majority of the dsDNA that successfully melted and translocated through the pore, strand dissociation was the rate-determining step; thus, consequently dominates the final value of the average event amplitude. We suspect that the slight decrease in value of event current blockage of dsDNA observed herein, in comparison to that of ssDNA, most likely was due to technical limitations. Specifically, translocation of dsDNA through the nanopore occurs in two steps of capture and dissociation, leading to a step-wise function in current blockage event amplitude. As shown in Figure 32D, capture of ssDNA overhang caused a shallow blockage amplitude of 73 ± 0.5%, while the strand dissociation step was characterized by a prominent long block at 87.5 ± 1% (similar to ssDNA translocation). When detected by Clampfit software, the current blockage amplitude was automatically calculated as an average of the whole event amplitude, thus leading to a slight decrease in %I/Io of dsDNA translocation.

7.4.3 Increased strand-dissociation time in methylated DNA and MBD2-bound DNA. Event translocation time reflects the structural stability and interaction with the pore. In Figure 34, we plotted the distribution of event dwell time for u100, m100 and m100•MBD2 DNAs in the same experimental condition. Under a +180mV applied voltage and an ionic concentration of 1M – 1M NaCl (in the *cis* and *trans* chambers), the average

translocation time, Δt, increased in the order of u100 < m100 < m100•MBD2 DNA (Δt = 1.2, 1.7, and 2s, respectively (Table 15)). In other words, Δt of m100 and m100•MBD2 DNA increased approximately 1.4 and 1.6 times for m100 and m100•MBD2, respectively, in comparison with Δt of u100 DNA.

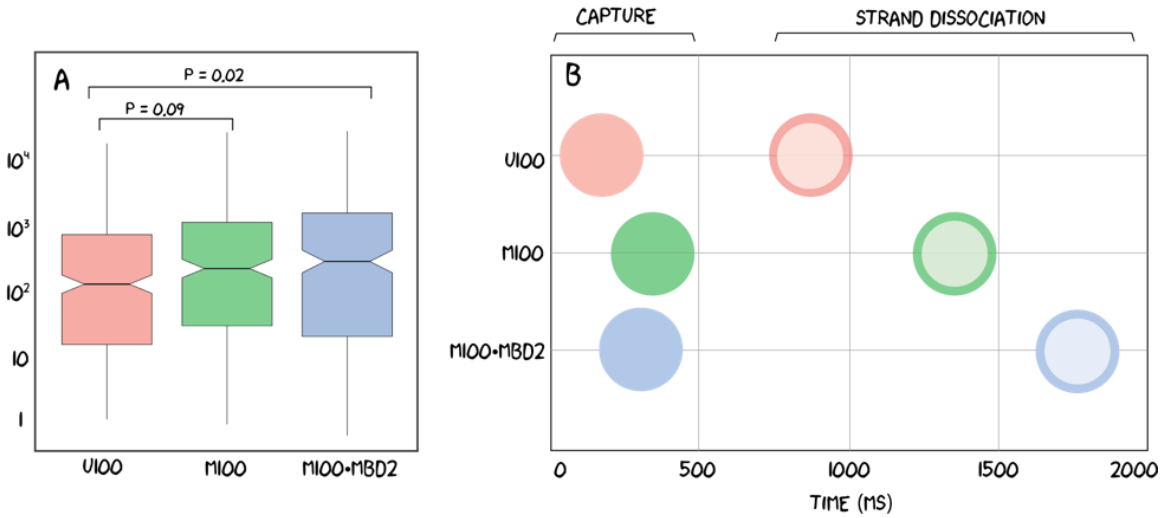


Figure 34. Changes in translocation time. (A) Boxplot of event dwell time distribution for unmethylated DNA, methylated DNA and MBD2-methylated DNA complex. Students' test determined a statistically significant change in dwell time of m100•MBD2, in comparison to u100 DNA> (B) Bubble plot of event dwell time, associated with two stages in dsDNA translocation through the nanopore: (1) capture of single-stranded overhangs and (2) dissociation of complimentary base-pairs in double-stranded section. The majority of change in dwell time exhibited in the strand dissociation time, reflecting an increase in duplex stability in m100 and m100•MBD2 DNA.

Table 15
Summary of current blockage event dwell time (ms)

Condition	Mean	Max	Min	Median	Std.
u100	1262.17	23077.65	1.25	153.55	3041.21
m100	1719.08	34113.85	1.05	267.45	3678.56
m100•MBD2	2026.84	44168.05	0.70	344.80	4198.51

Notably, Δt of identical DNA sequences from the same group of samples was greatly varied (from ms to tens of seconds), due to the random thermal activation required for separation.[256] The melting process of dsDNA occurs in two steps: the first is the initial separation to overcome the energy barrier, and the second is the unzipping of the rest of the strand.[316] The energy barrier refers to the amount of energy required to break the hydrogen bonds of the dsDNA region.[309] The energy that breaks the hydrogen bonds is fueled by the energy field within the nanopore.[321] The electrons in the energy field within the nanopore causes the dsDNA to denature and begin to unwind.[314] Sometimes, trouble arises during the transition from the initial metastable state to a final stable (lower-free-energy) state. When this happens, the dsDNA can get stuck inside the nanocavity, causing current blockage events lasting minutes or indefinitely.[315]

On the other hand, the overall increase in Δt among different samples, as observed from Figure 34A, was a consequential result of increased DNA duplex stability. As mentioned previously, each translocation event of dsDNA consists of two steps: event capture and strand dissociation. Through manual analysis of the data, we found the time of strand capture was similar among the three samples (~250 ±70ms), while strand dissociation times significantly increase in m100 and m100•MBD2 DNA (as shown in Figure 34B). As the methylated cytosine is at the 44th base of the DNA sequence, distance between MBD2 protein and the 5'- or 3'- end of the DNA was sufficiently long (~ 15nm) to prevent the binding of MBD2 from affecting the capture of ssDNA overhangs.

This explains the reason why capture time remained similar between three samples, while increases in Δt of the strand dissociation step indicated enhanced stability of dsDNA structure in m100 and m100•MBD2. Specifically, the strand dissociation time of u100,

m100 and m100•MBD2 were 0.87, 1.36 and 1.75s, respectively. In other words, compared to u100 DNA, there were 1.6 and 2 times increase in strand dissociation time in m100 and m100•MBD2, respectively. Using an unpaired t-test, we determined that there was statistically significant change in Δt of m100•MBD2 in comparison with u100 DNA (adjusted p value = 0.018). The main contributing factor to enhanced duplex stability was the presence of methylated cytosine on the 44th base. Previous studies showed that DNA templates with different content of cytosine methylation have different melting temperatures.[18] DNA methylation increases the DNA duplex stability, reflecting through both its thermal and mechanical properties. In the context of nanopore experiments, dsDNA translocating through the pore went through a force-induced strand separation, which is greatly affected by the strand mechanical stability.

In m100•MBD2 sample, both the presence of methylated cytosine and MBD2 contributed to the prolonged strand dissociation time. As mentioned earlier, in order for the m100•MBD2 complex to translocate through the pore, MBD2 protein must first dissociate from dsDNA. A previous study reported MBD2 to have a dissociation constant (K_D) of 2.1 **μ**M.[316] The structure of MBD2 consists of a long finger-like projection that is composed of three strands of **β** sheets.[316] When binding to DNA, MBD2 interacts with approximately 10-11 nucleotide base pairs.[316-317] Its finger-like projection extends into major grooves of the DNA where it contacts the CpG sequence. In particular, the region contacting the DNA is well structured and thus creates a stable connection.[316] This stable bond is created by interactions between specific amino acids that make up the MBD and bases of the CpG sequence. Specifically, Arginine (R), Tyrosine (Y), and Aspartic Acid (D) residues make specific base interactions with the CpG sequence that cause a flat

orientation.[316-318] The Arginine forms hydrogen bonds with symmetrically opposed Guanine bases. This allows its side chains to pack against neighboring methylated cytosine groups. This dense packing causes the Arginine side chains to form van der Waals contacts with the DNA, which further stabilizes the bond.[316-317] Finally, the Tyrosine and Aspartic Acid also form hydrogen bonds with the surrounding complex, which further strengthens the connection and locks MBD2 into place.[316-317] As dsDNA is pulled through the nanopore, MBD2 protein is anchored at the pore entrance until there is sufficient energy to break its connection to the DNA; thus, prolonging the event dwell time.

7.4.4 MBD2 enhance methylation detection accuracy. Having determined the changes in amplitude and stability of the three samples (u100, m100, and m100•MBD2), we tested the efficiency of using machine learning algorithms to distinguish methylated from unmethylated DNAs. Specifically, we performed comparative analysis using the same set of five different algorithms in duplicate on: (1) u100 and m100 DNA and (2) u100 and m100•MBD2 DNA. The final goal was to determine whether employing MBD2 protein enhanced the nanopore's ability to detect methylated cytosine on dsDNA.

The five algorithms we used included: K-Nearest Neighbor (KNN), Logistic Regression (LR), Linear Discrimination Analysis (LDA), Quadratic Discrimination Analysis (QDA), and Random Forest (RF). In each DNA sample group, we randomly selected 500 current-blockage events, then used 80% of the dataset for model training and the remaining 20% for validation purposes. Two features of the current-blockage event, Δt and %I/Io, were used as input for the models.

To evaluate the overall performance of the models on classifying u100 from m100 DNA and from m100•MBD2 DNA, we constructed two vertical box-and-whisker plots, as

shown in Figure 35A – B, respectively. In this plot, the box and whiskers show the median of ARUOC as a dot, and the interquartile range of each distribution as a box around the median. The x-axis represents the area under the Receiver Operating Characteristics curve (AUROC). The closer the AUROC value is to 1, the better the model at is correct classification. Generally, a better model would have higher median AUROC and smaller range between minimum and maximum AUROC. For a binary system, the threshold is AUROC is equal to 0.5 (shown as a red line) and maximum AUROC value is equal to 1.

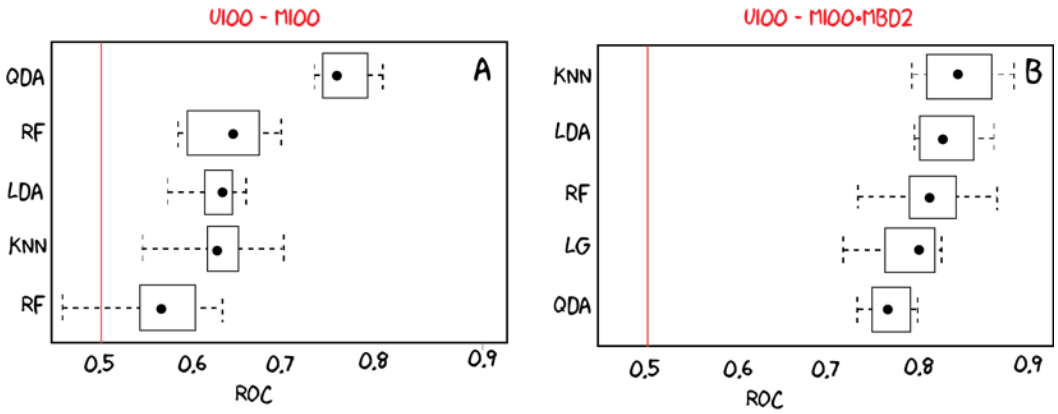


Figure 35. Machine learning classification of methylated and unmethylated DNA. Five machine learning models were employed to classify u100 from m100 or m100•MBD2. The models include: K-nearest neighbor (KNN), Random forest (RF), Logistic regression (LG), Linear discrimination analysis (LDA), and Quadratic discrimination analysis (QDA). Box-and-whisker plots represent distribution of area under the receiver operating characteristic curve (AUROC). Model's performance was ranked based on increasing median AUROC value (shown at the dot).

Classification models for u100 and m100 DNA yielded relatively poor results, with the majority of the models having a median ARUOC ≤ 0.65. The QDA classifier had the best performance with a median AUROC of 0.76. The classifier's optimal sensitivity and specificity were 78% and 54%, respectively. On the other hand, when applying the

classifying models on u100 and m100•MBD2 dataset, all classifiers returned a median AUROC value ≥ 0.75 (comparing to 0.65 for the u100 – m100 dataset).

KNN model had the highest median AUROC of 0.83, with the optimal sensitivity and specificity of 88% and 86%, respectively. It is worth noting that according to the decision tree (Figure 36 and Figure 37), the current-blockage event amplitude outweighed dwell time and played the most important role in classification of u100 from m100 and m100•MBD2 DNA.

Altogether, our study showed that the addition of MBD2 protein significantly improved the probability of correctly distinguishing unmethylated DNA from methylated DNA. A limitation in our prediction model comes from having nonhomogeneous control samples, specifically m100•MBD2 population. When working with dynamic experimental conditions, such as those investigated here, the process of MBD2 protein binding/unbinding to dsDNA happens constantly; thus, there is always a mixture of both bound and unbound dsDNA in the m100•MBD2 sample. This creates bias in the training set for the machine learning models and hindered the model performance.

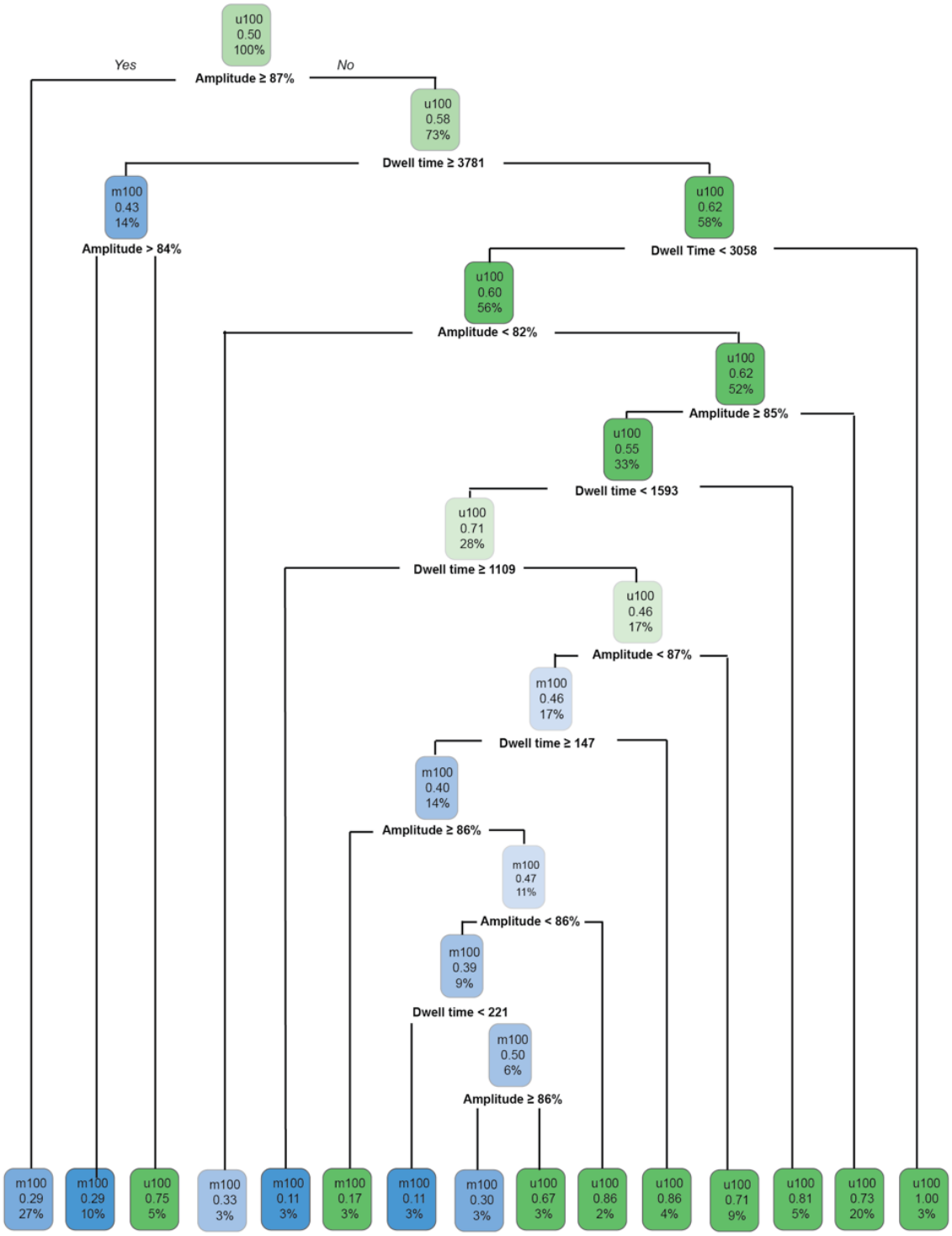


Figure 36. Decision tree for classification of u100 DNA from m100 DNA.

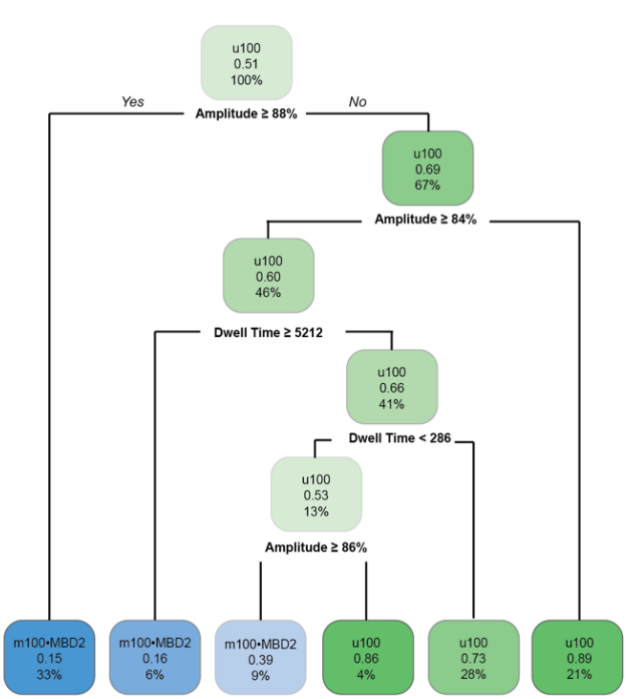


Figure 37. Decision tree for classification of u100 DNA from m100•MBD2 DNA.

7.5 Conclusions

In this work, we showed that the α-hemolysin nanopore can be used as a biosensor for detecting the methylation status of dsDNA. Since a level amount of methylation directly affects the DNA's stability and its corresponding translocation time, it is possible to build a panel of dsDNA translocation times in correlation with the sequence length, GC content, and methylation level. Previous studies have shown that methylation of a certain level (i.e. 40% CpG is methylated in *CDKN2B* gene) can result in complete gene silencing, regardless of the CpG methylation pattern.[22] While it will remain challenging to understand the exact number of methylation sites on a random DNA fragment, it is feasible to determine whether the sequence is hypermethylated or under methylated (compared to a threshold methylation percentage), using nanopore aided with machine learning prediction algorithms.

Chapter 8

Gone Fragments: Capturing Circulating-Tumor DNA

8.1 Abstract

Compared to traditional tissue biopsy methods, the use of plasma or urine samples (the so-called "liquid biopsy") is less invasive, improves the patient's quality of life, while offering a more reliable and frequent detection of cancer signal over time with longitudinal sampling. Unfortunately, current methods to analyze liquid biopsy samples are time-consuming, expensive and biased towards a specific fraction of the genome or mutations. In this chapter, we describe two novel sample-preparation strategies that enable sequencing of long cfDNA fragments, which are often missed by the current Next-generation sequencing (NGS) platforms (i.e. Short-read Illumina sequencers). These techniques, which employed rolling-circle amplification (RCA) or blunt-end ligation, transform cfDNA fragments into long strands that are easily captured by the nanopore sequencer. The proposed techniques can reduce the time required to process and analyze liquid biopsy samples (from weeks currently to less than 2 days). This study was carried out under the supervision of Dr. Florent Mouliere (Amsterdam UMC) and in collaboration with the group from Dr. Jeroen de Ridder (Utrecht UMC).

8.2 Introduction

Liquid biopsy is the study of tumor material collected in a minimally invasive manner through sampling of blood, urine and other bodily fluids.[319] Presence of cell-free DNA (cfDNA) in human blood was first described by Mandel and Metais in 1948. Abnormally high concentration of cfDNA in cancer patients was first observed in 1977.[320] In 1989, Stroun et al. reported that a fraction of cfDNA from the plasma of cancer patients

originated from tumor cells, giving rise to the term 'circulating tumor DNA' (ctDNA).[321-322]

Among the potential signals of cancer in the bloodstream,[323-325] ctDNA has strongly emerged as a major tool and biomarker for precision medicine in the past decade.[326-328] ctDNA is tumor-derived, fragmented DNA found in bodily fluids. Though the existence of ctDNA is now widely accepted, the mechanisms by which ctDNA enter the bloodstream are still unclear.[329] There are two main proposed mechanisms for cfDNA release: (1) apoptosis/necrosis, or (2) active secretion.[330-332] CfDNA, which was first found in the circulation, has been detected in various bodily fluids, including urine, saliva, pleural fluid, and cerebrospinal fluid.[333-338] Epigenetic and methylation analysis revealed that in healthy individuals and cancer patients, the majority of cfDNA originates from hematopoietic cells.[339-340] These cells have a modal size of approximately 166bps, which corresponds to the length of DNA wrapped around a nucleosome (~147bp) plus linker DNA.[319] Demonstration by PCR and sequencing revealed that ctDNA molecules are normally shorter than non-mutant cfDNA, with an average size of 134bp – 144bp.[341] However, the cause of shortening in ctDNA fragment is unclear.[319]

As ctDNA exhibits both the genetic and epigenetic alterations from their cancer cell origin,[339, 342-343] and can reflect tumor type and stage, it has therefore been hypothesized as a prognostic and predictive marker in cancer treatment selection[326, 344] and detection.[345] Specifically, various types of genomic and epigenomic DNA alterations have been observed in cfDNA, including for example mutation, DNA hypermethylation, copy number variation, microsatellites instability, rearrangements and losses of heterozygosity.[346] It is more beneficial to detect cancer earlier when the disease is easier to

treat. However, the potential use of ctDNA for detecting early-stage cancer still presents great challenges. For example, ctDNA amount in total cfDNA is usually very low; thus, demanding techniques with sensitivity for detecting low tumor fraction and high technical and biological background signal. In healthy individuals, cfDNA concentration tends to range between 1 and 10 ng per ml in plasma (Figure 38).[347-348] In cancer patients, the concentration of cfDNA in plasma or serum can increase over 10-fold (compared to healthy individuals), with an average of 180 ng per ml.[349] The exact amount of ctDNA released into the blood stream varies depending on the stage and type of cancer.[350-351] While the concentration of ctDNA content can be much higher in late-stage cancer patients,[350, 352-353] they usually make up less than 10% of total cfDNA.[354-356] Thus, it is still challenging to detect and diagnose tumor-related cancers with a good specificity and sensitivity, reducing the clinical implementation of such methods.

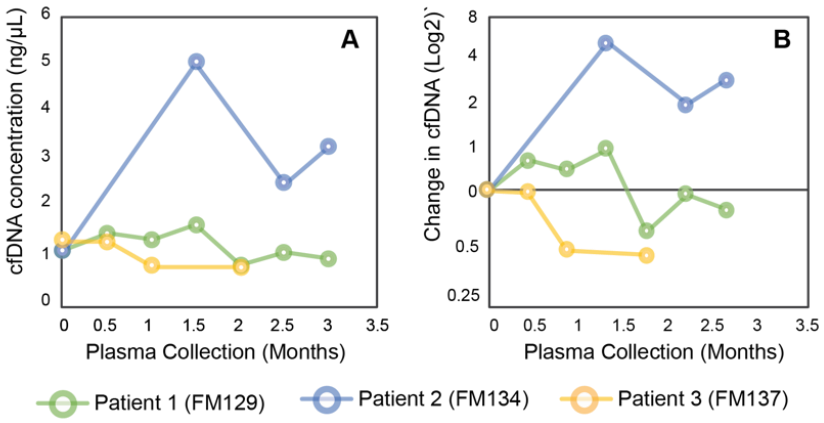


Figure 38. Changes in cfDNA concentration in plasma collected from three non-small cell lung cancer patients. Samples were collected multiple times for up to three months (longitudinal sampling). (A) Concentration of cfDNA ranges between 0 – 10ng/µL (Qubit). (B) Changes in cfDNA overtime and treatment, as normalized by the starting cfDNA concentration at the first collection time (t = 0). A high cfDNA concentration could be an indication for high ctDNA in the sample, though more tests are usually followed to confirm the exact concentration of ctDNA.

Currently, monitoring disease burden is performed using imaging or molecular methods, which are associated with radiation exposure or logistical burden.[319, 357] Conventional methods for sampling solid tumor, such as needle biopsies, are subjected to high procedure complication rate (1 out of 6 cases in some cancer types),[358] and sampling biases that arise from tumor genetic heterogeneity.[359-361] Compared to the traditional surgical biopsy, liquid biopsy is less invasive, improving the patient's quality of life, while offering a more reliable detection of tumors and mutations over time with longitudinal sampling. Since the half-life of cfDNA is between 16 minutes to 2.5 hours in circulation,[362-364] analysis of ctDNA can be considered as 'real-time' dynamic snapshot of disease burden.[319]

Currently, the majority of liquid biopsy samples are analyzed with PCR methods or Next Generation Sequencing systems (e.g. Illumina, Roche, or Ion Torrent), whose optimal DNA read length range between 75-500bp (i.e. short-read sequencing). As these systems sequence through enzymatic reaction, their efficiency significantly reduces when the DNA fragment is longer than 500bp.[365] While Next-generation sequencing can be highly accurate, this platform may not practical in clinic, due to extensive processing time and the restriction imposed by the gene panel used. Another common method for cfDNA analysis is digital droplet PCR, which has been proven to be a fast and efficient tool for studying a single locus (a fixed position on a chromosome). However, when the targeted gene does not have a hotspot (i.e. *TP53*), it is challenging to design and employ the correct set of probes, even more if the location of point mutation is unknown.

However, there is a population of cfDNA fragments that is currently missing from the picture: long cfDNA fragments (Figure 39). Early works using electrophoresis on low-

percentage agarose gel showed a variation in size of cfDNA, ranging between 0.18 – 21 kbs.[349, 366] Recently, long cfDNA fragments (>1000bp) have also been identified in healthy individuals using long-read sequencing techniques.[367] These fragments might be released to the circulation in association with exosomes,[368-369] or by tumor cells via necrosis.[370] A portion of long cfDNA fragment could come from fusion genes or oncogene-carrying extrachromosomal DNA (ecDNA, 1-3Mbs long),[371] which is a potent mechanism leading up very high copy number of oncogenes.[372] Unfortunately, these long cfDNA fragments are usually poorly recovered due to bias in DNA extraction methods,[373-374] sequencing library preparation,[375] and limitation in the enzymatic reaction of the current short-read Illumina sequencing platforms.[319]

Nanopore sequencing offers a shorter processing time, lower capital cost, and high flexibility. At the moment, MinION (Oxford Nanopore) is the most well-known nanopore sequencer that is commercially available and has been successfully employed in multiple fields, including oncology research, agriculture, and outbreak surveillance (i.e.CoVid19, Ebola, Zika virus outbreaks). Unlike other sequencing platforms, nanopore uses electrical signal disruption (as explained in previous chapters), instead of synthesis. Thus, it does not suffer from enzymatic reaction exhaustion when sequencing DNA fragments > 500bp (e.g. long-read sequencing). While nanopore can effectively capture long DNA fragments, it still suffers from (1) low base-calling accuracy fragments shorter than 1kbp,[304]and (2) high amount of DNA input required.

Compared to short-read sequencing platform like Illumina chemistry, read-length of nanopore is the same as the length of the native DNA fragment in cell nucleus. This offers a great advantage in analyzing cfDNA size distribution of the whole cfDNA size-

range, fragmentation, and also gene fusion. While the Oxford Nanopore is attractive with its ability to directly sequence DNA without PCR amplification, this is not yet an option when working with cfDNA. Specifically, most of nanopore library preparation kit requires at least 500ng - 1 µg gDNA input (or 100fmol of shorter DNA), which is approximately 100 times higher than the normal input for other NGS library preparation. As the average concentration of cfDNA is 1 – 10ng/ ml in plasma, whole genome amplification is a necessity to increase the starting amount of material for library preparation.

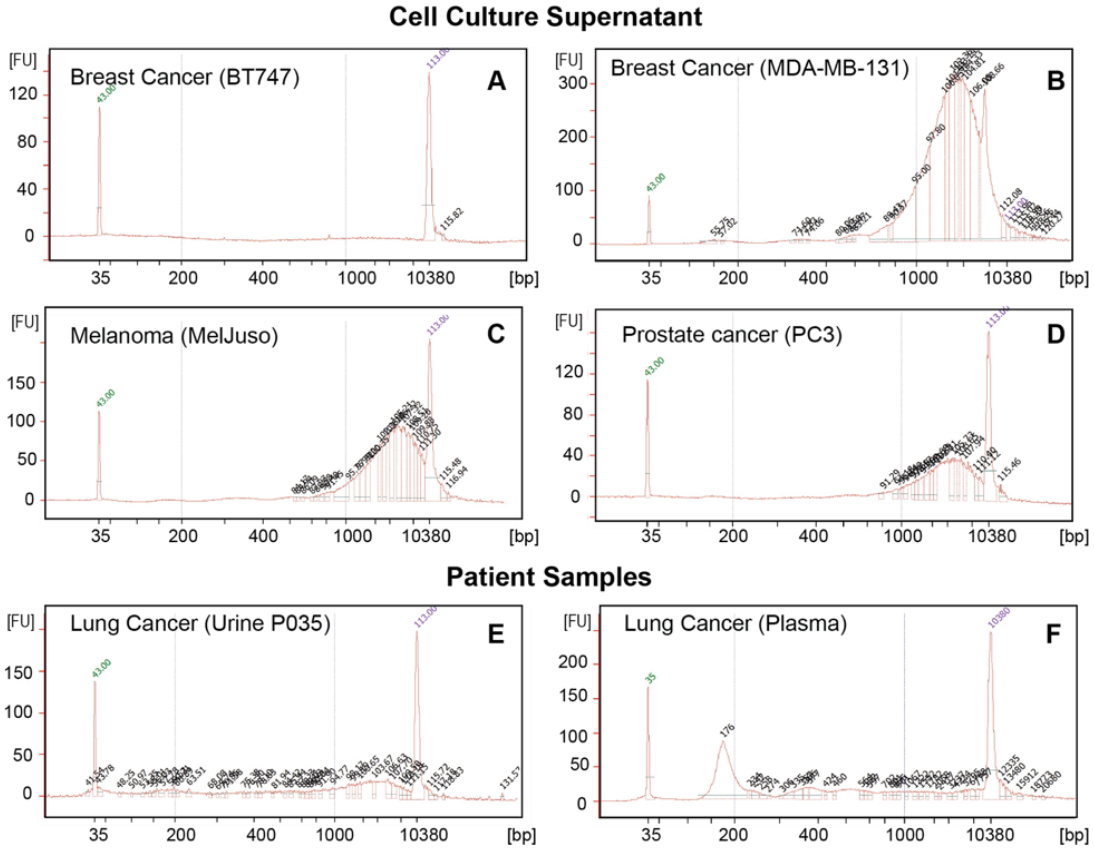


Figure 39. Presence of long cfDNA fragments. Bioanalyzer results of cfDNA extracted from (A-D) cancer cell cultures media, (E) urine sample, and (F) plasma sample of cancer patients. Preliminary results showed that the size distribution and concentration of cfDNA greatly varied among samples and sample types. Besides the characteristic peaks of cfDNA (~166bp), multiple samples exhibited a presence of long cfDNA fragments (1-3kbs), which are often missed in sequencing a downstream analysis.

Thus, it is essential to develop a new protocol for nanopore sequencing that could effectively capture both short and long fragments of cfDNA. As long cfDNA fragments (*e.g.* fusions, extrachromosomal DNA) are currently missed by standard Illumina sequencing, the nanopore could have here a conceptual advantage for analyzing them.

To tackle this issue, we developed two new protocols to perform the preprocessing cfDNA for nanopore sequencing, using: (1) rolling-circle amplification to create long fragments with tandem repeats, and (2) blunt-end ligation to create DNA concatemers. Specifically, we will discuss technology development and evaluate preliminary results of using these two techniques on plasma samples from non-small cell lung cancer patients.

8.3 Materials and Methods

8.3.1 Cell culture and sampling. 143B Human metastatic osteosarcoma cells were cultured in IMDM +/+ (+10% FBS + 1% penicillin/streptomycin). Cells were plated at 30% confluency and allowed to attached for 4hr. Once attached, media was changed to 4mL serum free media. Cells were serum starved for 24hr to synchronize. First collection was done (0h) at the end of serum starvation, and processed. After that, media was changed to 4mL regular growth media and media was collected in duplicate at 4h, 24h, 48h and 72h from the same flask. At 72h, all cells were dead due to lack of space and nutrient and we excluded this timepoint from the study.

8.3.2 Cell sample processing. Upon media collection, duplicated sample was divided into two groups (A and B), then immediately processed to remove supernatant. Specifically, samples in group A were centrifugated at 500 g for 5 minutes (removal of cells), then 2000 g for 5 minutes (removal of large apoptotic bodies). Samples in group B

were processed similar to group A, but had a third centrifugation step at 10,000g for 10 minutes (removal of micro-vesicles). All samples were then stored at -20˚C.

8.3.3 Extraction of cfDNA from cell culture. Samples were thawed at room temperature for 1 hour, then quickly vortexed and spinned down. For each sample, cfDNA was extracted from 500 µL cell media using QIAamp Blood MiniKit (following manufacturer's instruction). Sample was eluted in 200 µL nuclease-free water, and concentrated to 50µL using SpeedVac at 37˚C. Finally, all samples stored at -20˚C until quantification.

8.3.4 Extraction of cfDNA from plasma. Blood samples of non-small cell lung cancer patients were collected aseptically and processed using the standard operating procedure validated by the Liquid Biopsy Center, Cancer Center Amsterdam (the Netherlands). For each sample, cfDNA was extracted from 3 mL plasma using QIASymphony automatic extraction system and magnetic beads chemistry, followed by elution in 90 µL nuclease-free water and stored at -20˚C. Concentration of cfDNA extracted checked with Qubit according to the manufacturer's instructions.

8.3.5 Agilent Bioanalyzer 2100. Size distribution analysis of cfDNA were performed by capillary electrophoresis using Agilent 2100 Bioanalyzer (Santa Clara, CA), in combination with a High-sensitivity DNA Chip (Agilent Technologies). Agilent 2100 Bioanalyzer is a microfluidic-based platform for sizing and quantification of DNA and RNA, using fluorescent dye molecules. The assay was performed according to manufacturer's specific instruction for High-sensitivity DNA chip. After separated analogously to CE, nucleic acids are normalized for size to a ladder and two DNA markers,

which are then represented graphically as a virtual band. Size and concentration of each band are automatically calculated by 2100 Expert Software (B.02.08.SI648, Agilent).

8.3.6 Gel electrophoresis. 2% agarose gel (w/v) was prepared by dissolving 1g agarose in 50mL 1X Tris-Borate-EDTA (TBE) buffer, and 2.5 uL Ethidium Bromine. Gel was casted and settled at room temperature for at least 30 minutes prior to experiment. FastRuler Low Range DNA ladder (Thermo Scientific) and O'GeneRuler 1kb Plus DNA ladder (Thermo Scientific) were used as references. All electrophoresis experiments were performed at 100mV for 15-45 minutes.

8.4 Methods Development

8.4.1 Detection of long cfDNA in cell culture. Release of cfDNA was characterized over time (Figure 40). We divided the collected cell culture media into two group of samples— group A and group B (see Materials and Methods for detailed information on the difference in sample processing). Comparing between the two groups, there was no significant change in peaks positions at any given time point. Specifically, at the end of serum starvation, there was a minimal amount of long cfDNA fragments (~1-3kbps) observed (collection time t = 0, Figure 40A and E). After growth medium renewal, the amount of cfDNA increased incrementally, notably peaked after 24h and continued to increase at 48h. This could be explained by the increasing number of cells in the culture over time. In addition, previous reports reported that more cfDNA was released also per cell.[376-378]

At t = 24h and 48h, we observed an expected peak at 166bp and ladder pattern, which could indicates an apoptotic origin.[379-380] The pattern was reminiscent to that obtained by apoptotic cells was also observed in both sample groups at 24h and 48h.

However, besides a peak at ~160 bps, there was a much more prominent peak near ~2kbps that consistently appeared in all samples, except for at 4h. Previous studies on the same cell line (143B) showed that these long cfDNA fragments were neither from apoptotic or necrotic origin.[380]

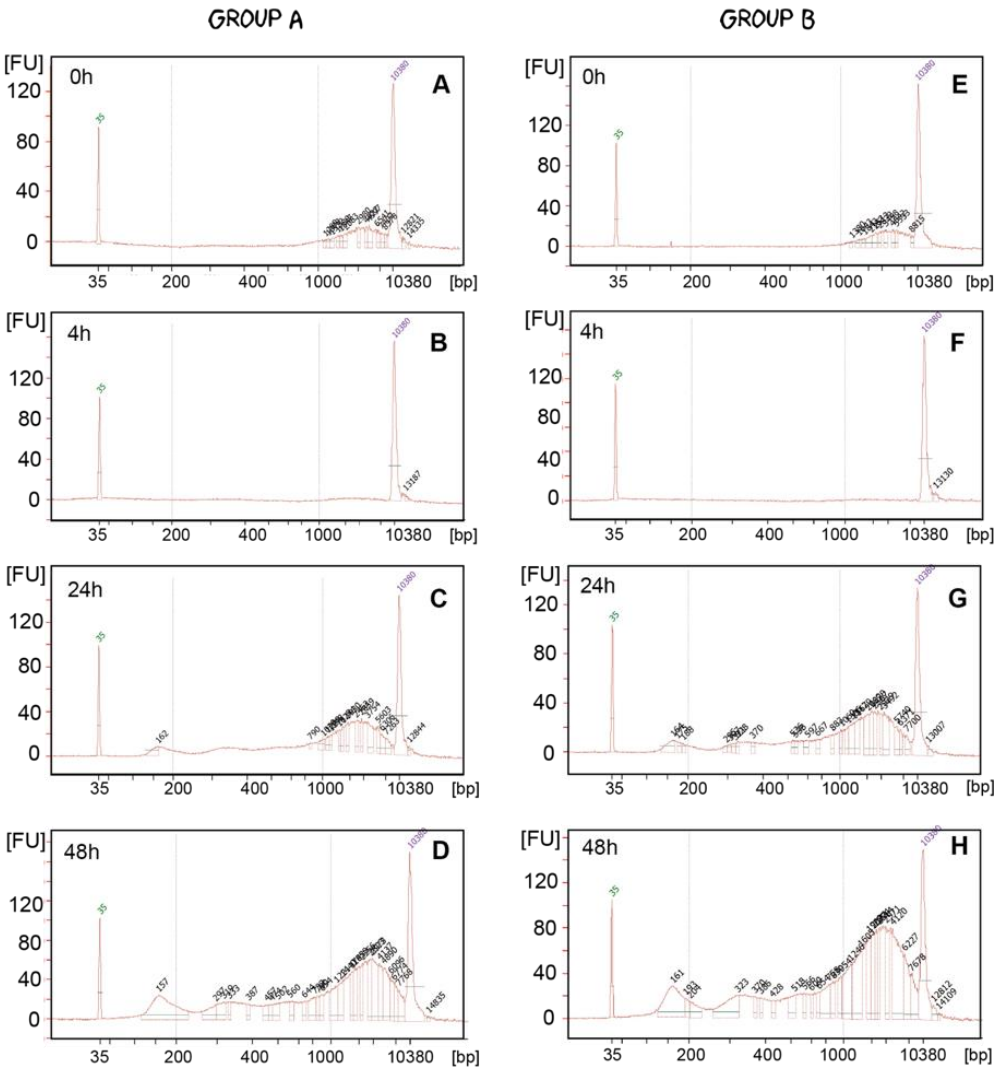


Figure 40. Short and long cfDNA fragments in cell culture overtime. Bioanalyzer results of cfDNA extracted from 143B osteosarcoma cells culture media, collected at 0h, 4h, 24h, and 72h. (A-D) samples in groups A was centrifuged at 0.5g x 5min and 2g x 5min. (E-H) samples in group B was centrifuged at 0.5g x 5min, 2g x 5min, and 10g x 10min. Presence of long cfDNA fragments (1-3kbs) from both groups indicated that they may not come from dying cells, apoptotic bodies, or microvesicles.

Samples in both groups were collected at the same time, but processed slightly differently (Materials and Methods). Because of the difference in centrifugations, samples in group A contained intermediate-size extracellular vesicles (e.g. microvesicles, 100-1000nm), while those in group B only contained smaller extracellular vesicles (e.g. exosomes, 30-100 nm). Thus, it was interesting that concentrations of long fragments (~2kbps) at 24h and 48h are higher in group B than in group A. We suspected that these long fragments might be DNA bound to surfaces of extracellular vesicles that were released with centrifugation.[381] Currently, it is still challenging to access these long fragments using short-read sequencing platforms.

8.4.2 Methods to capture both short and long cfDNA fragments. Herein, we describe two sample-preparation strategies that convert cfDNA into long fragments that can be sequenced on the nanopore systems. The main goal of both protocols was to produce > 500ng of long cfDNA fragments (>1000bps), through two main steps: (a) whole-genome amplification and (b) long-fragment synthesis.

8.4.2.1 Method 1: Rolling-circle amplification. The first sample-preparation strategy enables capturing and sequencing of both short and long cfDNA fragments through circularization, followed by whole-genome amplification of cfDNA using rolling-circle amplification technique. Rolling-circle amplification is an isothermal enzymatic process where short, circularized DNA or RNA is quickly amplified to form a long single-stranded fragment with multiple tandem repeats.[382] As the RCA products can be tailor-designed through manipulation of the original circular template, this technology has been widely utilized for biotechnology, drug delivery, diagnosis, and many other techniques.[382] For each sample, approximately 40-50ng extracted cfDNA from the plasma of lung cancer

patients were first end-repaired to create blunt end fragments, then cleaned up with a commercial reaction cleanup kit. After that, the blunt-end fragments were a phosphorylated at the 5' end. This population will be referred to as the insert (I). At the end of each step, the sample was cleaned up using a column-based reaction clean-up kit in order to remove excess enzymes and waste products. After the second clean up, we used Qubit to quantify the concentration of end-repaired DNA samples. The Qubit results indicated that there were approximately 4.5ng of end-repaired DNA product, reflecting approximatively a 90% loss of DNA input.

As nanopore sequencing requires at least 500ng -1µg of input DNA (or 50-100 fmol of shorter DNA fragments), it was essential to perform a whole genome amplification. In this protocol, the circular template for amplification consisted of two components: the cfDNA insert (I) and the barcoded backbone sequence (BB). The BB sequence (a 250bp blunt-end, synthetic DNA fragment) was designed to contain a phi29 DNA polymerase binding domain, a restriction enzyme cutting site, DNA barcode, and a CRISPR cutting domain. Phi29 DNA polymerase was chosen for its exceptional processivity and strand displacement ability.[383]

Note that in order for ligation to happen, at least one fragment (either I or BB) needs to be phosphorylated. Backbone and insert cfDNA were mixed and incubated together for 5.6 hr, allowing ligation and circularization. The BB was unphosphorylated, in order to minimize the amount of backbone-backbone ligation (e.g. containing no cfDNA). After ligation, there were four possible products: circularized BB:I, circularized BB:BB, circularized I:I, and linear fragments. As the circularized BB:BB population contained no biological information of the tested sample, we discarded this group to reduce waste in

sequencing. In order to do so, backbone DNA was designed so that when two BB:BB ligated, they would form a sequence recognized and cut by CRISPR. This step prevented BB:BB products from being amplified by phi29 and sequenced.

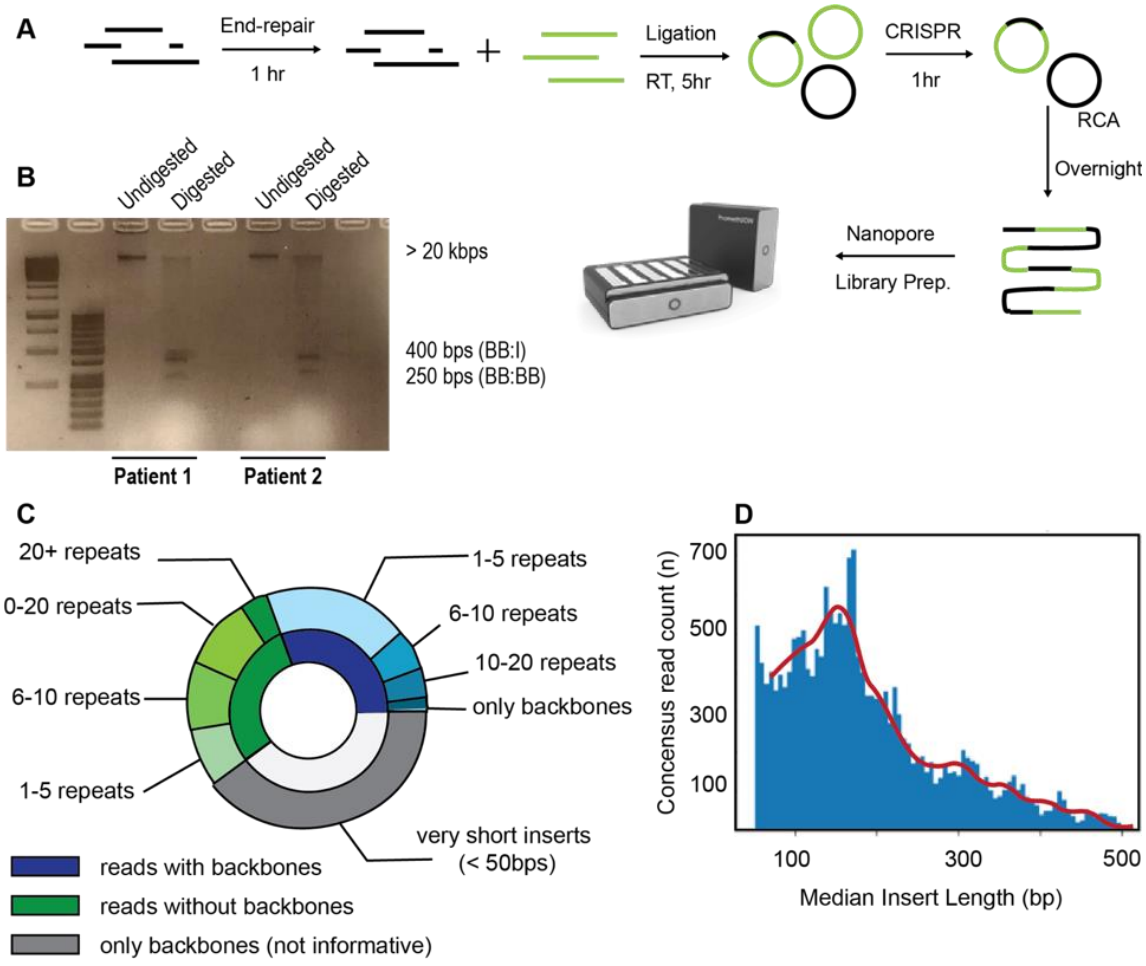


Figure 41. Flow chart of sequencing using a guided system. (A) Each cfDNA fragment is end-repaired and attached to a backbone, then amplified with rolling circle amplification. This procedure takes approximately two days, yielding an end product containing long DNA fragments, each of which contains tandem repeats of the backbone and one insert cfDNA. After that, the sample is ready for nanopore library preparation and sequencing (24-48hrs). (B) Gel image of DNA after RCA reaction. (C) Preliminary results of all sequencing reads (Patient 1) from a Flongle flow cell after 48 hours. There were approximately 70,000 reads in total. (D) Histogram of the insert (cfDNA) length distribution.

The results from the gel electrophoresis (Figure 41B) indicated that for both samples (Patient 1 and 2), RCA reaction creates DNA fragments of over 20kbps (undigested). Upon digestion with restricted enzyme, digested samples exhibited three distinct bands at 250, 400, and 20kbps+. The first two bands corresponded to BB:BB and BB:I fragments, respectively. Long fragments that remained after restriction enzyme digestion could result from inefficient digestion, or simply because they did not contain a backbone sequence (e.g. multiple insert cfDNA ligated together).

Next, RCA primers were annealed to the circular template. RCA reaction was carried out with 12h incubation (overnight) at 30˚C, then heat inactivated at 60˚C for 10 minutes. RCA reaction yielded > 100µg final product, providing a sufficient amount of DNA for quality control and multiple sequencing runs. After a final clean up with Dynabeads, the RCA products were ready for library preparation using 1D Ligation Kit (Oxford Nanopore Technology) and sequencing.

To evaluate the efficiency of the technique, we performed whole-genome sequencing on an Oxford Flongle using the RCA product from Patient 1. Compared to the MinION's traditional flowcell, a Flongle only has 126 channels, instead of 512; thus, the output yield is also significantly reduced. However, as the cost per sample is low (approximately $200), it is suitable for smaller experiments, such as preliminary testing and quality check. Thus, we first sequenced the RCA product from Patient 1 sample on a Flongle flow cell. A 48 hr run yielded approximately 70,000 reads, each of which contained information from a unique fragment post-RCA. Figure 41C shows that about 10% of the sequenced fragments had no insert (B:B). Only one third of all the reads (blue) had the expected combination (BB:I, with I > 50bps), whereas the other one third (green) contained

no backbone (I:I). The rest of the reads (gray) contained extremely short inserts (BB:I, with I < 50bps) and were discarded from analysis. The number of repeats indicated how many tandem repeats a fragment had. As RCA is an enzymatic reaction, the number of tandem repeats could not be precisely controlled.

We selected reads with more than five tandem repeats (approximately 10,000 of the 70,000 total reads) to examine the distribution of cfDNA fragment length (Figure 41D). As can be seen, the histogram is left-skewed, with a prominent peak near 166 bps, similar to read length distribution obtained from Illumina sequencing (data not shown). However, there is a long right tail representing cfDNA fragments of up to 500bps. Also, it is important to note that Figure 41D only reflects 10% of the total sequencing data.

At the moment, we are working on our bioinformatic pipeline to include RCA fragments that only contained inserts (I:I). Since the backbone sequence was 250bp, the optimal length of insert DNA ranged between 90 – 160 bp. While the system could capture insert fragments outside of the optimal range (as shown in Figure 41D), its efficiency had not been determined. Thus, we expected the fragment size distribution of fragments without BB would be quite different from the one shown in Figure 41D. As preliminary data obtained from Flongle flow cell showed promising results of capturing long cfDNA fragments, we multiplexed and sequenced RCA products of both patients 1 and 2 using a PromethION flow cell. While the cost of running a PromethION flow cell is about 10 times higher than a Flongle, the actual cost per Gb of data is much lower, due to higher output and ability of multiplexing.

8.4.2.2 Method 2: DNA Concatenation. The second method of sample-preparation produces a long cfDNA concatemers through a straightforward procedure of whole-

genome amplification followed by ligation (Figure 42A). First, we used a commercially available NGS library preparation kit to perform whole-genome amplification of extracted cfDNA (Plasma-Seq, Takara). For this reaction, up to 10 µL of isolated cfDNA (e.g. 1-10ng total) were used for reaction input. Briefly, isolated cfDNA was first end-repaired, then blunt-end ligated to NGS sequencing adaptors (i.e. library synthesis). Sequencing indexes were then added through high-fidelity amplification using PCR (i.e. library amplification). After the addition of sequencing adaptors and indexes to the two ends, length of cfDNA fragments increased by 150bps (length of adapters + indexes). Finally, we used magnetic beads to clean up the post PCR product. This process of NGS library preparation took approximately 3 hours, yielding 100-400ng DNA output.

Then, we performed phosphorylated the NGS library for 30 minutes at room temperature, allowing blunt-end ligation in the next step. Long DNA concatemers were created by adding NEB Blunt-end ligation mixture to the sample and incubated at room temperature for four hours, to maximize the amount and length of DNA concatemers. Finally, we cleaned up the reaction and eluted final product in nuclease-free water. To check for concatemers quantity and size distribution after the reaction, we used Qubit and a DNA high sensitivity chip with Bioanalyzer Agilent 2100, respectively.

The final clean up step was necessary to remove T4 ligase enzyme and any reaction chemicals that could potentially interfere with sequencing. During the development of this protocol, we tested both column-based and gel-based purification methods. The latter one was quickly ruled out due to low recovery (only up to 80%) and long processing time (over 20 minutes). As our target was whole-genome sequencing, rather than a size-specific fragment, column-based purification was referred for a faster procedure and higher

recovery. It is important to note that the DNA size selection range of the selected column should be carefully considered, in order to maximize recovery of long fragments while efficiently removing T4 ligase enzymes. In this protocol, we used NucleoSpin Gel and PCR Clean up kit (Takara Bio), which allowed recovery of fragments between 50 – 20,000bps (within 10 minutes).

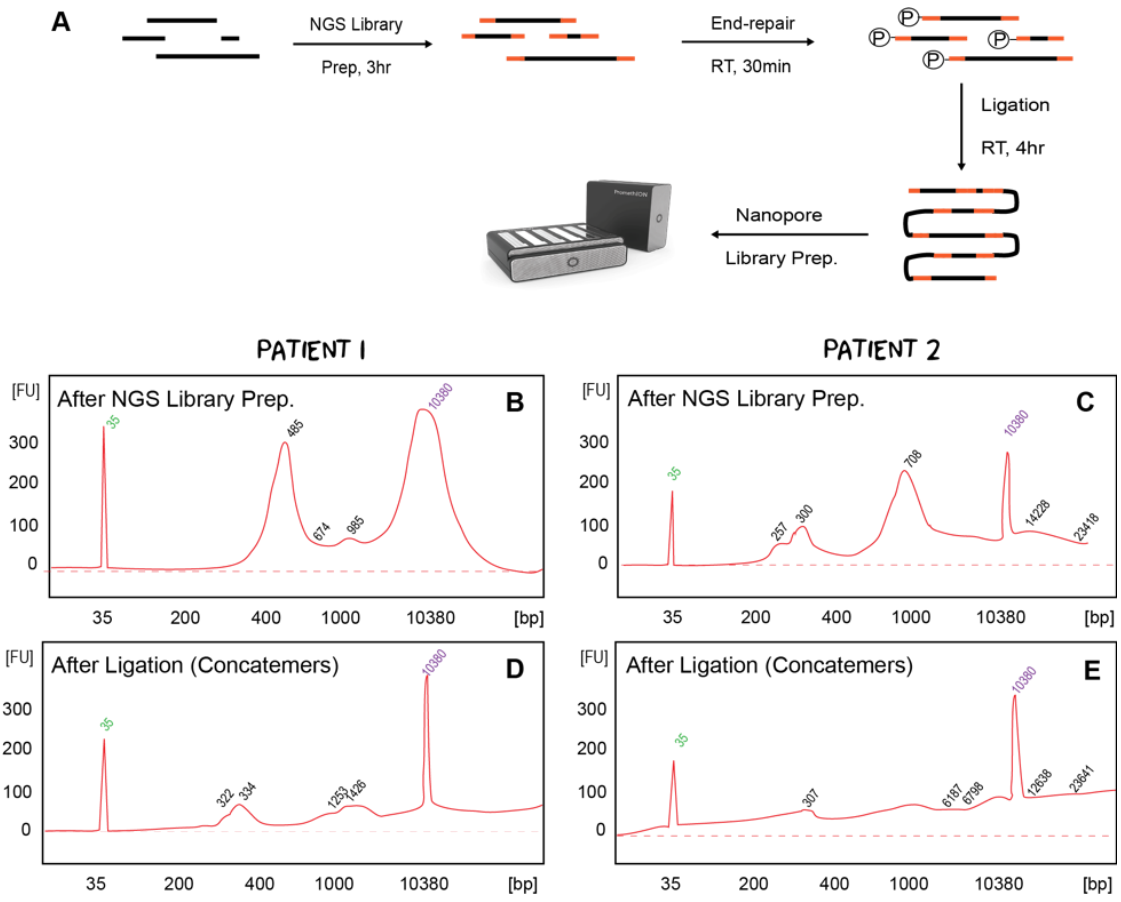


Figure 42. Sample-preparation method: Concatenation. (A) Flow chart of sequencing using DNA concatenation. Sample containing extract cfDNA undergoes the standard Next-generation sequencing library preparation, followed by end-repair and blunt-end ligation to create long DNA fragments. The procedure takes approximately one day. (B-E) Bioanalyzer traces of Patients 1 and 2 samples after step 1 (NGS library preparation) and step 2 (blunt-end ligation / concatenation). Disappearance of the prominent peaks, and upward-shifting baseline in the second step indicated successful formation of long DNA fragments.

Overall, this second strategy of sample-preparation takes approximately 8 hours and yield between 100-400ng of DNA that can be directly used for ONT 1D ligation library preparation and nanopore sequencing. The final amount of concatemer obtained strongly depends on the output of whole-genome amplification with PCR.

We applied this protocol to prepare samples from lung cancer Patient 1 and 2 for nanopore sequencing. Comparing the bioanalyzer trace at each step (isolated cfDNA – after library amplification – after DNA concatenation), we observed a clear shift in fragment size distribution (Figure 42B – G). After DNA concatenation (Figure 42D and G), the prominent peaks that previously showed Figure 42B-C and E-F disappeared, replaced by multiple smaller peaks that stretched passed the end marker (10380 bp).

These results indicated successful ligation and formation of DNA concatemers with random length. Since ONT 1D ligation library preparation kit requires at least 500ng – 1µg DNA input per sequencing run, we plan to pool these two samples (from Patient 1 and 2) together in one nanopore flow cell. As these two samples contained two different NGS index combinations, it is possible to sequence them as is, without a second step of multiplexing using ONT barcode. Currently, a new bioinformatic pipeline is under development to analyze the sequencing results of DNA concatemers prepared using this second strategy. As each DNA concatemers consist of several individual cfDNA fragments that are separated to each other by a duplicate pair of adaptor and index, we need this bioinformatic tool to: (1) divide one concatenated strand into its component fragments, (2) trim the adaptor and index at 3' and 5' end of each fragment, (3) align the fragments to a reference human genome, and (4) perform downstream analysis (i.e. copy number variation, SNP, fragmentation, methylation and bacterial DNA.).

Table 16

Comparison between sequencing methods

	Reference: Illumina (NGS)	Method 1: RCA	Method 2: Concatenation
Amount cfDNA input	< 10 ng	40 ng[1]	< 10 ng
Amount DNA output[2]	100 – 500 ng	> 100 µg	100 – 500 ng
Sample preparation time	3 – 5 hours	~ 24 hours	8 – 10 hours
Sequencing time	24 – 80 hours[3]	24 – 48 hours	24 – 48 hours
Wait time	~ 3 weeks	-	-
Read length	Short (< 500 bps)	Fragment length	Fragment length

* Note:

1 About 90% of cfDNA input was lost before ligation of cfDNA and backbone, due to reaction cleanups and fragment blunting.

2 Amount of DNA output after whole-genome amplification.

3 Run time depends on the sequencer model.

8.5 Results and Discussion

So far, we have shown evidences of long cfDNA fragments, and described in detailed two recently developed strategies for sample-preparation before nanopore sequencing. These protocols enabled capture of both short and long cfDNA fragments. In Table 16, we give a brief comparison between the two strategies and Illumina sequencing. These two protocols, while having the same end goal, had different wet lab approaches that brought in unique advantages and disadvantages. Specifically:

8.5.1 Method 1: Rolling-circle amplification. Summary:

(1) High accuracy: End product consisted of long fragments with tandem repeats, allowing a highly accurate sequencing result that could be used for SNP detection (i.e. mutation).

(2) High amount of output: RCA reaction yielded over 100 µg DNA, which exceeds the input requirement of nanopore sequencing; thus, allowing quality controls and multiple testing.

(3) Fragment length: ability to capture long-fragments that previously missed by short-read platforms, as indicated by preliminary sequencing with the Flongle flow cell.

(4) Long processing time: Sample preparation time took 24 hr total. However, there was a long waiting period between steps, and the process actually took 2-3 working days, which is not practical in clinics.

(5) High amount of isolated cfDNA input: depending on the concentration of cfDNA, 4 – 40mL of plasma is needed, in order to obtain 40-50ng cfDNA input per sample. Moreover, as 90% of the DNA input was lost before RCA reaction, the diversity of the RCA product was significantly reduced.

8.5.2 Method 2: DNA concatenation. Summary:

(1) Simpler and shorter processing time: sample preparation plus ONT library preparation takes approximately 8 hours. Thus, DNA sequencing can be started within the same day.

(2) Adaptable procedure from short-read technology: since the first half of this protocol used a common NGS library preparation method, this protocol is highly convenient for laboratories that currently rely on short-read platforms, but would like to try out long-read sequencing. Specifically, the sample can be divided for sequencing with both short-read platform (i.e. Illumina) and long-read platform (i.e. Nanopore).

(3) Low DNA output: as nanopore sequencing requires at least 50-100fmol (or 500ng – 1µg) of DNA input, at least two or more samples will have to be multiplexed for a sequencing run.

(4) Unavailable bioinformatic pipeline: This protocol is a combination of both NGS and nanopore sequencing, thus it requires a new, adapted pipeline. We estimate that it will take approximately 1-3 months to develop and employ this new bioinformatic pipeline.

8.6 Conclusions

Up-to-date, long fragments of cfDNA are still missing from the analysis, due to bias in sample preparation and sequencing limitations.[326] Nanopore technology can be a potential tool to retrieve these missing fragments, thanks to its ability to perform long-read sequencing. However, sequencing ctDNA with nanopore sequencer is still challenging, due to the high input requirements and low base-calling accuracy with fragments shorter than 1kbps. In this study, we showed two novel sample-preparation methods that can capture both short and long cfDNA fragments for sequencing with high accuracy using nanopore technology. At the moment, these methods are still under development and optimization to lower the cost and maximize sequencing output. We have sent samples from patients 1 and 2 to sequence with both Illumina and nanopore sequencers. In the future, it is our goal to conduct a full-scale comparison of the advantages and disadvantages of WGS using three techniques: Next-generation sequencing (Illumina), Nanopore sequencing (using RCA and DNA concatenation strategies).

Chapter 9

Summary and Conclusion

This thesis focuses on exploring the α-hemolysin nanopore's capability to detect and study molecular markers of cancer for precision medicine. The studies described herein, and summarized below, showcase different techniques to improve nanopore sensitivity in molecular detection and demonstrate how to bring the application of this technique from the lab-bench to the clinic.

In **Chapter 2**, current work on applying nanopore technology in studying cancer biomarkers is summarized. We provided an overview of the essential genetic and epigenetic modifications in cancerous tissue and the progression of cancer cells. With the complexity of the human body and more specifically cancer tissues, many of the mechanisms for cancer proliferation remain unknown. Nanopore membranes have shown their ability to detect various biomolecules' chemical and structural modifications, as well as genetic and epigenetic modifications. There is a growing opportunity for more significant research in this field to be conducted in the next few decades.

Unlike solid-state nanopores, whose size can vary, α-hemolysin is a bacterial toxin that forms a nanopore with well-defined structure and consistent dimensions. This feature allows a reproducible sensing of molecular structure. Taking advantage of the α-hemolysin nanopore's size selection, in **Chapter 3**, we evaluated the pore's ability to sense DNA secondary structures that were different in size. The employed C-rich single-stranded DNA were able to adopt an i-motif structure in solutions with pH less than 6.15 and a linear strand when pH is higher than 6.15. However, we found that in highly concentrated buffers (1M KCl and 1M NaCl) cations can bind to and induce the compaction process on C-rich

DNA, as the exocyclic amino group of cytosine allows strong cation-nucleotide binding because of resonance effects. This process created secondary structures larger than 2.6nm (nanopore *cis* opening), resulting in shallow current blockage events (%I/Io < 30%) that could be sensed by the pore. In either type of solution (1M NaCl or 1M KCl), both DNA compacting and translocating events were observed, hinting that the 15C-ssDNA adopted both elongated and compacted forms and could spontaneously shift from one conformation to the other.

Formation of G-quadruplexes in the promoter region of Ckit1 proto-oncogene suppresses gene expression and prevents cancer cell replication. Excessive activation of the C-kit1 DNA sequence is implicated in several human malignancies and its expression is downregulated by the stabilization of its G-quadruplex structure. In **Chapter 4**, we studied the formability, stability and responsiveness to CX-5461 cancer drug of Ckit1 G-quadruplex in K^+ and Na^+ environment. Using circular dichroism, we determined that Ckit1 G-quadruplex adopted a parallel folding topology in both environments, with a more tightly packed structure in the presence of K^+. While formation of a stable G-quadruplex was observed in both environments, Na^+ G-quadruplex folded much slower and unfolded more rapidly than K^+ G-quadruplex. Binding of CX-5461 effectively increased the structural stability and structural volume of G-quadruplex in both Na^+ and K^+ environments. Subsequent validation by thermal denaturation with circular dichroism revealed a 13˚C and 11˚C increase in melting temperature of (CX-5461 bound versus unbound) K^+ and Na^+ G-quadruplex, respectively. Moreover, through exploratory analysis using clustering, we identified two distinct populations of current blockage events, representing the two stages of (i) capturing (%I/Io = 67.4±3.9%) and (ii) translocation of G-quadruplex through the

pore (%I/Io = 85.1±2.9%). The volume of G-quadruplex is correlated with the cation species and presence of CX-5461. As the stability and volume of Ckit1 G-quadruplex were dependent on cations and ligand-binding status, it was feasible to distinguish different structures of Ckit1 G-quadruplex using nanopore data alone (e.g. predicting the binding state of CX-5461 to Ckit1 G-quadruplex in Na^+ environment: 92.4% sensitivity, 70.3% specificity, and 81.3% average accuracy).

In addition to binding of ligands (e.g. CX5461), other factors, such as point mutation, can also strongly affect the stability of G-quadruplex. Activating mutation on Ckit1 has been observed in several types of malignancy; however, their direct effect on cancer drug efficacy (i.e. CX5461) remains unclear. Therefore, in **Chapter 5**, we evaluated the ability of mutated Ckit1 sequences to form G-quadruplex, as well as their stability with respect to ligand binding. We employed six samples that were variations of the Ckit1 sequence, including one wild-type (control) and five sequences containing 1-6 guanine mutations. Nanopore dwell time analysis of these mutated structures revealed a general decrease in G-quadruplex structural stability, with the most significant changes present when there was more than one mutation of the same G-quartet. Thermal denaturation study with circular dichroism showed that sequences with more than one mutation on the same G-quartet lost their characteristic positive peak (at 263nm), hinting a partial or uncomplete formation of G-quadruplex. While addition of CX5461 to mutated Ckit1 DNA sequences increased the overall stability of G-quadruplex structure formed by most samples, sequences with mutations on the outer G-quartet did not exhibit statistically significant changes. This result indicated the important role of the outer G-quartet for ligand-binding and stabilizing the structure. Furthermore, we implemented random forest classifier to

nanopore sensing data to distinguish mutated from unmutated Ckit1 DNA (AUC = 0.89, sensitivity = 81.9%, specificity = 83.4% and balanced accuracy = 82.7%). Understanding the potential effect of mutation on drug efficiency can have implications in the implementation of cancer therapies. Our results indicate the potential of using a nanopore biosensor to study DNA molecular stability and their interaction with cancer drugs for constructing methods to predict therapeutic response, as well as personalizing cancer treatment.

In addition to DNA secondary structure, other types of DNA modifications, such as methylation of DNA, are important markers for cancer diagnosis and therapeutic response prediction. While conventional methods can detect cytosine methylation with high accuracy, they still require a high amount of input and extensive sample processing. In **Chapter 6**, we performed, for the first time, label-free detection of cytosine methylation on single-stranded DNA, using an α-hemolysin nanopore sensor. One of the main challenges with nanopore sensing is low accuracy due to fast translocation speed of biomolecules through the pore. To tackle this issue, we employed a concentration gradient of salt across the experimental chambers. As the cations move down the concentration gradient, they move against DNA translocation direction, thus, significantly lowering the electrophoretic drive of the system and DNA translocation speed. With a 15-fold increase of LiCl concentration going from the *cis* to the *trans* chamber, DNA translocation speed was effectively decreased by 5 times, allowing a clear distinction between methylated and unmethylated DNA. Distribution of current blockage amplitude for methylated – unmethylated DNA mixtures exhibited bimodule distributions, with peaks' height relatively proportional to the mixture content (unmethylated : methylated DNA ratio).

Results of this study indicate the potential of simplifying label-free cytosine methylation detection with nanopore, using LiCl salt concentration asymmetry.

While label-free detection of cytosine methylation is feasible, current blockage amplitude of methylated DNA varies based on the local-sequence context, thus, complicating the computational aspect. Therefore, in **Chapter 7**, we employed methyl-binding protein (MBD2) to enhance the signal of methylated cytosine going through the nanopore. Binding of MBD2 to duplex DNA increased the structural stability and significantly prolonged the event translocation time through the pore. Applying machine learning classification models on nanopore results revealed that utilizing MBD2 protein provided a good discriminatory power for cytosine methylation detection with an optimal AUC of 0.83, 88% sensitivity and 86% specificity (using KNN algorithm).

Finally, we explored the potential of applying nanopore technology to capture cancer biomarkers in a clinical setting. Liquid biopsy—a technique to study cancer through circulating-tumor DNA—offers a minimally invasive and reliable detection of tumors' genomic evolution through longitudinal sampling. Current work with cell-free DNA obtained from cancer cell culture media reflected the existence of long cfDNA fragment (>1kbps). However, current sequencing techniques, which depend on enzymatic reaction, are limited by the read-length (<500bps). Compared to conventional methods, the nanopore offers a shorter sample processing time and the ability to capture long cfDNA fragments. However, because of their rapid translocation speed, short DNA fragments (<1kbps) are often missed by the nanopore. In **Chapter 8**, we developed two methods to capture both short and long fragments of cell-free DNA using commercially-available nanopore assay. In the first method, we utilized a guide system to capture cfDNA, then performed rolling-

circle amplification to amplify and convert them into long fragments with tandem repeats. In the second method, we used standard sequencing library preparation (NGS library) to perform whole-genome amplification, after which, long fragments were generated through blunt-end ligation (Gibson assembly). We tested the two methods using plasma samples obtained from the same lung cancer patients. Sequencing results with Flongle showed successful capture of both short and long cfDNA fragments (50bps to over 600bps). While further validations and optimization are needed, our study demonstrated the potential benefits and applications of nanopore in sequencing cfDNA for monitoring treating progression.

Overall, in this thesis, we aimed to validate and improve α-hemolysin nanopore assays toward precision medicine in oncology. Specifically, we investigated the potential of lab-based and commercially available nanopore assays for molecular sensing and ctDNA sequencing. While lab-based assay offers the flexibility for controlling and fine-tuning of parameters, commercial assay provides higher sensing resolution and throughput. Each method certainly has its own potential area in precision medicine. Thanks to its low capital cost and small footprint, nanopore technology offers a higher mobility and more flexibility than conventional techniques. Our study demonstrated the nanopore's ability for rapid molecular sensing and sequencing (within a few days, compared to the standard three weeks' waiting period for NGS sequencing). The biggest challenge remaining for both types of assays is their high input requirement, compared to clinical standards. Therefore, there is a need for future advancements in both wet lab and technological development for a multi-targets nanopore assay.

References

1. Cooper, G. M.; Hausman, R. E., The development and causes of cancer. *The cell: A molecular approach* **2000**, 725-766.

2. Anand, P.; Kunnumakara, A. B.; Sundaram, C.; Harikumar, K. B.; Tharakan, S. T.; Lai, O. S.; Sung, B.; Aggarwal, B. B., Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical research* **2008,** *25* (9), 2097-2116.

3. Croce, C. M., Oncogenes and cancer. *New England journal of medicine* **2008,** *358* (5), 502-511.

4. Libermann, T. A.; Zerbini, L. F., Targeting transcription factors for cancer gene therapy. *Current gene therapy* **2006,** *6* (1), 17-33.

5. Herceg, Z.; Hainaut, P., Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Molecular oncology* **2007,** *1* (1), 26-41.

6. Bielas, J. H.; Loeb, K. R.; Rubin, B. P.; True, L. D.; Loeb, L. A., Human cancers express a mutator phenotype. *Proceedings of the National Academy of Sciences* **2006,** *103* (48), 18238-18242.

7. Feinberg, A. P.; Ohlsson, R.; Henikoff, S., The epigenetic progenitor origin of human cancer. *Nature reviews genetics* **2006,** *7* (1), 21-33.

8. Jaenisch, R.; Bird, A., Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics* **2003,** *33* (3), 245-254.

9. Egger, G.; Liang, G.; Aparicio, A.; Jones, P. A., Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **2004,** *429* (6990), 457-463.

10. Takeshima, H.; Ushijima, T., Accumulation of genetic and epigenetic alterations in normal cells and cancer risk. *NPJ Precision Oncology* **2019,** *3* (1), 1-8.

11. Friedman, A. A.; Letai, A.; Fisher, D. E.; Flaherty, K. T., Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer* **2015,** *15* (12), 747-756.

12. Ku, C.-S.; Roukos, D. H., From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic medicine. *Expert review of medical devices* **2013,** *10* (1), 1-6.

13. Tran, B.; Dancey, J. E.; Kamel-Reid, S.; McPherson, J. D.; Bedard, P. L.; Brown, A. M.; Zhang, T.; Shaw, P.; Onetto, N.; Stein, L., Cancer genomics: technology, discovery, and translation. *Journal of Clinical Oncology* **2012,** *30* (6), 647-660.

14. Metzker, M. L., Sequencing technologies—the next generation. *Nature reviews genetics* **2010,** *11* (1), 31-46.

15. Shah, S. P.; Roth, A.; Goya, R.; Oloumi, A.; Ha, G.; Zhao, Y.; Turashvili, G.; Ding, J.; Tse, K.; Haffari, G., The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **2012,** *486* (7403), 395-399.

16. Banerji, S.; Cibulskis, K.; Rangel-Escareno, C.; Brown, K. K.; Carter, S. L.; Frederick, A. M.; Lawrence, M. S.; Sivachenko, A. Y.; Sougnez, C.; Zou, L., Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **2012,** *486* (7403), 405-409.

17. Ellis, M. J.; Ding, L.; Shen, D.; Luo, J.; Suman, V. J.; Wallis, J. W.; Van Tine, B. A.; Hoog, J.; Goiffon, R. J.; Goldstein, T. C., Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **2012,** *486* (7403), 353-360.

18. Curtis, C.; Shah, S. P.; Chin, S.-F.; Turashvili, G.; Rueda, O. M.; Dunning, M. J.; Speed, D.; Lynch, A. G.; Samarajiwa, S.; Yuan, Y., The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **2012,** *486* (7403), 346-352.

19. Ross-Innes, C. S.; Stark, R.; Teschendorff, A. E.; Holmes, K. A.; Ali, H. R.; Dunning, M. J.; Brown, G. D.; Gojis, O.; Ellis, I. O.; Green, A. R., Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **2012,** *481* (7381), 389-393.

20. Fojo, T. In *Precision oncology: a strategy we were not ready to deploy*, Seminars in oncology, 2016; p 9.

21. Prasad, V., Perspective: The precision-oncology illusion. *Nature* **2016,** *537* (7619), S63-S63.

22. Prasad, V.; Fojo, T.; Brada, M., Precision oncology: origins, optimism, and potential. *The Lancet Oncology* **2016,** *17* (2), e81-e86.

23. Tannock, I. F.; Hickman, J. A., Limits to personalized cancer medicine. *N Engl J Med* **2016,** *375* (13), 1289-1294.

24. Gupta, G. P.; Massagué, J., Cancer metastasis: building a framework. *Cell* **2006,** *127* (4), 679-695.

25. Chaffer, C. L.; Weinberg, R. A., A perspective on cancer cell metastasis. *science* **2011,** *331* (6024), 1559-1564.

26. Seyfried, T. N.; Huysentruyt, L. C., On the origin of cancer metastasis. *Critical reviews in oncogenesis* **2013,** *18* (1-2), 43.

27. Yates, L. R.; Gerstung, M.; Knappskog, S.; Desmedt, C.; Gundem, G.; Van Loo, P.; Aas, T.; Alexandrov, L. B.; Larsimont, D.; Davies, H., Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature medicine* **2015,** *21* (7), 751.

28. Gerlinger, M.; Rowan, A. J.; Horswell, S.; Larkin, J.; Endesfelder, D.; Gronroos, E.; Martinez, P.; Matthews, N.; Stewart, A.; Tarpey, P., Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England journal of medicine* **2012,** *366* (10), 883-892.

29. Gerlinger, M.; Horswell, S.; Larkin, J.; Rowan, A. J.; Salm, M. P.; Varela, I.; Fisher, R.; McGranahan, N.; Matthews, N.; Santos, C. R., Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature genetics* **2014,** *46* (3), 225.

30. Zhang, J.; Fujimoto, J.; Zhang, J.; Wedge, D. C.; Song, X.; Zhang, J.; Seth, S.; Chow, C.-W.; Cao, Y.; Gumbs, C., Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **2014,** *346* (6206), 256-259.

31. Castro-Giner, F.; Gkountela, S.; Donato, C.; Alborelli, I.; Quagliata, L.; Ng, C.; Piscuoglio, S.; Aceto, N., Cancer diagnosis using a liquid biopsy: challenges and expectations. *Diagnostics* **2018,** *8* (2), 31.

32. Schweiger, M. R.; Kerick, M.; Timmermann, B.; Isau, M., The power of NGS technologies to delineate the genome organization in cancer: from mutations to structural variations and epigenetic alterations. *Cancer and Metastasis Reviews* **2011,** *30* (2), 199-210.

33. Venkatesan, B. M.; Bashir, R., Nanopore sensors for nucleic acid analysis. *Nature nanotechnology* **2011,** *6* (10), 615-624.

34. Howorka, S.; Siwy, Z., Nanopore analytics: sensing of single molecules. *Chemical Society Reviews* **2009,** *38* (8), 2360-2384.

35. Deamer, D. W.; Branton, D., Characterization of nucleic acids by nanopore analysis. *Accounts of chemical research* **2002,** *35* (10), 817-825.

36. Gu, L.-Q.; Wanunu, M.; Wang, M. X.; McReynolds, L.; Wang, Y., Detection of miRNAs with a nanopore single-molecule counter. *Expert review of molecular diagnostics* **2012,** *12* (6), 573-584.

37. Shim, J.; Humphreys, G. I.; Venkatesan, B. M.; Munz, J. M.; Zou, X.; Sathe, C.; Schulten, K.; Kosari, F.; Nardulli, A. M.; Vasmatzis, G., Detection and quantification of methylation in DNA using solid-state nanopores. *Scientific reports* **2013,** *3*.

38. Venkatesan, B. M.; Dorvel, B.; Yemenicioglu, S.; Watkins, N.; Petrov, I.; Bashir, R., Highly sensitive, mechanically stable nanopore sensors for DNA analysis. *Advanced Materials* **2009,** *21* (27), 2771-2776.

39. Feng, Y.; Zhang, Y.; Ying, C.; Wang, D.; Du, C., Nanopore-based fourth-generation DNA sequencing technology. *Genomics, proteomics & bioinformatics* **2015,** *13* (1), 4-16.

40. Rhee, M.; Burns, M. A., Nanopore sequencing technology: nanopore preparations. *TRENDS in Biotechnology* **2007,** *25* (4), 174-181.

41. Deng, T.; Li, M.; Wang, Y.; Liu, Z., Development of solid-state nanopore fabrication technologies. *Science bulletin* **2015,** *60* (3), 304-319.

42. Kwok, H.; Briggs, K.; Tabard-Cossa, V., Nanopore fabrication by controlled dielectric breakdown. *PloS one* **2014,** *9* (3), e92880.

43. Miles, B. N.; Ivanov, A. P.; Wilson, K. A.; Doğan, F.; Japrung, D.; Edel, J. B., Single molecule sensing with solid-state nanopores: novel materials, methods, and applications. *Chemical Society Reviews* **2013,** *42* (1), 15-28.

44. Desai, T. A.; Hansford, D. J.; Kulinsky, L.; Nashat, A. H.; Rasi, G.; Tu, J.; Wang, Y.; Zhang, M.; Ferrari, M., Nanopore technology for biomedical applications. *Biomedical Microdevices* **1999,** *2* (1), 11-40.

45. Gu, L.-Q.; Shim, J. W., Single molecule sensing by nanopores and nanopore devices. *Analyst* **2010,** *135* (3), 441-451.

46. Banerjee, S.; Shim, J.; Rivera, J.; Jin, X.; Estrada, D.; Solovyeva, V.; You, X.; Pak, J.; Pop, E.; Aluru, N., Electrochemistry at the Edge of a Single Graphene Layer in a Nanopore. *ACS nano* **2012,** *7* (1), 834-843.

47. Shim, J.; Gu, L.-Q., Single-molecule investigation of G-quadruplex using a nanopore sensor. *Methods* **2012,** *57* (1), 40-46.

48. Banerjee, S.; Wilson, J.; Shim, J.; Shankla, M.; Corbin, E. A.; Aksimentiev, A.; Bashir, R., Slowing DNA transport using graphene–DNA interactions. *Advanced functional materials* **2015,** *25* (6), 936-946.

49. Branton, D.; Deamer, D. W.; Marziali, A.; Bayley, H.; Benner, S. A.; Butler, T.; Di Ventra, M.; Garaj, S.; Hibbs, A.; Huang, X., The potential and challenges of nanopore sequencing. *Nature biotechnology* **2008,** *26* (10), 1146-1153.

50. Vu, T.; Davidson, S.-L.; Borgesi, J.; Maksudul, M.; Jeon, T.-J.; Shim, J., Piecing together the puzzle: nanopore technology in detection and quantification of cancer biomarkers. *RSC advances* **2017,** *7* (68), 42653-42666.

51. Bello, J.; Kim, Y.-R.; Kim, S. M.; Jeon, T.-J.; Shim, J., Lipid bilayer membrane technologies: A review on single-molecule studies of DNA sequencing by using membrane nanopores. *Microchimica Acta* **2017**, 1-15.

52. Song, L.; Hobaugh, M. R.; Shustak, C.; Cheley, S.; Bayley, H.; Gouaux, J. E., Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore. *Science* **1996,** *274* (5294), 1859-1865.

53. Shim, J. W.; Tan, Q.; Gu, L.-Q., Single-molecule detection of folding and unfolding of the G-quadruplex aptamer in a nanopore nanocavity. *Nucleic acids research* **2008,** *37* (3), 972-982.

54. Ding, Y.; Fleming, A. M.; He, L.; Burrows, C. J., Unfolding kinetics of the human telomere i-motif under a 10 pN force imposed by the α-hemolysin nanopore identify transient folded-state lifetimes at physiological pH. *Journal of the American Chemical Society* **2015,** *137* (28), 9053-9060.

55. Guéron, M.; Leroy, J.-L., The i-motif in nucleic acids. *Current opinion in structural biology* **2000,** *10* (3), 326-331.

56. Shim, J.; Kim, Y.; Humphreys, G. I.; Nardulli, A. M.; Kosari, F.; Vasmatzis, G.; Taylor, W. R.; Ahlquist, D. A.; Myong, S.; Bashir, R., Nanopore-based assay for detection of methylation in double-stranded DNA fragments. *Acs Nano* **2015,** *9* (1), 290-300.

57. Gilboa, T.; Torfstein, C.; Juhasz, M.; Grunwald, A.; Ebenstein, Y.; Weinhold, E.; Meller, A., Single-Molecule DNA Methylation Quantification Using Electro-optical Sensing in Solid-State Nanopores. *ACS nano* **2016,** *10* (9), 8861-8870.

58. Nakane, J. J.; Akeson, M.; Marziali, A., Nanopore sensors for nucleic acid analysis. *Journal of Physics: Condensed Matter* **2003,** *15* (32), R1365.

59. Rhee, M.; Burns, M. A., Nanopore sequencing technology: research trends and applications. *Trends in biotechnology* **2006,** *24* (12), 580-586.

60. Wanunu, M., Nanopores: A journey towards DNA sequencing. *Physics of Life Reviews* **2012,** *9* (2), 125-158.

61. Maitra, R. D.; Kim, J.; Dunbar, W. B., Recent advances in nanopore sequencing. *Electrophoresis* **2012,** *33* (23), 3418-3428.

62. Namazie, A.; Alavi, S.; Olopade, O. I.; Pauletti, G.; Aghamohammadi, N.; Aghamohammadi, M.; Gornbein, J. A.; Calcaterra, T. C.; Slamon, D. J.; Wang, M. B., Cyclin D1 amplification and p16 (MTS1/CDK4I) deletion correlate with poor prognosis in head and neck tumors. *The Laryngoscope* **2002,** *112* (3), 472-481.

63. Albertson, D. G.; Collins, C.; McCormick, F.; Gray, J. W., Chromosome aberrations in solid tumors. *Nature genetics* **2003,** *34* (4), 369-376.

64. Nowell PC, H. D., Chromosome studies on normal and leukemic human leukocytes. *J Natl Cancer Inst* **1960,** *25*, 85-109.

65. Dryja TP, R. J., Joyce JM, Petersen RA., Molecular detection of deletions involving band q14 of chromosome 13 in retinoblastomas. *Proc Natl Acad Sci U S A* **1986,** *83*, 7391-4.

66. Slamon DJ, C. G., Wong SG, Levin WJ, Ullrich A, McGuire WL., Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **1987,** *82*, 177-182.

67. Norris, A. e. a., Nanopore sequencing detects structural variants in cancer. *Cancer Biology & Therapy* **2016,** *17*, 246-253.

68. Jones S, Z. X., Parsons DW, Lin JC, Leary RJ, Angenendt P, Man-koo P, Carter H, Kamiyama H, Jimeno A, et al. , Core signaling path-ways in human pancreatic cancers revealed by global genomicanalyses. *Science* **2008**, 1801-6.

69. Biankin AV, W. N., Kassahn KS, Gingras MC, Muthuswamy LB, Johns AL, Miller DK, Wilson PJ, Patch AM, Wu J, et al, Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **2012,** *491*, 399-405.

70. Norris AL, K. H., Makohon-Moore A, Pallavajjala A, Morsberger LA, Lee K, Batista D, Iacobuzio-Donahue CA, Lin MT, Klein AP, et al, TransFlip mutations produce deletions in pancreatic cancer. *Genes, Chromosomes Cancer* **2015,** *54*, 472-481.

71. Diehl F, S. K., Choti MA, Romans K, Goodman S, Li M, Thornton K, Agrawal N, Sokoll L, Szabo SA, et al., Circulating mutant DNA to assess tumor dynamics. *Nat Med* **2008,** *14* (985-990).

72. Ritz, A.; Bashir, A.; Sindi, S.; Hsu, D.; Hajirasouliha, I.; Raphael, B. J., Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics* **2014,** *30* (24), 3458-3466.

73. Medvedev, P.; Stanciu, M.; Brudno, M., Computational methods for discovering structural variation with next-generation sequencing. *Nature methods* **2009,** *6*, S13-S20.

74. Ewing, B.; Green, P., Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* **1998,** *8* (3), 186-194.

75. Alkan, C.; Coe, B. P.; Eichler, E. E., Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **2011,** *12* (5), 363-376.

76. Darnell, J. E., Transcription factors as targets for cancer therapy. *Nature Reviews Cancer* **2002,** *2* (10), 740-749.

77. Blume-Jensen, P.; Hunter, T., Oncogenic kinase signalling. *Nature* **2001,** *411* (6835), 355-365.

78. Squires, A. e. a., Nanopore sensing of individual transcription factors bound to DNA. *Sci. Rep* **2015**.

79. Strong, L. C.; Riccardi, V. M.; Ferrell, R. E.; Sparkes, R. S., Familial retinoblastoma and chromosome 13 deletion transmitted via an insertional translocation. *Science* **1981,** *213* (4515), 1501-1503.

80. Konopka, J. B.; Watanabe, S. M.; Singer, J. W.; Collins, S. J.; Witte, O. N., Cell lines and clinical isolates derived from Ph1-positive chronic myelogenous leukemia patients express c-abl proteins with a common structural alteration. *Proceedings of the National Academy of Sciences* **1985,** *82* (6), 1810-1814.

81. Tsujimoto, Y.; Gorham, J.; Cossman, J.; Jaffe, E.; Croce, C. M., The t (14; 18) chromosome translocations involved in B-cell neoplasms result from mistakes in VDJ joining. *Science* **1985,** *229* (4720), 1390-1393.

82. Catlett-Falcone, R.; Landowski, T. H.; Oshiro, M. M.; Turkson, J.; Levitzki, A.; Savino, R.; Ciliberto, G.; Moscinski, L.; Fernández-Luna, J. L.; Nuñez, G., Constitutive activation of Stat3 signaling confers resistance to apoptosis in human U266 myeloma cells. *Immunity* **1999,** *10* (1), 105-115.

83. Bowman, T.; Garcia, R.; Turkson, J.; Jove, R., STATs in oncogenesis. *Oncogene* **2000,** *19* (21).

84. Coffer, P. J.; Koenderman, L.; de Groot, R. P., The role of STATs in myeloid differentiation and leukemia. *Oncogene* **2000,** *19* (21), 2511-2522.

85. Song, J. I.; Grandis, J. R., STAT signaling in head and neck cancer. *Oncogene* **2000,** *19* (21).

86. Bromberg, J., Stat proteins and oncogenesis. *The Journal of clinical investigation* **2002,** *109* (9), 1139-1142.

87. Mijušković, M.; Brown, S. M.; Tang, Z.; Lindsay, C. R.; Efstathiadis, E.; Deriano, L.; Roth, D. B., A streamlined method for detecting structural variants in cancer genomes by short read paired-end sequencing. *PloS one* **2012,** *7* (10), e48314.

88. Garner, M. M.; Revzin, A., A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic acids research* **1981,** *9* (13), 3047-3060.

89. Frederick, C. A.; Grable, J.; Melia, M.; Samudzi, C.; Jen-Jacobson, L.; Bi-Cheng, W.; Greene, P.; Boyer, H. W.; Rosenberg, J. M., Kinked DNA in crystalline complex with EcoRI endonuclease. *Nature* **1984,** *309* (5966), 327-331.

90. Pavletich, N. P.; Pabo, C. O., Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **1991,** *252* (5007), 809-817.

91. Johnson, D. S.; Mortazavi, A.; Myers, R. M.; Wold, B., Genome-wide mapping of in vivo protein-DNA interactions. *Science* **2007,** *316* (5830), 1497-1502.

92. Takayama, Y.; Sahu, D.; Iwahara, J., NMR studies of translocation of the Zif268 protein between its target DNA Sites. *Biochemistry* **2010,** *49* (37), 7998-8005.

93. Spitz, F. F., E. E. M, Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **2012,** *13*.

94. Bettegowda C, S. M., Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM, et al, Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* **2014,** *6*.

95. Brouilette, S.; Singh, R. K.; Thompson, J. R.; Goodall, A. H.; Samani, N. J., White cell telomere length and risk of premature myocardial infarction. *Arteriosclerosis, thrombosis, and vascular biology* **2003,** *23* (5), 842-846.

96. Maubaret, C. G.; Salpea, K. D.; Jain, A.; Cooper, J. A.; Hamsten, A.; Sanders, J.; Montgomery, H.; Neil, A.; Nair, D.; Humphries, S. E., Telomeres are shorter in myocardial infarction patients compared to healthy subjects: correlation with environmental risk factors. *Journal of molecular medicine* **2010,** *88* (8), 785-794.

97. Weischer, M.; Bojesen, S. E.; Cawthon, R. M.; Freiberg, J. J.; Tybjærg-Hansen, A.; Nordestgaard, B. G., Short telomere length, myocardial infarction, ischemic heart disease, and early death. *Arteriosclerosis, thrombosis, and vascular biology* **2012,** *32* (3), 822-829.

98. Weischer, M.; Nordestgaard, B. G.; Cawthon, R. M.; Freiberg, J. J.; Tybjærg-Hansen, A.; Bojesen, S. E., Short telomere length, cancer survival, and cancer risk in 47102 individuals. *Journal of the National Cancer Institute* **2013**, djt016.

99. Hande, M. P.; Samper, E.; Lansdorp, P.; Blasco, M. A., Telomere length dynamics and chromosomal instability in cells derived from telomerase null mice. *The Journal of cell biology* **1999,** *144* (4), 589-601.

100. Hahn, W. C.; Stewart, S. A.; Brooks, M. W.; York, S. G.; Eaton, E.; Kurachi, A.; Beijersbergen, R. L.; Knoll, J. H.; Meyerson, M.; Weinberg, R. A., Inhibition of telomerase limits the growth of human cancer cells. *Nature medicine* **1999,** *5* (10), 1164-1170.

101. Morales, C. P.; Holt, S. E.; Ouellette, M.; Kaur, K. J.; Yan, Y.; Wilson, K. S.; White, M. A.; Wright, W. E.; Shay, J. W., Absence of cancer–associated changes in human fibroblasts immortalized with telomerase. *Nature genetics* **1999,** *21* (1), 115-118.

102. Cawthon, R. M., Telomere measurement by quantitative PCR. *Nucleic acids research* **2002,** *30* (10), e47-e47.

103. Shim, J. W.; Gu, L.-Q., Encapsulating a single G-quadruplex aptamer in a protein nanocavity. *The journal of physical chemistry. B* **2008,** *112* (28), 8354.

104. Shim, J. W.; Tan, Q.; Gu, L.-Q., Single-molecule detection of folding and unfolding of the G-quadruplex aptamer in a nanopore nanocavity. *Nucleic acids research* **2009,** *37* (3), 972-982.

105. An, N.; Fleming, A. M.; Middleton, E. G.; Burrows, C. J., Single-molecule investigation of G-quadruplex folds of the human telomere sequence in a protein nanocavity. *Proceedings of the National Academy of Sciences* **2014,** *111* (40), 14325-14331.

106. Robertson, J. W. F.; Kasianowicz, J. J.; Banerjee, S., Analytical Approaches for Studying Transporters, Channels and Porins. *Chemical Reviews* **2012,** *112* (12), 6227-6249.

107. Jung, Y. C., S.; Braha, O.; Bayley, H. , The internal cavity of the staphylococcal a-Hemolysin pore accommodates approximately 175 exogenous amino acid residues. *Biochemistry* **2005,** *44* (25), 8918-8929.

108. Brooks, T. A.; Kendrick, S.; Hurley, L., Making sense of G-quadruplex and i-motif functions in oncogene promoters. *FEBS Journal* **2010,** *277* (17), 3459-3469.

109. Liang, G.; Qureshi, A. A.; Guo, Q.; De Vivo, I.; Han, J., No association between telomere length in peripheral blood leukocytes and the risk of nonmelanoma skin cancer. *Cancer Epidemiology Biomarkers & Prevention* **2011,** *20* (5), 1043-1045.

110. Liu, Z.; Ma, H.; Wei, S.; Li, G.; Sturgis, E. M.; Wei, Q., Telomere length and TERT functional polymorphisms are not associated with risk of squamous cell carcinoma of the head and neck. *Cancer Epidemiology Biomarkers & Prevention* **2011,** *20* (12), 2642-2645.

111. Hou, L.; Zhang, X.; Gawron, A. J.; Liu, J., Surrogate tissue telomere length and cancer risk: shorter or longer? *Cancer letters* **2012,** *319* (2), 130-135.

112. Hanahan, D.; Weinberg, R. A., The hallmarks of cancer. *cell* **2000,** *100* (1), 57-70.

113. Haffner, M. C.; Chaux, A.; Meeker, A. K.; Esopi, D.; Gerber, J.; Pellakuru, L. G.; Toubaji, A.; Argani, P.; Iacobuzio-Donahue, C.; Nelson, W. G., Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. *Oncotarget* **2011,** *2* (8), 627-637.

114. Holmfeldt, L.; Mullighan, C. G., The role of TET2 in hematologic neoplasms. *Cancer cell* **2011,** *20* (1), 1-2.

115. Song, C.-X.; Szulwach, K. E.; Fu, Y.; Dai, Q.; Yi, C.; Li, X.; Li, Y.; Chen, C.-H.; Zhang, W.; Jian, X., Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nature biotechnology* **2011,** *29* (1), 68-72.

116. Yang, H.; Liu, Y.; Bai, F.; Zhang, J.; Ma, S.; Liu, J.; Xu, Z.; Zhu, H.; Ling, Z.; Ye, D., Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene* **2013,** *32* (5), 663-669.

117. Bacolla, A.; Cooper, D. N.; Vasquez, K. M., Mechanisms of base substitution mutagenesis in cancer genomes. *Genes* **2014,** *5* (1), 108-146.

118. Esteller, M.; Corn, P. G.; Baylin, S. B.; Herman, J. G., A gene hypermethylation profile of human cancer. *Cancer research* **2001,** *61* (8), 3225-3229.

119. Strathdee, G.; Brown, R., Aberrant DNA methylation in cancer: potential clinical interventions. *Expert reviews in molecular medicine* **2002,** *4* (04), 1-17.

120. Laird, P. W., The power and the promise of DNA methylation markers. *Nature Reviews Cancer* **2003,** *3* (4), 253-266.

121. Vanaja, D. K.; Ehrich, M.; Van den Boom, D.; Cheville, J. C.; Karnes, R. J.; Tindall, D. J.; Cantor, C. R.; Young, C. Y., Hypermethylation of genes for diagnosis and risk stratification of prostate cancer. *Cancer investigation* **2009,** *27* (5), 549-560.

122. Nakamura, N.; Takenaga, K., Hypomethylation of the metastasis-associated S100A4 gene correlates with gene activation in human colon adenocarcinoma cell lines. *Clinical & experimental metastasis* **1998,** *16* (5), 471-479.

123. Adrianadecapoa, A. M.; Della Rosa, S.; Caiafa, P.; Mariani, L.; Del Nonno, F.; Vocaturo, A.; Donnorso, R. A. P.; Niveleau, A.; Grappelli, C., DNA demethylation is directly related to tumour progression: evidence in normal, pre-malignant and malignant cells from uterine cervix samples. *Oncology reports* **2003,** *10*, 545-549.

124. Akiyama, Y.; Maesawa, C.; Ogasawara, S.; Terashima, M.; Masuda, T., Cell-type-specific repression of the maspin gene is disrupted frequently by demethylation at the promoter region in gastric intestinal metaplasia and cancer cells. *The American journal of pathology* **2003,** *163* (5), 1911-1919.

125. Iacobuzio-Donahue, C. A.; Maitra, A.; Olsen, M.; Lowe, A. W.; Van Heek, N. T.; Rosty, C.; Walter, K.; Sato, N.; Parker, A.; Ashfaq, R., Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *The American journal of pathology* **2003,** *162* (4), 1151-1162.

126. Oshimo, Y.; Nakayama, H.; Ito, R.; Kitadai, Y.; Yoshida, K.; Chayama, K.; Yasui, W., Promoter methylation of cyclin D2 gene in gastric carcinoma. *International journal of oncology* **2003,** *23* (6), 1663-1670.

127. Sato, N.; Maitra, A.; Fukushima, N.; van Heek, N. T.; Matsubayashi, H.; Iacobuzio-Donahue, C. A.; Rosty, C.; Goggins, M., Frequent hypomethylation of multiple genes overexpressed in pancreatic ductal adenocarcinoma. *Cancer Research* **2003,** *63* (14), 4158-4166.

128. Susan, J. C.; Harrison, J.; Paul, C. L.; Frommer, M., High sensitivity mapping of methylated cytosines. *Nucleic acids research* **1994,** *22* (15), 2990-2997.

129. Herman, J. G.; Graff, J. R.; Myöhänen, S.; Nelkin, B. D.; Baylin, S. B., Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proceedings of the National Academy of Sciences* **1996,** *93* (18), 9821-9826.

130. Eads, C. A.; Danenberg, K. D.; Kawakami, K.; Saltz, L. B.; Blake, C.; Shibata, D.; Danenberg, P. V.; Laird, P. W., MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic acids research* **2000,** *28* (8), e32-00.

131. Fraga, M. F.; Esteller, M., DNA methylation: a profile of methods and applications. *Biotechniques* **2002,** *33* (3), 632-649.

132. Weber, M.; Davies, J. J.; Wittig, D.; Oakeley, E. J.; Haase, M.; Lam, W. L.; Schuebeler, D., Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics* **2005,** *37* (8), 853-862.

133. Keshet, I.; Schlesinger, Y.; Farkash, S.; Rand, E.; Hecht, M.; Segal, E.; Pikarski, E.; Young, R. A.; Niveleau, A.; Cedar, H., Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nature genetics* **2006,** *38* (2), 149-153.

134. Khulan, B.; Thompson, R. F.; Ye, K.; Fazzari, M. J.; Suzuki, M.; Stasiek, E.; Figueroa, M. E.; Glass, J. L.; Chen, Q.; Montagna, C., Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome research* **2006,** *16* (8), 1046-1055.

135. Shim, J.; Rivera, J. A.; Bashir, R., Electron beam induced local crystallization of HfO 2 nanopores for biosensing applications. *Nanoscale* **2013,** *5* (22), 10887-10893.

136. Tahiliani, M.; Koh, K. P.; Shen, Y.; Pastor, W. A.; Bandukwala, H.; Brudno, Y.; Agarwal, S.; Iyer, L. M.; Liu, D. R.; Aravind, L., Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **2009,** *324* (5929), 930-935.

137. Ito, S.; D'Alessio, A. C.; Taranova, O. V.; Hong, K.; Sowers, L. C.; Zhang, Y., Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **2010,** *466* (7310), 1129-1133.

138. Zhang, H.; Zhang, X.; Clark, E.; Mulcahey, M.; Huang, S.; Shi, Y. G., TET1 is a DNA-binding protein that modulates DNA methylation and gene transcription via hydroxylation of 5-methylcytosine. *Cell research* **2010,** *20* (12), 1390-1393.

139. Wu, H.; Zhang, Y., Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes & development* **2011,** *25* (23), 2436-2452.

140. Szwagierczak, A.; Bultmann, S.; Schmidt, C. S.; Spada, F.; Leonhardt, H., Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic acids research* **2010,** *38* (19), e181-e181.

141. Wossidlo, M.; Nakamura, T.; Lepikhov, K.; Marques, C. J.; Zakhartchenko, V.; Boiani, M.; Arand, J.; Nakano, T.; Reik, W.; Walter, J., 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature communications* **2011,** *2,* 241.

142. Booth, M. J.; Branco, M. R.; Ficz, G.; Oxley, D.; Krueger, F.; Reik, W.; Balasubramanian, S., Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **2012,** *336* (6083), 934-937.

143. Shankar, T. S.; Willems, L., Epigenetic modulators mitigate angiogenesis through a complex transcriptomic network. *Vascular pharmacology* **2014,** *60* (2), 57-66.

144. Fu, Y.; He, C., Nucleic acid modifications with epigenetic significance. *Current opinion in chemical biology* **2012,** *16* (5), 516-524.

145. Lian, C. G.; Xu, Y.; Ceol, C.; Wu, F.; Larson, A.; Dresser, K.; Xu, W.; Tan, L.; Hu, Y.; Zhan, Q., Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. *Cell* **2012,** *150* (6), 1135-1146.

146. Yu, M.; Hon, G. C.; Szulwach, K. E.; Song, C.-X.; Zhang, L.; Kim, A.; Li, X.; Dai, Q.; Shen, Y.; Park, B., Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **2012,** *149* (6), 1368-1380.

147. Wanunu, M.; Cohen-Karni, D.; Johnson, R. R.; Fields, L.; Benner, J.; Peterman, N.; Zheng, Y.; Klein, M. L.; Drndic, M., Discrimination of Methylcytosine from Hydroxymethylcytosine in DNA Molecules. *Journal of the American Chemical Society* **2011,** *133* (3), 486-492.

148. Laszlo, A. H.; Derrington, I. M.; Brinkerhoff, H.; Langford, K. W.; Nova, I. C.; Samson, J. M.; Bartlett, J. J.; Pavlenok, M.; Gundlach, J. H., Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences* **2013,** *110* (47), 18904-18909.

149. Wescoe, Z. L.; Schreiber, J.; Akeson, M., Nanopores discriminate among five C5-cytosine variants in DNA. *Journal of the American Chemical Society* **2014,** *136* (47), 16582-16587.

150. Nguyen, C. T.; Gonzales, F. A.; Jones, P. A., Altered chromatin structure associated with methylation-induced gene silencing in cancer cells: correlation of accessibility, methylation, MeCP2 binding and acetylation. *Nucleic acids research* **2001,** *29* (22), 4598-4606.

151. Fahrner, J. A.; Eguchi, S.; Herman, J. G.; Baylin, S. B., Dependence of histone modifications and gene expression on DNA hypermethylation in cancer. *Cancer research* **2002,** *62* (24), 7213-7218.

152. Ballestar, E.; Paz, M. F.; Valle, L.; Wei, S.; Fraga, M. F.; Espada, J.; Cigudosa, J. C.; Huang, T. H. M.; Esteller, M., Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. *The EMBO journal* **2003,** *22* (23), 6335-6345.

153. Fraga, M. F.; Ballestar, E.; Villar-Garea, A.; Boix-Chornet, M.; Espada, J.; Schotta, G.; Bonaldi, T.; Haydon, C.; Ropero, S.; Petrie, K., Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature genetics* **2005,** *37* (4), 391-400.

154. Esteller, M., Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics* **2007,** *8* (4), 286-298.

155. Wang, Y.; Fischle, W.; Cheung, W.; Jacobs, S.; Khorasanizadeh, S.; Allis, C. D. In *Beyond the double helix: writing and reading the histone code*, Novartis Foundation Symposium, Chichester; New York; John Wiley; 1999: 2004; pp 3-21.

156. Dobosy, J.; Selker, E., Emerging connections between DNA methylation and histone acetylation. *Cellular and Molecular Life Sciences CMLS* **2001,** *58* (5-6), 721-727.

157. Wade, P. A., Methyl CpG-binding proteins and transcriptional repression. *Bioessays* **2001,** *23* (12), 1131-1137.

158. Soni, G. V.; Dekker, C., Detection of nucleosomal substructures using solid-state nanopores. *Nano letters* **2012,** *12* (6), 3180-3186.

159. Andrey Ivankin, S. C., Fast, Label-Free Force Spectroscopy of Histone-DNA Interactions in Individual Nucleosomes Using Nanopores. *Journal of the American Chemical Society* **2016**, 15350-15352.

160. Langecker, M.; Ivankin, A.; Carson, S.; Kinney, S. R.; Simmel, F. C.; Wanunu, M., Nanopores Suggest a Negligible Influence of CpG Methylation on Nucleosome Packaging and Stability. *Nano letters* **2014,** *15* (1), 783-790.

161. Wanunu, M.; Sutin, J.; Meller, A., DNA profiling using solid-state nanopores: detection of DNA-binding molecules. *Nano letters* **2009,** *9* (10), 3498-3502.

162. Ivankin, A.; Carson, S.; Kinney, S. R.; Wanunu, M., Fast, Label-Free Force Spectroscopy of Histone–DNA Interactions in Individual Nucleosomes Using Nanopores. *Journal of the American Chemical Society* **2013,** *135* (41), 15350-15352.

163. Suzuki, H.; Gabrielson, E.; Chen, W.; Anbazhagan, R.; van Engeland, M.; Weijenberg, M. P.; Herman, J. G.; Baylin, S. B., A genomic screen for genes upregulated by demethylation and histone deacetylase inhibition in human colorectal cancer. *Nature genetics* **2002,** *31* (2), 141-149.

164. Yamashita, K.; Upadhyay, S.; Osada, M.; Hoque, M. O.; Xiao, Y.; Mori, M.; Sato, F.; Meltzer, S. J.; Sidransky, D., Pharmacologic unmasking of epigenetically silenced tumor suppressor genes in esophageal squamous cell carcinoma. *Cancer cell* **2002,** *2* (6), 485-495.

165. Carthew, R. W.; Sontheimer, E. J., Origins and Mechanisms of miRNAs and siRNAs. *Cell* **2009,** *136* (4), 642-55.

166. Winter, J.; Jung, S.; Keller, S.; Gregory, R. I.; Diederichs, S., Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature cell biology* **2009,** *11* (3), 228-234.

167. Jansson, M. D.; Lund, A. H., MicroRNA and cancer. *Molecular oncology* **2012,** *6* (6), 590-610.

168. Chen, C.; Ridzon, D. A.; Broomer, A. J.; Zhou, Z.; Lee, D. H.; Nguyen, J. T.; Barbisin, M.; Xu, N. L.; Mahuvakar, V. R.; Andersen, M. R., Real-time quantification of microRNAs by stem–loop RT–PCR. *Nucleic acids research* **2005,** *33* (20), e179-e179.

169. Hunt, E. A.; Goulding, A. M.; Deo, S. K., Direct detection and quantification of microRNAs. *Analytical biochemistry* **2009,** *387* (1), 1.

170. Li, W.; Ruan, K., MicroRNA detection by microarray. *Anal Bioanal Chem* **2009,** *394* (4), 1117-24.

171. Yendamuri, S.; Kratzke, R., MicroRNA biomarkers in lung cancer: MiRacle or quagMiRe? *Translational Research* **2011,** *157* (4), 209-215.

172. Wang, Y.; Zheng, D.; Tan, Q.; Wang, M. X.; Gu, L. Q., Nanopore-based detection of circulating microRNAs in lung cancer patients. *Nat Nanotechnol* **2011,** *6* (10), 668-74.

173. Silvestri, G. A.; Alberg, A. J.; Ravenel, J., The changing epidemiology of lung cancer with a focus on screening. *Bmj* **2009,** *339* (6), 30-53.

174. Yanaihara, N.; Caplen, N.; Bowman, E.; Seike, M.; Kumamoto, K.; Yi, M.; Stephens, R. M.; Okamoto, A.; Yokota, J.; Tanaka, T., Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer cell* **2006,** *9* (3), 189-198.

175. Patnaik, S. K.; Kannisto, E.; Knudsen, S.; Yendamuri, S., Evaluation of microRNA expression profiles that may predict recurrence of localized stage I non–small cell lung cancer after surgical resection. *Cancer research* **2010,** *70* (1), 36-45.

176. Wanunu, M.; Dadosh, T.; Ray, V.; Jin, J.; McReynolds, L.; Drndić, M., Rapid electronic detection of probe-specific microRNAs using thin nanopore sensors. *Nature nanotechnology* **2010,** *5* (11), 807-814.

177. Jin, S. G.; Kadam, S.; Pfeifer, G. P., Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res* **2010,** *38* (11), e125.

178. Estevez-Torres, A.; Baigl, D., DNA compaction: fundamentals and applications. *Soft Matter* **2011,** *7* (15), 6746-6756.

179. Bloomfield, V. A., Condensation of DNA by multivalent cations: considerations on mechanism. *Biopolymers* **1991,** *31* (13), 1471-1481.

180. Zinchenko, A. A.; Yoshikawa, K., Na+ shows a markedly higher potential than K+ in DNA compaction in a crowded environment. *Biophysical journal* **2005,** *88* (6), 4118-4123.

181. Yoshikawa, K.; Takahashi, M.; Vasilevskaya, V.; Khokhlov, A., Large discrete transition in a single DNA molecule appears continuous in the ensemble. *Physical review letters* **1996,** *76* (16), 3029.

182. Akitaya, T.; Seno, A.; Nakai, T.; Hazemoto, N.; Murata, S.; Yoshikawa, K., Weak interaction induces an ON/OFF switch, whereas strong interaction causes gradual change: folding transition of a long duplex DNA chain by poly-L-lysine. *Biomacromolecules* **2007,** *8* (1), 273-278.

183. Vasilevskaya, V.; Khokhlov, A.; Matsuzawa, Y.; Yoshikawa, K., Collapse of single DNA molecule in poly (ethylene glycol) solutions. *The Journal of chemical physics* **1995,** *102* (16), 6595-6602.

184. Zinchenko, A. A.; Yoshikawa, K.; Baigl, D., Compaction of single-chain DNA by histone-inspired nanoparticles. *Physical review letters* **2005,** *95* (22), 228101.

185. Zinchenko, A. A.; Sakaue, T.; Araki, S.; Yoshikawa, K.; Baigl, D., Single-chain compaction of long duplex DNA by cationic nanoparticles: modes of interaction and comparison with chromatin. *The Journal of Physical Chemistry B* **2007,** *111* (11), 3019-3031.

186. Mel'nikov, S. M.; Sergeyev, V. G.; Yoshikawa, K., Discrete coil-globule transition of large DNA induced by cationic surfactant. *Journal of the American Chemical Society* **1995,** *117* (9), 2401-2408.

187. Sollogoub, M.; Guieu, S.; Geoffroy, M.; Yamada, A.; Estévez-Torres, A.; Yoshikawa, K.; Baigl, D., Photocontrol of Single-Chain DNA Conformation in Cell-Mimicking Microcompartments. *ChemBioChem* **2008,** *9* (8), 1201-1206.

188. Dias, R.; Mel'nikov, S.; Lindman, B.; Miguel, M. G., DNA phase behavior in the presence of oppositely charged surfactants. *Langmuir* **2000,** *16* (24), 9577-9583.

189. Livolant, F.; Leforestier, A., Condensed phases of DNA: structures and phase transitions. *Progress in Polymer Science* **1996,** *21* (6), 1115-1164.

190. Robinson, C., Liquid-crystalline structures in polypeptide solutions. *Tetrahedron* **1961,** *13* (1-3), 219-234.

191. Robinson, C., The cholesteric phase in polypeptide solutions and biological structures. *Molecular Crystals and Liquid Crystals* **1966,** *1* (4), 467-494.

192. Luzzati, V.; Nicolaieff, A., Etude par diffusion des rayons X aux petits angles des gels d'acide désoxyribonucléique et de nucléoprotéines:(note préliminaire). *Journal of Molecular Biology* **1959,** *1* (2), 127IN4-133IN5.

193. Jmdan, C.; Lerman, L.; Venable, J., Structure and circular dichroism of DNA in concentrated polymer solutions. *Nature* **1972,** *236* (64), 67-70.

194. Zimmerman, S. B., Macromolecular crowding effects on macromolecular interactions: some implications for genome structure and function. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression* **1993,** *1216* (2), 175-185.

195. Maniatis, T.; Venable, J. H.; Lerman, L. S., The structure of Ψ DNA. *Journal of molecular biology* **1974,** *84* (1), 37IN157-5664.

196. Shim, J. W.; Gu, L.-Q., Encapsulating a single G-quadruplex aptamer in a protein nanocavity. *The Journal of Physical Chemistry B* **2008,** *112* (28), 8354-8360.

197. Gouaux, E., α-Hemolysin fromStaphylococcus aureus: an archetype of β-barrel, channel-forming toxins. *Journal of structural biology* **1998,** *121* (2), 110-122.

198. Braha, O.; Walker, B.; Cheley, S.; Kasianowicz, J. J.; Song, L.; Gouaux, J. E.; Bayley, H., Designed protein pores as components for biosensors. *Chemistry & biology* **1997,** *4* (7), 497-505.

199. Chen, C.; Li, M.; Xing, Y.; Li, Y.; Joedecke, C.-C.; Jin, J.; Yang, Z.; Liu, D., Study of pH-induced folding and unfolding kinetics of the DNA i-motif by stopped-flow circular dichroism. *Langmuir* **2012,** *28* (51), 17743-17748.

200. Kypr, J.; Kejnovská, I.; Renčiuk, D.; Vorlíčková, M., Circular dichroism and conformational polymorphism of DNA. *Nucleic acids research* **2009,** *37* (6), 1713-1725.

201. Choi, J.; Kim, S.; Tachikawa, T.; Fujitsuka, M.; Majima, T., pH-induced intramolecular folding dynamics of i-motif DNA. *Journal of the American Chemical Society* **2011,** *133* (40), 16146-16153.

202. Vorlíčková, M.; Kejnovská, I.; Bednářová, K.; Renčiuk, D.; Kypr, J., Circular dichroism spectroscopy of DNA: from duplexes to quadruplexes. *Chirality* **2012,** *24* (9), 691-698.

203. Plesa, C.; Verschueren, D.; Pud, S.; van der Torre, J.; Ruitenberg, J. W.; Witteveen, M. J.; Jonsson, M. P.; Grosberg, A. Y.; Rabin, Y.; Dekker, C., Direct observation of DNA knots using a solid-state nanopore. *Nature nanotechnology* **2016,** *11* (12), 1093-1097.

204. Phan, A. T.; Guéron, M.; Leroy, J.-L., The solution structure and internal motions of a fragment of the cytidine-rich strand of the human telomere. *Journal of molecular biology* **2000,** *299* (1), 123-144.

205. Zhao, Y.; Zeng, Z. x.; Kan, Z. y.; Hao, Y. h.; Tan, Z., The Folding and Unfolding Kinetics of the i-Motif Structure Formed by the C-Rich Strand of Human Telomere DNA. *ChemBioChem* **2005,** *6* (11), 1957-1960.

206. Manzini, G.; Yathindra, N.; Xodo, L., Evidence for intramolecularly folded i-DNA structures in biologically relevant CCC-repeat sequences. *Nucleic acids research* **1994,** *22* (22), 4634-4640.

207. Savelyev, A.; Papoian, G. A., Electrostatic, steric, and hydration interactions favor Na+ condensation around DNA compared with K+. *Journal of the American Chemical Society* **2006,** *128* (45), 14506-14518.

208. Rouzina, I.; Bloomfield, V. A., Macroion attraction due to electrostatic correlation between screening counterions. 1. Mobile surface-adsorbed ions and diffuse ion cloud. *The Journal of Physical Chemistry* **1996,** *100* (23), 9977-9989.

209. Rodgers, M.; Armentrout, P., Noncovalent interactions of nucleic acid bases (uracil, thymine, and adenine) with alkali metal ions. Threshold collision-induced dissociation and theoretical studies. *Journal of the American Chemical Society* **2000,** *122* (35), 8548-8558.

210. Burda, J. V.; Šponer, J.; Hobza, P., Ab initio study of the interaction of guanine and adenine with various mono-and bivalent metal cations (Li+, Na+, K+, Rb+, Cs+; Cu+, Ag+, Au+; Mg2+, Ca2+, Sr2+, Ba2+; Zn2+, Cd2+, and Hg2+). *The Journal of Physical Chemistry* **1996,** *100* (17), 7250-7255.

211. Cerda, B. A.; Wesdemiotis, C., Li+, Na+, and K+ binding to the DNA and RNA nucleobases. Bond energies and attachment sites from the dissociation of metal ion-bound heterodimers. *Journal of the American Chemical Society* **1996,** *118* (47), 11884-11892.

212. Hoyau, S.; Norrman, K.; McMahon, T.; Ohanessian, G., A quantitative basis for a scale of Na+ affinities of organic and small biological molecules in the gas phase. *Journal of the American Chemical Society* **1999,** *121* (38), 8864-8875.

213. de Martimprey, H.; Vauthier, C.; Malvy, C.; Couvreur, P., Polymer nanocarriers for the delivery of small fragments of nucleic acids: Oligonucleotides and siRNA. *Eur. J. Pharm. Biopharm.* **2009,** *71* (3), 490--504.

214. Merdan, T.; Kopeček, J.; Kissel, T., Prospects for cationic polymers in gene and oligonucleotide therapy against cancer. *Adv. Drug Delivery Rev.* **2002,** *54* (5), 715--758.

215. Vinogradov, S. V.; Batrakova, E. V.; Li, S.; Kabanov, A. V., Mixed Polymer Micelles of Amphiphilic and Cationic Copolymers for Delivery of Antisense Oligonucleotides. *J. Drug Targeting* **2004,** *12* (8), 517--526.

216. Lu, X.; Tran, T.-H.; Jia, F.; Tan, X.; Davis, S.; Krishnan, S.; Amiji, M. M.; Zhang, K., Providing Oligonucleotides with Steric Selectivity by Brush-Polymer-Assisted Compaction. *J. Am. Chem. Soc.* **2015,** *137* (39), 12466--12469.

217. Huppert, J. L.; Balasubramanian, S., Prevalence of quadruplexes in the human genome. *Nucleic acids research* **2005,** *33* (9), 2908-2916.

218. Armas, P.; David, A.; Calcaterra, N. B., Transcriptional control by G-quadruplexes: In vivo roles and perspectives for specific intervention. *Transcription* **2017,** *8* (1), 21-25.

219. Rhodes, D.; Lipps, H. J., G-quadruplexes and their regulatory roles in biology. *Nucleic acids research* **2015,** *43* (18), 8627-8637.

220. David, A. P.; Margarit, E.; Domizi, P.; Banchio, C.; Armas, P.; Calcaterra, N. B., G-quadruplexes as novel cis-elements controlling transcription during embryonic development. *Nucleic acids research* **2016,** *44* (9), 4163-4173.

221. Fleming, A. M.; Zhu, J.; Ding, Y.; Burrows, C. J., 8-Oxo-7, 8-dihydroguanine in the context of a gene promoter G-quadruplex is an on–off switch for transcription. *ACS chemical biology* **2017,** *12* (9), 2417-2426.

222. Neidle, S., Quadruplex nucleic acids as targets for anticancer therapeutics. *Nature Reviews Chemistry* **2017,** *1* (5), 0041.

223. d'Auriol, L.; Mattei, M.-G.; Andre, C.; Galibert, F., Localization of the human c-kit protooncogene on the q11–q12 region of chromosome 4. *Human genetics* **1988,** *78* (4), 374-376.

224. Yarden, Y.; Kuang, W.-J.; Yang-Feng, T.; Coussens, L.; Munemitsu, S.; Dull, T.; Chen, E.; Schlessinger, J.; Francke, U.; Ullrich, A., Human proto-oncogene c-kit: a new cell surface receptor tyrosine kinase for an unidentified ligand. *The EMBO journal* **1987,** *6* (11), 3341-3351.

225. Yamamoto, K.; Tojo, A.; Aoki, N.; Shibuya, M., Characterization of the Promoter Region of the Human c-kit Proto-oncogene. *Japanese journal of cancer research* **1993,** *84* (11), 1136-1144.

226. Kotar, A.; Rigo, R.; Sissi, C.; Plavec, J., Two-quartet kit* G-quadruplex is formed via double-stranded pre-folded structure. *Nucleic acids research* **2018,** *47* (5), 2641-2653.

227. Metcalfe, D. D., Mast cells and mastocytosis. *Blood* **2008,** *112* (4), 946-956.

228. Gregory-Bryson, E.; Bartlett, E.; Kiupel, M.; Hayes, S.; Yuzbasiyan-Gurkan, V., Canine and human gastrointestinal stromal tumors display similar mutations in c-KIT exon 11. *BMC cancer* **2010,** *10* (1), 559.

229. Li, X.; Heyer, W. D., Homologous recombination in DNA repair and DNA damage tolerance. *Cell research* **2008,** *18* (1), 99-113.

230. Yamamoto, K.; Tojo, A.; Aoki, N.; Shibuya, M., Characterization of the promoter region of the human c-kit proto-oncogene. *Japanese journal of cancer research : Gann* **1993,** *84* (11), 1136-44.

231. Balasubramanian, S.; Hurley, L. H.; Neidle, S., Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nature reviews Drug discovery* **2011,** *10* (4), 261.

232. Collie, G. W.; Parkinson, G. N., The application of DNA and RNA G-quadruplexes to therapeutic medicines. *Chemical Society Reviews* **2011,** *40* (12), 5867-5892.

233. Xu, H.; Di Antonio, M.; McKinney, S.; Mathew, V.; Ho, B.; O'Neil, N. J.; Santos, N. D.; Silvester, J.; Wei, V.; Garcia, J.; Kabeer, F.; Lai, D.; Soriano, P.; Banáth, J.; Chiu, D. S.; Yap, D.; Le, D. D.; Ye, F. B.; Zhang, A.; Thu, K.; Soong, J.; Lin, S.-c.; Tsai, A. H. C.; Osako, T.; Algara, T.; Saunders, D. N.; Wong, J.; Xian, J.; Bally, M. B.; Brenton, J. D.; Brown, G. W.; Shah, S. P.; Cescon, D.; Mak, T. W.; Caldas, C.; Stirling, P. C.; Hieter, P.; Balasubramanian, S.; Aparicio, S., CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. *Nature Communications* **2017,** *8*, 14432.

234. Andrews, J. O.; Conway, W.; Cho, W. K.; Narayanan, A.; Spille, J. H.; Jayanth, N.; Inoue, T.; Mullen, S.; Thaler, J.; Cissé, I. I., qSR: a quantitative super-resolution analysis tool reveals the cell-cycle dependent organization of RNA Polymerase I in live human cells. *Scientific Reports* **2018,** *8* (1), 7424.

235. Yoshida, K.; Miki, Y., Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer science* **2004,** *95* (11), 866-71.

236. Shim, J.; Gu, L.-Q., Encapsulating a single G-quadruplex aptamer in a protein nanocavity. *The Journal of Physical Chemistry B* **2008,** *112* (28), 8354-8360.

237. Vercoutere, W.; Winters-Hilt, S.; Olsen, H.; Deamer, D.; Haussler, D.; Akeson, M., Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel. *Nature Biotechnology* **2001,** *19*, 248.

238. Vu, T.; Borgesi, J.; Soyring, J.; D'Alia, M.; Davidson, S.-L.; Shim, J., Employing LiCl salt gradient in the wild-type α-hemolysin nanopore to slow down DNA translocation and detect methylated cytosine. *Nanoscale* **2019,** *11* (21), 10536-10545.

239. Meller, A.; Nivon, L.; Brandin, E.; Golovchenko, J.; Branton, D., Rapid nanopore discrimination between single polynucleotide molecules. *Proceedings of the National Academy of Sciences* **2000,** *97* (3), 1079-1084.

240. Bochman, M. L.; Paeschke, K.; Zakian, V. A., DNA secondary structures: stability and function of G-quadruplex structures. *Nature Reviews Genetics* **2012,** *13* (11), 770.

241. Largy, E.; Mergny, J.-L.; Gabelica, V., Role of alkali metal ions in G-quadruplex nucleic acid structure and stability. In *The Alkali Metal Ions: Their Role for Life*, Springer: 2016; pp 203-258.

242. Misra, V. K.; Draper, D. E., On the role of magnesium ions in RNA stability. *Biopolymers: Original Research on Biomolecules* **1998,** *48* (2-3), 113-135.

243. Gu, J.; Leszczynski, J., Origin of Na+/K+ selectivity of the guanine tetraplexes in water: the theoretical rationale. *The Journal of Physical Chemistry A* **2002,** *106* (3), 529-532.

244. Hud, N. V.; Smith, F. W.; Anet, F. A.; Feigon, J., The selectivity for K+ versus Na+ in DNA quadruplexes is dominated by relative free energies of hydration: a thermodynamic analysis by 1H NMR. *Biochemistry* **1996,** *35* (48), 15383-15390.

245. Gu, L.-Q.; Braha, O.; Conlan, S.; Cheley, S.; Bayley, H., Stochastic sensing of organic analytes by a pore-forming protein containing a molecular adapter. *Nature* **1999,** *398* (6729), 686-690.

246. Ashman, L. K.; Griffith, R., Therapeutic targeting of c-KIT in cancer. *Expert opinion on investigational drugs* **2013,** *22* (1), 103-115.

247. Lennartsson, J.; Rönnstrand, L., Stem cell factor receptor/c-Kit: from basic science to clinical implications. *Physiological reviews* **2012,** *92* (4), 1619-1649.

248. Islam, B.; Stadlbauer, P.; Krepl, M.; Koca, J.; Neidle, S.; Haider, S.; Sponer, J., Extended molecular dynamics of a c-kit promoter quadruplex. *Nucleic acids research* **2015,** *43* (18), 8673-8693.

249. Strick, T.; Bensimon, D.; Croquette, V., Micro-mechanical measurement of the torsional modulus of DNA. In *Structural Biology and Functional Genomics*, Springer: 1999; pp 87-96.

250. Danilowicz, C.; Coljee, V. W.; Bouzigues, C.; Lubensky, D. K.; Nelson, D. R.; Prentiss, M., DNA unzipped under a constant force exhibits multiple metastable intermediates. *Proceedings of the National Academy of Sciences* **2003,** *100* (4), 1694-1699.

251. Huguet, J. M.; Forns, N.; Ritort, F., Statistical properties of metastable intermediates in DNA unzipping. *Physical review letters* **2009,** *103* (24), 248106.

252. Xu, H.; Di Antonio, M.; McKinney, S.; Mathew, V.; Ho, B.; O'Neil, N. J.; Dos Santos, N.; Silvester, J.; Wei, V.; Garcia, J., CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. *Nature communications* **2017,** *8* (1), 1-18.

253. Marathias, V. M.; Bolton, P. H., Structures of the potassium-saturated, 2: 1, and intermediate, 1: 1, forms of a quadruplex DNA. *Nucleic acids research* **2000,** *28* (9), 1969-1977.

254. Latysheva, N. S.; Babu, M. M., Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic acids research* **2016,** *44* (10), 4487-4503.

255. Esteller, M.; Toyota, M.; Sanchez-Cespedes, M.; Capella, G.; Peinado, M. A.; Watkins, D. N.; Issa, J.-P. J.; Sidransky, D.; Baylin, S. B.; Herman, J. G., Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is associated with G to A mutations in K-ras in colorectal tumorigenesis. *Cancer research* **2000,** *60* (9), 2368-2371.

256. Mikeska, T.; Bock, C.; Do, H.; Dobrovic, A., DNA methylation biomarkers in cancer: progress towards clinical implementation. *Expert review of molecular diagnostics* **2012,** *12* (5), 473-487.

257. Fearon, E. R.; Vogelstein, B., A genetic model for colorectal tumorigenesis. *Cell* **1990,** *61* (5), 759-767.

258. Baylln, S. B.; Herman, J. G.; Graff, J. R.; Vertino, P. M.; Issa, J.-P., Alterations in DNA methylation: a fundamental aspect of neoplasia. *Advances in cancer research* **1997,** *72*, 141-196.

259. Herman, J. G.; Baylin, S. B., Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine* **2003,** *349* (21), 2042-2054.

260. Cantara, W. A.; Crain, P. F.; Rozenski, J.; McCloskey, J. A.; Harris, K. A.; Zhang, X.; Vendeix, F. A.; Fabris, D.; Agris, P. F., The RNA modification database, RNAMDB: 2011 update. *Nucleic acids research* **2010,** *39* (suppl_1), D195-D201.

261. Squires, J. E.; Patel, H. R.; Nousch, M.; Sibbritt, T.; Humphreys, D. T.; Parker, B. J.; Suter, C. M.; Preiss, T., Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic acids research* **2012,** *40* (11), 5023-5033.

262. Trixl, L.; Lusser, A., The dynamic RNA modification 5-methylcytosine and its emerging role as an epitranscriptomic mark. *Wiley Interdisciplinary Reviews: RNA* **2019,** *10* (1), e1510.

263. Amort, T.; Rieder, D.; Wille, A.; Khokhlova-Cubberley, D.; Riml, C.; Trixl, L.; Jia, X.-Y.; Micura, R.; Lusser, A., Distinct 5-methylcytosine profiles in poly (A) RNA from mouse embryonic stem cells and brain. *Genome biology* **2017,** *18* (1), 1.

264. Toiyama, Y.; Okugawa, Y.; Tanaka, K.; Araki, T.; Uchida, K.; Hishida, A.; Uchino, M.; Ikeuchi, H.; Hirota, S.; Kusunoki, M., A panel of methylated microRNA biomarkers for identifying high-risk patients with ulcerative colitis-associated colorectal cancer. *Gastroenterology* **2017,** *153* (6), 1634-1646. e8.

265. Miura, F.; Enomoto, Y.; Dairiki, R.; Ito, T., Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic acids research* **2012,** *40* (17), e136-e136.

266. Schatz, M. C., Nanopore sequencing meets epigenetics. *Nature methods* **2017,** *14* (4), 347.

267. Gigante, S.; Gouil, Q.; Lucattini, A.; Keniry, A.; Beck, T.; Tinning, M.; Gordon, L.; Woodruff, C.; Speed, T. P.; Blewitt, M., Using long-read sequencing to detect imprinted DNA methylation. *bioRxiv* **2019**, 445924.

268. Shim, J.; Humphreys, G. I.; Venkatesan, B. M.; Munz, J. M.; Zou, X.; Sathe, C.; Schulten, K.; Kosari, F.; Nardulli, A. M.; Vasmatzis, G.; Bashir, R., Detection and quantification of methylation in DNA using solid-state nanopores. *Scientific reports* **2013,** *3*, 1389.

269. de Zoysa, R. S. S.; Jayawardhana, D. A.; Zhao, Q.; Wang, D.; Armstrong, D. W.; Guan, X., Slowing DNA translocation through nanopores using a solution containing organic salts. *The Journal of Physical Chemistry B* **2009,** *113* (40), 13332-13336.

270. Piguet, F.; Discala, F.; Breton, M.-F.; Pelta, J.; Bacri, L.; Oukhaled, A., Electroosmosis through α-hemolysin that depends on alkali cation type. *J. Phys. Chem. Lett* **2014,** *5* (24), 4362-4367.

271. Kowalczyk, S. W.; Wells, D. B.; Aksimentiev, A.; Dekker, C., Slowing down DNA translocation through a nanopore in lithium chloride. *Nano Letters* **2012,** *12* (2), 1038-1044.

272. Stellwagen, E.; Dong, Q.; Stellwagen, N. C., Monovalent cations affect the free solution mobility of DNA by perturbing the hydrogen-bonded structure of water. *Biopolymers* **2005,** *78* (2), 62-68.

273. Stellwagen, E.; Stellwagen, N. C., The free solution mobility of DNA in Tris-acetate-EDTA buffers of different concentrations, with and without added NaCl. *Electrophoresis* **2002,** *23* (12), 1935-1941.

274. Stellwagen, E.; Stellwagen, N. C., Probing the electrostatic shielding of DNA with capillary electrophoresis. *Biophysical journal* **2003,** *84* (3), 1855-1866.

275. Wanunu, M.; Cohen-Karni, D.; Johnson, R. R.; Fields, L.; Benner, J.; Peterman, N.; Zheng, Y.; Klein, M. L.; Drndic, M., Discrimination of Methylcytosine from Hydroxymethylcytosine in DNA Molecules. *J. Am. Chem. Soc.* **2011,** *133* (3), 486--492.

276. Marcus, Y., *Ion properties*. Marcel Dekker: New York, 1997.

277. McFail-Isom, L.; Sines, C. C.; Williams, L. D., DNA structure: cations in charge? *Current opinion in structural biology* **1999,** *9* (3), 298-304.

278. McConnell, K. J.; Beveridge, D., DNA structure: what's in charge? *Journal of molecular biology* **2000,** *304* (5), 803-820.

279. Hud, N. V.; Plavec, J., A unified model for the origin of DNA sequence-directed curvature. *Biopolymers* **2003,** *69* (1), 144-158.

280. Stellwagen, N. C.; Gelfi, C.; Righetti, P. G., The free solution mobility of DNA. *Biopolymers* **1997,** *42* (6), 687-703.

281. Schiedt, B.; Healy, K.; Morrison, A. P.; Neumann, R.; Siwy, Z., Transport of ions and biomolecules through single asymmetric nanopores in polymer films. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* **2005,** *236* (1), 109-116.

282. Stellwagen, E.; Dong, Q.; Stellwagen, N. C., Monovalent cations affect the free solution mobility of DNA by perturbing the hydrogen-bonded structure of water. *Biopolymers* **2005,** *78* (2), 62-68.

283. Kawano, R.; Schibel, A. E.; Cauley, C.; White, H. S., Controlling the translocation of single-stranded DNA through α-hemolysin ion channels using viscosity. *Langmuir* **2008,** *25* (2), 1233-1237.

284. Fologea, D.; Uplinger, J.; Thomas, B.; McNabb, D. S.; Li, J., Slowing DNA translocation in a solid-state nanopore. *Nano letters* **2005,** *5* (9), 1734-1737.

285. Ralph, H.; William, S.; Herring, F. G.; Madura Jeffrey, D., General chemistry: Principles and modern applications. *Prentice Hall* **2007**, 606.

286. Falkenhagen, H., The principal ideas in the interionic attraction theory of strong electrolytes. *Reviews of Modern Physics* **1931,** *3* (3), 412.

287. Wanunu, M.; Morrison, W.; Rabin, Y.; Grosberg, A. Y.; Meller, A., Electrostatic focusing of unlabelled DNA into nanoscale pores using a salt gradient. *Nature nanotechnology* **2010,** *5* (2), 160-165.

288. Hatlo, M. M.; Panja, D.; van Roij, R., Translocation of DNA Molecules through Nanopores with Salt Gradients: The Role of Osmotic Flow. *Physical Review Letters* **2011,** *107* (6), 068101.

289. Chou, T., Enhancement of charged macromolecule capture by nanopores in a salt gradient. *The Journal of chemical physics* **2009,** *131* (3), 034703.

290. Grosberg, A. Y.; Rabin, Y., DNA capture into a nanopore: Interplay of diffusion and electrohydrodynamics. *J. Chem. Phys.* **2010,** *133* (16), 165102.

291. Grosberg, A. Y.; Rabin, Y., DNA capture into a nanopore: interplay of diffusion and electrohydrodynamics. *The Journal of chemical physics* **2010,** *133* (16), 10B617.

292. Siwy, Z.; Kosińska, I.; Fuliński, A.; Martin, C., Asymmetric diffusion through synthetic nanopores. *Physical review letters* **2005,** *94* (4), 048102.

293. Wang, Y.; Zheng, D.; Tan, Q.; Wang, M. X.; Gu, L.-Q., Nanopore-based detection of circulating microRNAs in lung cancer patients. *Nature nanotechnology* **2011,** *6* (10), 668.

294. Wen, S.; Zeng, T.; Liu, L.; Zhao, K.; Zhao, Y.; Liu, X.; Wu, H.-C., Highly sensitive and selective DNA-based detection of mercury (II) with α-hemolysin nanopore. *Journal of the American Chemical Society* **2011,** *133* (45), 18312-18317.

295. Simpson, J. T.; Workman, R. E.; Zuzarte, P.; David, M.; Dursi, L.; Timp, W., Detecting DNA cytosine methylation using nanopore sequencing. *nature methods* **2017,** *14* (4), 407-410.

296. Rand, A. C.; Jain, M.; Eizenga, J. M.; Musselman-Brown, A.; Olsen, H. E.; Akeson, M.; Paten, B., Mapping DNA methylation with high-throughput nanopore sequencing. *nature methods* **2017,** *14* (4), 411-413.

297. Hodges-Garcia, Y.; Hagerman, P. J., Cytosine methylation can induce local distortions in the structure of duplex DNA. *Biochemistry* **1992,** *31* (33), 7595-9.

298. Hagerman, P. J., Pyrimidine 5-methyl groups influence the magnitude of DNA curvature. *Biochemistry* **1990,** *29* (8), 1980-3.

299. Norberg, J.; Vihinen, M., Molecular dynamics simulation of the effects of cytosine methylation on structure of oligonucleotides. *Journal of Molecular Structure-Theochem* **2001,** *546*, 51-62.

300. Dantas Machado, A. C.; Zhou, T.; Rao, S.; Goel, P.; Rastogi, C.; Lazarovici, A.; Bussemaker, H. J.; Rohs, R., Evolving insights on how cytosine methylation affects protein–DNA binding. *Briefings in functional genomics* **2014,** *14* (1), 61-73.

301. Rao, S.; Chiu, T.-P.; Kribelbauer, J. F.; Mann, R. S.; Bussemaker, H. J.; Rohs, R., Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein–DNA binding. *Epigenetics & chromatin* **2018,** *11* (1), 6.

302. Keyser, U. F.; Koeleman, B. N.; van Dorp, S.; Krapf, D.; Smeets, R. M. M.; Lemay, S. G.; Dekker, N. H.; Dekker, C., Direct force measurements on DNA in a solid-state nanopore. **2006,** *2*, 473.

303. Klose, R. J.; Bird, A. P., Genomic DNA methylation: the mark and its mediators. *Trends in biochemical sciences* **2006,** *31* (2), 89-97.

304. Krishnakumar, R.; Sinha, A.; Bird, S. W.; Jayamohan, H.; Edwards, H. S.; Schoeniger, J. S.; Patel, K. D.; Branda, S. S.; Bartsch, M. S., Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. *Scientific reports* **2018,** *8* (1), 1-13.

305. Wilson, B. D.; Eisenstein, M.; Soh, H. T., High-fidelity nanopore sequencing of ultra-short DNA targets. *Analytical chemistry* **2019,** *91* (10), 6783-6789.

306. Marcus, Y., Ion Properties, Marcus Dekker. *Inc, New York* **1997**.

307. Wang, Y.; Tian, K.; Hunter, L. L.; Ritzo, B.; Gu, L.-Q., Probing molecular pathways for DNA orientational trapping, unzipping and translocation in nanopores by using a tunable overhang sensor. *Nanoscale* **2014,** *6* (19), 11372-11379.

308. Jin, Q.; Fleming, A. M.; Burrows, C. J.; White, H. S., Unzipping kinetics of duplex DNA containing oxidized lesions in an α-hemolysin nanopore. *Journal of the American Chemical Society* **2012,** *134* (26), 11006-11011.

309. Sauer-Budge, A. F.; Nyamwanda, J. A.; Lubensky, D. K.; Branton, D., Unzipping Kinetics of Double-Stranded DNA in a Nanopore. *Physical Review Letters* **2003,** *90* (23), 238101.

310. Mathé, J.; Visram, H.; Viasnoff, V.; Rabin, Y.; Meller, A., Nanopore unzipping of individual DNA hairpin molecules. *Biophysical Journal* **2004,** *87* (5), 3205-3212.

311. Dudko, O. K.; Mathé, J.; Szabo, A.; Meller, A.; Hummer, G., Extracting kinetics from single-molecule force spectroscopy: nanopore unzipping of DNA hairpins. *Biophysical journal* **2007,** *92* (12), 4188-4195.

312. Viasnoff, V.; Chiaruttini, N.; Muzard, J.; Bockelmann, U., Force fluctuations assist nanopore unzipping of DNA. *Journal of Physics: Condensed Matter* **2010,** *22* (45), 454122.

313. Muzard, J.; Martinho, M.; Mathé, J.; Bockelmann, U.; Viasnoff, V., DNA Translocation and Unzipping through a Nanopore: Some Geometrical Effects. *Biophysical Journal* **2010,** *98* (10), 2170-2178.

314. Alishahi, M.; Kamali, R.; Abouali, O. J. T. E. P. J. E., Rigorous study of molecular dynamics of a single dsDNA confined in a nanochannel: Introduction of a critical mobility behaviour. **2015,** *38* (8), 92.

315. Strick, T. R.; Bensimon, D.; Croquette, V., Micro-mechanical measurement of the torsional modulus of DNA. *Genetica* **1999,** *106* (1), 57-62.

316. Scarsdale, J. N.; Webb, H. D.; Ginder, G. D.; Williams Jr, D. C., Solution structure and dynamic analysis of chicken MBD2 methyl binding domain bound to a target-methylated DNA sequence. *Nucleic acids research* **2011,** *39* (15), 6741-6752.

317. Torchy, M. P.; Hamiche, A.; Klaholz, B. P., Structure and function insights into the NuRD chromatin remodeling complex. *Cellular and Molecular Life Sciences* **2015,** *72* (13), 2491-2507.

318. Liu, Y.; Zhang, X.; Blumenthal, R. M.; Cheng, X., A common mode of recognition for methylated CpG. *Trends in biochemical sciences* **2013,** *38* (4), 177-183.

319. Wan, J. C.; Massie, C.; Garcia-Corbacho, J.; Mouliere, F.; Brenton, J. D.; Caldas, C.; Pacey, S.; Baird, R.; Rosenfeld, N., Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer* **2017,** *17* (4), 223.

320. Leon, S.; Shapiro, B.; Sklaroff, D.; Yaros, M., Free DNA in the serum of cancer patients and the effect of therapy. *Cancer research* **1977,** *37* (3), 646-650.

321. Thierry, A.; El Messaoudi, S.; Gahan, P.; Anker, P.; Stroun, M., Origins, structures, and functions of circulating DNA in oncology. *Cancer and metastasis reviews* **2016,** *35* (3), 347-376.

322. Stroun, M.; Anker, P.; Maurice, P.; Lyautey, J.; Lederrey, C.; Beljanski, M., Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* **1989,** *46* (5), 318-322.

323. Bardelli, A.; Pantel, K., Liquid biopsies, what we do not know (yet). *Cancer cell* **2017,** *31* (2), 172-179.

324. Diaz Jr, L. A.; Bardelli, A., Liquid biopsies: genotyping circulating tumor DNA. *Journal of clinical oncology* **2014,** *32* (6), 579.

325. Perakis, S.; Speicher, M. R., Emerging concepts in liquid biopsies. *BMC medicine* **2017,** *15* (1), 75.

326. Van Der Pol, Y.; Mouliere, F., Toward the early detection of cancer by decoding the epigenetic and environmental fingerprints of cell-free DNA. *Cancer cell* **2019,** *36* (4), 350-368.

327. Siravegna, G.; Marsoni, S.; Siena, S.; Bardelli, A., Integrating liquid biopsies into the management of cancer. *Nature reviews Clinical oncology* **2017,** *14* (9), 531.

328. Stewart, C. M.; Kothari, P. D.; Mouliere, F.; Mair, R.; Somnay, S.; Benayed, R.; Zehir, A.; Weigelt, B.; Dawson, S. J.; Arcila, M. E., The value of cell-free DNA for molecular pathology. *The Journal of pathology* **2018,** *244* (5), 616-627.

329. Cheng, F.; Su, L.; Qian, C., Circulating tumor DNA: a promising biomarker in the liquid biopsy of cancer. *Oncotarget* **2016,** *7* (30), 48832.

330. Botezatu, I.; Serdyuk, O. g.; Potapova, G.; Shelepov, V.; Alechina, R.; Molyaka, Y.; Anan'ev, V.; Bazin, I.; Garin, A.; Narimanov, M., Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism. *Clinical chemistry* **2000,** *46* (8), 1078-1084.

331. Chan, K. A.; Leung, S. F.; Yeung, S. W.; Chan, A. T.; Lo, Y. D., Quantitative analysis of the transrenal excretion of circulating EBV DNA in nasopharyngeal carcinoma patients. *Clinical Cancer Research* **2008,** *14* (15), 4809-4813.

332. Birkenkamp-Demtröder, K.; Nordentoft, I.; Christensen, E.; Høyer, S.; Reinert, T.; Vang, S.; Borre, M.; Agerbæk, M.; Jensen, J. B.; Ørntoft, T. F., Genomic alterations in liquid biopsies from patients with bladder cancer. *European urology* **2016,** *70* (1), 75-82.

333. Reckamp, K. L.; Melnikova, V. O.; Karlovich, C.; Sequist, L. V.; Camidge, D. R.; Wakelee, H.; Perol, M.; Oxnard, G. R.; Kosco, K.; Croucher, P., A highly sensitive and quantitative test platform for detection of NSCLC EGFR mutations in urine and plasma. *Journal of Thoracic Oncology* **2016,** *11* (10), 1690-1700.

334. Su, Y.-H.; Wang, M.; Brenner, D. E.; Ng, A.; Melkonyan, H.; Umansky, S.; Syngal, S.; Block, T. M., Human urine contains small, 150 to 250 nucleotide-sized, soluble DNA derived from the circulation and may be useful in the detection of colorectal cancer. *The journal of molecular diagnostics* **2004,** *6* (2), 101-107.

335. Mithani, S. K.; Smith, I. M.; Zhou, S.; Gray, A.; Koch, W. M.; Maitra, A.; Califano, J. A., Mitochondrial resequencing arrays detect tumor-specific mutations in salivary rinses of patients with head and neck cancer. *Clinical Cancer Research* **2007,** *13* (24), 7335-7340.

336. Wang, Y.; Springer, S.; Zhang, M.; McMahon, K. W.; Kinde, I.; Dobbyn, L.; Ptak, J.; Brem, H.; Chaichana, K.; Gallia, G. L., Detection of tumor-derived DNA in cerebrospinal fluid of patients with primary tumors of the brain and spinal cord. *Proceedings of the National Academy of Sciences* **2015,** *112* (31), 9704-9709.

337. Pan, W.; Gu, W.; Nagpal, S.; Gephart, M. H.; Quake, S. R., Brain tumor mutations detected in cerebral spinal fluid. *Clinical chemistry* **2015,** *61* (3), 514-522.

338. De Mattos-Arruda, L.; Mayor, R.; Ng, C. K.; Weigelt, B.; Martínez-Ricarte, F.; Torrejon, D.; Oliveira, M.; Arias, A.; Raventos, C.; Tang, J., Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. *Nature communications* **2015,** *6*, 8839.

339. Lehmann-Werman, R.; Zick, A.; Paweletz, C.; Welch, M.; Hubert, A.; Maoz, M.; Davidy, T.; Magenheim, J.; Piyanzin, S.; Neiman, D., Specific detection of cell-free DNA derived from intestinal epithelial cells using methylation patterns. *bioRxiv* **2018**, 409219.

340. Sun, K.; Jiang, P.; Chan, K. A.; Wong, J.; Cheng, Y. K.; Liang, R. H.; Chan, W.-k.; Ma, E. S.; Chan, S. L.; Cheng, S. H., Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences* **2015,** *112* (40), E5503-E5512.

341. Underhill, H. R.; Kitzman, J. O.; Hellwig, S.; Welker, N. C.; Daza, R.; Baker, D. N.; Gligorich, K. M.; Rostomily, R. C.; Bronner, M. P.; Shendure, J., Fragment length of circulating tumor DNA. *PLoS genetics* **2016,** *12* (7).

342. Lehmann-Werman, R.; Magenheim, J.; Moss, J.; Neiman, D.; Abraham, O.; Piyanzin, S.; Zemmour, H.; Fox, I.; Dor, T.; Grompe, M., Monitoring liver damage using hepatocyte-specific methylation markers in cell-free circulating DNA. *JCI insight* **2018,** *3* (12).

343. Ju, Y. S.; Alexandrov, L. B.; Gerstung, M.; Martincorena, I.; Nik-Zainal, S.; Ramakrishna, M.; Davies, H. R.; Papaemmanuil, E.; Gundem, G.; Shlien, A., Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife* **2014,** *3*, e02935.

344. Corcoran, R. B.; Chabner, B. A., Application of cell-free DNA analysis to cancer treatment. *New England Journal of Medicine* **2018,** *379* (18), 1754-1765.

345. Chen, X. Q.; Stroun, M.; Magnenat, J.-L.; Nicod, L. P.; Kurt, A.-M.; Lyautey, J.; Lederrey, C.; Anker, P., Microsatellite alterations in plasma DNA of small cell lung cancer patients. *Nature medicine* **1996,** *2* (9), 1033-1035.

346. Gormally, E.; Caboux, E.; Vineis, P.; Hainaut, P., Circulating free DNA in plasma or serum as biomarker of carcinogenesis: practical aspects and biological significance. *Mutation Research/Reviews in Mutation Research* **2007,** *635* (2-3), 105-117.

347. Mouliere, F.; El Messaoudi, S.; Pang, D.; Dritschilo, A.; Thierry, A. R., Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer. *Molecular oncology* **2014,** *8* (5), 927-941.

348. Mouliere, F.; Robert, B.; Peyrotte, E. A.; Del Rio, M.; Ychou, M.; Molina, F.; Gongora, C.; Thierry, A. R., High fragmentation characterizes tumour-derived circulating DNA. *PloS one* **2011,** *6* (9).

349. Shapiro, B.; Chakrabarty, M.; Cohn, E. M.; Leon, S. A., Determination of circulating DNA levels in patients with benign or malignant gastrointestinal disease. *Cancer* **1983,** *51* (11), 2116-2120.

350. Bettegowda, C.; Sausen, M.; Leary, R. J.; Kinde, I.; Wang, Y.; Agrawal, N.; Bartlett, B. R.; Wang, H.; Luber, B.; Alani, R. M., Detection of circulating tumor DNA in early-and late-stage human malignancies. *Science translational medicine* **2014,** *6* (224), 224ra24-224ra24.

351. Zill, O. A.; Banks, K. C.; Fairclough, S. R.; Mortimer, S. A.; Vowles, J. V.; Mokhtari, R.; Gandara, D. R.; Mack, P. C.; Odegaard, J. I.; Nagy, R. J., The landscape of actionable genomic alterations in cell-free circulating tumor DNA from 21,807 advanced cancer patients. *Clinical Cancer Research* **2018,** *24* (15), 3528-3538.

352. Diehl, F.; Li, M.; Dressman, D.; He, Y.; Shen, D.; Szabo, S.; Diaz, L. A.; Goodman, S. N.; David, K. A.; Juhl, H., Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proceedings of the National Academy of Sciences* **2005,** *102* (45), 16368-16373.

353. Dawson, S.-J.; Tsui, D. W.; Murtaza, M.; Biggs, H.; Rueda, O. M.; Chin, S.-F.; Dunning, M. J.; Gale, D.; Forshew, T.; Mahler-Araujo, B., Analysis of circulating tumor DNA to monitor metastatic breast cancer. *New England Journal of Medicine* **2013,** *368* (13), 1199-1209.

354. Tie, J.; Wang, Y.; Tomasetti, C.; Li, L.; Springer, S.; Kinde, I.; Silliman, N.; Tacey, M.; Wong, H.-L.; Christie, M., Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Science translational medicine* **2016,** *8* (346), 346ra92-346ra92.

355. Chaudhuri, A. A.; Chabon, J. J.; Lovejoy, A. F.; Newman, A. M.; Stehr, H.; Azad, T. D.; Khodadoust, M. S.; Esfahani, M. S.; Liu, C. L.; Zhou, L., Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. *Cancer discovery* **2017,** *7* (12), 1394-1403.

356. Cohen, J. D.; Li, L.; Wang, Y.; Thoburn, C.; Afsari, B.; Danilova, L.; Douville, C.; Javed, A. A.; Wong, F.; Mattox, A., Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **2018,** *359* (6378), 926-930.

357. Fazel, R.; Krumholz, H. M.; Wang, Y.; Ross, J. S.; Chen, J.; Ting, H. H.; Shah, N. D.; Nasir, K.; Einstein, A. J.; Nallamothu, B. K., Exposure to low-dose ionizing radiation from medical imaging procedures. *New England Journal of Medicine* **2009,** *361* (9), 849-857.

358. Overman, M. J.; Modak, J.; Kopetz, S.; Murthy, R.; Yao, J. C.; Hicks, M. E.; Abbruzzese, J. L.; Tam, A. L., Use of research biopsies in clinical trials: are risks and benefits adequately discussed? *Journal of clinical oncology* **2013,** *31* (1), 17.

359. Popper, H. H., Commentary on tumor heterogeneity. *Translational lung cancer research* **2016,** *5* (4), 433.

360. De Mattos-Arruda, L.; Weigelt, B.; Cortes, J.; Won, H.; Ng, C.; Nuciforo, P.; Bidard, F.-C.; Aura, C.; Saura, C.; Peg, V., Capturing intra-tumor genetic heterogeneity by de novo mutation profiling of circulating cell-free tumor DNA: a proof-of-principle. *Annals of oncology* **2014,** *25* (9), 1729-1735.

361. Jamal-Hanjani, M.; Wilson, G.; Horswell, S.; Mitter, R.; Sakarya, O.; Constantin, T.; Salari, R.; Kirkizlar, E.; Sigurjonsson, S.; Pelham, R., Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer. *Annals of Oncology* **2016,** *27* (5), 862-867.

362. To, E. W.; Chan, K. A.; Leung, S.-F.; Chan, L. Y.; To, K.-F.; Chan, A. T.; Johnson, P. J.; Lo, Y. D., Rapid clearance of plasma Epstein-Barr virus DNA after surgical treatment of nasopharyngeal carcinoma. *Clinical Cancer Research* **2003,** *9* (9), 3254-3259.

363. Lo, Y. D.; Zhang, J.; Leung, T. N.; Lau, T. K.; Chang, A. M.; Hjelm, N. M., Rapid clearance of fetal DNA from maternal plasma. *The American Journal of Human Genetics* **1999,** *64* (1), 218-224.

364. Yao, W.; Mei, C.; Nan, X.; Hui, L., Evaluation and comparison of in vitro degradation kinetics of DNA in serum, urine and saliva: a qualitative study. *Gene* **2016,** *590* (1), 142-148.

365. Bronner, I. F.; Quail, M. A.; Turner, D. J.; Swerdlow, H., Improved protocols for illumina sequencing. *Current protocols in human genetics* **2013,** *79* (1), 18.2. 1-18.2. 42.

366. Stroun, M.; Anker, P.; Lyautey, J.; Lederrey, C.; Maurice, P. A., Isolation and characterization of DNA from the plasma of cancer patients. *European Journal of Cancer and Clinical Oncology* **1987,** *23* (6), 707-712.

367. Cheng, S. H.; Jiang, P.; Sun, K.; Cheng, Y. K.; Chan, K. A.; Leung, T. Y.; Chiu, R. W.; Lo, Y. D., Noninvasive prenatal testing by nanopore sequencing of maternal plasma DNA: feasibility assessment. *Clinical chemistry* **2015,** *61* (10), 1305-1306.

368. Thakur, B. K.; Zhang, H.; Becker, A.; Matei, I.; Huang, Y.; Costa-Silva, B.; Zheng, Y.; Hoshino, A.; Brazier, H.; Xiang, J., Double-stranded DNA in exosomes: a novel biomarker in cancer detection. *Cell research* **2014,** *24* (6), 766-769.

369. Kahlert, C.; Melo, S. A.; Protopopov, A.; Tang, J.; Seth, S.; Koch, M.; Zhang, J.; Weitz, J.; Chin, L.; Futreal, A., Identification of double-stranded genomic DNA spanning all chromosomes with mutated KRAS and p53 DNA in the serum exosomes of patients with pancreatic cancer. *Journal of Biological Chemistry* **2014,** *289* (7), 3869-3875.

370. Jahr, S.; Hentze, H.; Englisch, S.; Hardt, D.; Fackelmayer, F. O.; Hesch, R.-D.; Knippers, R., DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer research* **2001,** *61* (4), 1659-1665.

371. Khatami, F.; Larijani, B.; Tavangar, S. M., The presence of tumor extrachomosomal circular DNA (ecDNA) as a component of liquid biopsy in blood. *Medical hypotheses* **2018,** *114*, 5-7.

372. Verhaak, R. G.; Bafna, V.; Mischel, P. S., Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nature Reviews Cancer* **2019,** *19* (5), 283-288.

373. Breitbach, S.; Tug, S.; Helmig, S.; Zahn, D.; Kubiak, T.; Michal, M.; Gori, T.; Ehlert, T.; Beiter, T.; Simon, P., Direct quantification of cell-free, circulating DNA from unpurified plasma. *PloS one* **2014,** *9* (3).

374. Beránek, M.; Sirák, I.; Vosmik, M.; Petera, J.; Drastíková, M.; Palicka, V., Carrier molecules and extraction of circulating tumor DNA for next generation sequencing in colorectal cancer. *Acta Medica (Hradec Kralove)* **2016,** *59* (2), 54-8.

375. Gansauge, M.-T.; Meyer, M., Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature protocols* **2013,** *8* (4), 737.

376. Morozkin, E.; Sil'nikov, V.; Rykova, E. Y.; Vlassov, V.; Laktionov, P., Extracellular DNA in culture of primary and transformed cells, infected and not infected with mycoplasma. *Bulletin of experimental biology and medicine* **2009,** *147* (1), 63-65.

377. Morozkin, E. S.; Laktionov, P. P.; Rykova, E. Y.; Bryzgunova, O. E.; Vlassov, V. V., Release of nucleic acids by eukaryotic cells in tissue culture. *Nucleosides, Nucleotides and Nucleic Acids* **2004,** *23* (6-7), 927-930.

378. Choi, J. J.; Reich Iii, C.; Pisetsky, D., Release of DNA from dead and dying lymphocyte and monocyte cell lines in vitro. *Scandinavian journal of immunology* **2004,** *60* (1-2), 159-166.

379. Jiang, P.; Chan, C. W.; Chan, K. A.; Cheng, S. H.; Wong, J.; Wong, V. W.-S.; Wong, G. L.; Chan, S. L.; Mok, T. S.; Chan, H. L., Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proceedings of the National Academy of Sciences* **2015,** *112* (11), E1317-E1325.

380. Bronkhorst, A. J.; Wentzel, J. F.; Aucamp, J.; Van Dyk, E.; Du Plessis, L.; Pretorius, P. J., Characterization of the cell-free DNA released by cultured cancer cells. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **2016,** *1863* (1), 157-165.

381. Bitto, N. J.; Chapman, R.; Pidot, S.; Costin, A.; Lo, C.; Choi, J.; D'cruze, T.; Reynolds, E. C.; Dashper, S. G.; Turnbull, L., Bacterial membrane vesicles transport their DNA cargo into host cells. *Scientific reports* **2017,** *7* (1), 1-11.

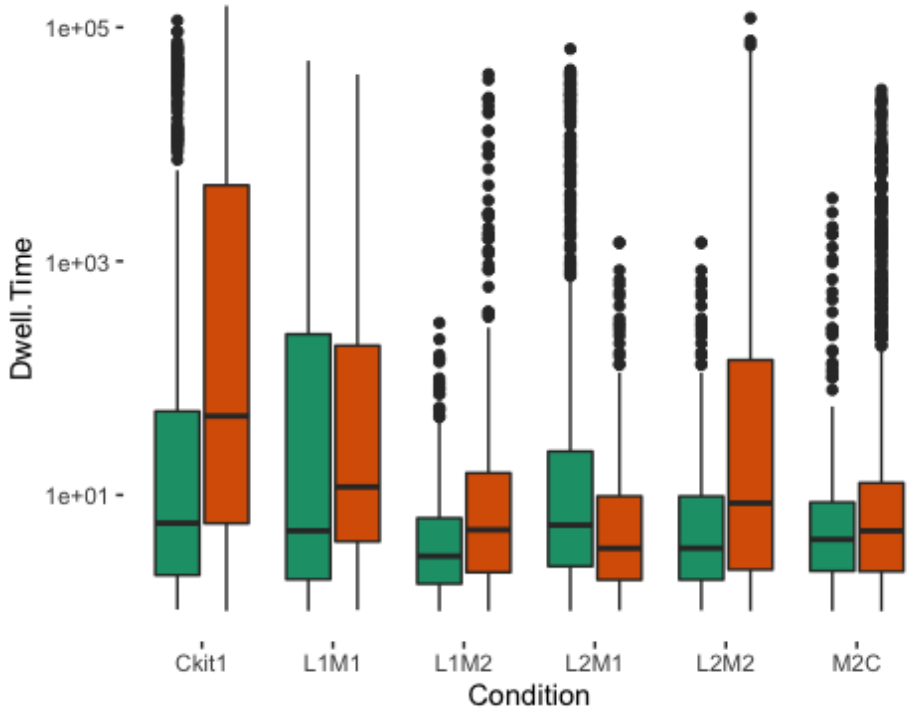
382. Ali, M. M.; Li, F.; Zhang, Z.; Zhang, K.; Kang, D.-K.; Ankrum, J. A.; Le, X. C.; Zhao, W., Rolling circle amplification: a versatile tool for chemical biology, materials science and medicine. *Chemical Society Reviews* **2014,** *43* (10), 3324-3341.

383. Lin, C.; Xie, M.; Chen, J. J.; Liu, Y.; Yan, H., Rolling-Circle Amplification of a DNA Nanojunction. *Angewandte Chemie International Edition* **2006,** *45* (45), 7537-7539.

Appendix A

Example of R Code for Dwell Time Boxplots

```
################ Dwell time distribution - Grouped Boxplot ################
dwelltime <-read.csv("Amplitude_Dwelltime_allconditions_mut_4columns_new.csv", sep = ",",
header = TRUE)
dwelltime.f <- dwelltime %>%filter(dwelltime$Amplitude > 60 & dwelltime$Dwell.Time > 1)
head(dwelltime.f)

##   Amplitude Dwell.Time Condition Treatment
## 1    86.263      1.80     L1M1    No CX
## 2    86.226      7.92     L1M1    No CX
## 3    85.990      1.84     L1M1    No CX
## 4    85.956    818.26     L1M1    No CX
## 5    85.549  27910.62     L1M1    No CX
## 6    85.416      2.54     L1M1    No CX

# annotation table with adjusted pvals and y-position of the labels
anno_df = compare_means(Dwell.Time ~ Treatment, group.by = "Condition", data = dwelltime.
f) %>%
  mutate(y_pos = 40)

## Warning: `cols` is now required.
## Please use `cols = c(p)`

p <- ggplot(dwelltime.f, aes(x=Condition, y=Dwell.Time)) +
  geom_boxplot(aes(fill = Treatment)) +
  scale_fill_brewer(palette = "Dark2")+
  scale_y_continuous(trans = 'log10')

p <- p + stat_compare_means(comparisons = list(c("No CX", "With CX")), size = 4)+
  theme(legend.position="none")+ # Remove legend
  # set transparency
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_rect(fill = "transparent",colour = NA),
    plot.background = element_rect(fill = "transparent",colour = NA)
  )
p
```




```
ggsave(p, filename = "Ckit_Mut_Dwelltime_GroupedBoxplot.png",  bg = "transparent")

## Saving 5 x 4 in image
```

Appendix B

Example of R Code for Summary Statistic and Pairwise Comparison

```
##########Summary of the dwell time table: mean, max, min, median## #########
dwelltime <-read.csv("Amplitude_Dwelltime_allconditions_mut_new.csv", sep = ",", header = T
RUE)
dwelltime.f <- dwelltime %>%filter(dwelltime$Amplitude > 50 & dwelltime$Dwell.Time > 1)
stat.sum <- as.data.frame(dwelltime.f %>%
                    group_by(Condition)%>%
                    summarise(Mean = mean(Dwell.Time), Max = max(Dwell.Time), Min =min(
Dwell.Time), Median = median(Dwell.Time), Std = sd(Dwell.Time) ))

stat.sum

##    Condition        Mean       Max    Min  Median       Std
## 1      Ckit1  3677.865950 114411.352 1.0500  5.5500 12755.3138
## 2   Ckit1_CX 10245.271661 152850.000 1.0200 31.1900 25301.9010
## 3      L1M1  3455.983506  51802.340 1.0200  3.5500  9639.9690
## 4    L1M1_CX  1664.956641  39545.141 1.0230  8.8000  5819.8137
## 5      L1M2     8.650711    296.945 1.0160  2.7190    22.9824
## 6    L1M2_CX   563.404257  39771.090 1.0200  4.4885  3364.7391
## 7      L2M1  2055.286784  65396.621 1.0200  5.2800  7247.3926
## 8    L2M1_CX    30.167402   1447.891 1.0310  3.2660   132.3704
## 9      L2M2    30.100872   1447.891 1.0310  3.2830   132.2039
## 10   L2M2_CX  6388.554461 119939.500 1.0200  8.9800 16950.6187
## 11      M2C    14.967050   3449.960 1.0200  4.0800   121.7973
## 12    M2C_CX   312.115897  29303.414 1.0155  4.8599  1863.5142

# Compute the analysis of variance
res.aov <- aov(Dwell.Time ~ Condition, data = dwelltime.f)
# Summary of the analysis
summary(res.aov)

##               Df    Sum Sq   Mean Sq F value Pr(>F)
## Condition     11 4.461e+10 4.056e+09   78.04 <2e-16 ***
## Residuals   9294 4.830e+11 5.197e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Multiple pairwise comparison
TukeyHSD(res.aov)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Dwell.Time ~ Condition, data = dwelltime.f)
##
## $Condition
##                       diff        lwr        upr     p adj
```

```
## Ckit1_CX-Ckit1    6.567406e+03   4934.7940  8200.017464 0.0000000
## L1M1-Ckit1       -2.218824e+02  -2252.6909  1808.925998 0.9999999
## L1M1_CX-Ckit1    -2.012909e+03  -3553.2472  -472.571381 0.0011828
## L1M2-Ckit1       -3.669215e+03  -4968.2458 -2370.184726 0.0000000
## L1M2_CX-Ckit1    -3.114462e+03  -4391.5832 -1837.340205 0.0000000
## L2M1-Ckit1       -1.622579e+03  -2778.6092  -466.549135 0.0002834
## L2M1_CX-Ckit1    -3.647699e+03  -5097.9737 -2197.423364 0.0000000
## L2M2-Ckit1       -3.647765e+03  -5096.7379 -2198.792281 0.0000000
## L2M2_CX-Ckit1     2.710689e+03   1130.0860  4291.291062 0.0000014
## M2C-Ckit1        -3.662899e+03  -4600.9346 -2724.863192 0.0000000
## M2C_CX-Ckit1     -3.365750e+03  -4313.9376 -2417.562523 0.0000000
## L1M1-Ckit1_CX    -6.789288e+03  -9130.8543 -4447.721959 0.0000000
## L1M1_CX-Ckit1_CX -8.580315e+03 -10511.9957 -6648.634339 0.0000000
## L1M2-Ckit1_CX    -1.023662e+04 -11981.9661 -8491.275828 0.0000000
## L1M2_CX-Ckit1_CX -9.681867e+03 -11410.9679 -7952.766882 0.0000000
## L2M1-Ckit1_CX    -8.189985e+03  -9831.6775 -6548.292229 0.0000000
## L2M1_CX-Ckit1_CX -1.021510e+04 -12075.7621 -8354.446464 0.0000000
## L2M2-Ckit1_CX    -1.021517e+04 -12074.8136 -8355.527950 0.0000000
## L2M2_CX-Ckit1_CX -3.856717e+03  -5820.6555 -1892.778875 0.0000000
## M2C-Ckit1_CX     -1.023030e+04 -11726.5199 -8734.089282 0.0000000
## M2C_CX-Ckit1_CX  -9.933156e+03 -11435.7565 -8430.555048 0.0000000
## L1M1_CX-L1M1     -1.791027e+03  -4069.2174   487.163654 0.2970685
## L1M2-L1M1        -3.447333e+03  -5569.8293 -1324.836297 0.0000073
## L1M2_CX-L1M1     -2.892579e+03  -5001.7379  -783.420553 0.0004611
## L2M1-L1M1        -1.400697e+03  -3438.8127   637.419207 0.5154649
## L2M1_CX-L1M1     -3.425816e+03  -5644.1058 -1207.526398 0.0000293
## L2M2-L1M1        -3.425883e+03  -5643.3211 -1208.444185 0.0000289
## L2M2_CX-L1M1      2.932571e+03    626.9657  5238.176197 0.0019255
## M2C-L1M1         -3.441016e+03  -5363.8847 -1518.148260 0.0000003
## M2C_CX-L1M1      -3.143868e+03  -5071.7086 -1216.026666 0.0000065
## L1M2-L1M1_CX     -1.656306e+03  -3315.6580     3.046169 0.0509311
## L1M2_CX-L1M1_CX  -1.101552e+03  -2743.8095   540.704725 0.5550697
## L2M1-L1M1_CX      3.903301e+02  -1159.6294  1940.289674 0.9996296
## L2M1_CX-L1M1_CX  -1.634789e+03  -3415.0328   145.454347 0.1080353
## L2M2-L1M1_CX     -1.634856e+03  -3414.0385   144.326988 0.1074817
## L2M2_CX-L1M1_CX   4.723598e+03   2835.6694  6611.526193 0.0000000
## M2C-L1M1_CX      -1.649990e+03  -3044.9377  -255.041525 0.0062430
## M2C_CX-L1M1_CX   -1.352841e+03  -2754.6356    48.954072 0.0703972
## L1M2_CX-L1M2      5.547535e+02   -863.6426  1973.149735 0.9818354
## L2M1-L1M2         2.046636e+03    736.2110  3357.061151 0.0000217
## L2M1_CX-L1M2      2.151669e+01  -1554.5882  1597.621562 1.0000000
## L2M2-L1M2         2.145016e+01  -1553.4564  1596.356705 1.0000000
## L2M2_CX-L1M2      6.379904e+03   4683.1089  8076.698634 0.0000000
## M2C-L1M2          6.316339e+00  -1116.5013  1129.133975 1.0000000
## M2C_CX-L1M2       3.034652e+02   -827.8473  1434.777718 0.9993243
## L2M1-L1M2_CX      1.491883e+03    203.1727  2780.592311 0.0085447
## L2M1_CX-L1M2_CX  -5.332369e+02  -2091.3336  1024.859929 0.9939262
## L2M2-L1M2_CX     -5.333034e+02  -2090.1880  1023.581211 0.9938793
## L2M2_CX-L1M2_CX   5.825150e+03   4145.0693  7505.231125 0.0000000
## M2C-L1M2_CX      -5.484372e+02  -1645.8334   548.959014 0.8968496
## M2C_CX-L1M2_CX   -2.512884e+02  -1357.3747   854.798014 0.9998642
```

```
## L2M1_CX-L2M1     -2.025119e+03  -3485.6096  -564.629147 0.0003670
## L2M2-L2M1        -2.025186e+03  -3484.3829  -565.988946 0.0003600
## L2M2_CX-L2M1      4.333268e+03   2743.2872  5923.248193 0.0000000
## M2C-L2M1         -2.040320e+03  -2994.0726 -1086.566828 0.0000000
## M2C_CX-L2M1      -1.743171e+03  -2706.9101  -779.431707 0.0000002
## L2M2-L2M1_CX     -6.652969e-02  -1701.8739  1701.740880 1.0000000
## L2M2_CX-L2M1_CX   6.358387e+03   4543.1927  8173.581461 0.0000000
## M2C-L2M1_CX      -1.520035e+01  -1310.0120  1279.611326 1.0000000
## M2C_CX-L2M1_CX    2.819485e+02  -1020.2365  1584.133537 0.9999157
## L2M2_CX-L2M2      6.358454e+03   4544.2996  8172.607599 0.0000000
## M2C-L2M2         -1.513382e+01  -1308.4866  1278.218929 1.0000000
## M2C_CX-L2M2       2.820150e+02  -1018.7194  1582.749409 0.9999145
## M2C-L2M2_CX      -6.373587e+03  -7812.8732 -4934.301617 0.0000000
## M2C_CX-L2M2_CX   -6.076439e+03  -7522.3612 -4630.515953 0.0000000
## M2C_CX-M2C        2.971488e+02   -390.0911   984.388807 0.9612128
```

Appendix C

Example of R Code for Machine Learning Classification Models

```
####Machine learning: predicting methylated from unmethylated DNA #####

## With MBD2
bothend.ml <-read.csv("bothend_nome_MBD2_ml.csv", sep = ",", header = TRUE)
head(bothend.ml)

##   Amplitude Dwell.Time Condition
## 1  82.35037   20704.35      nome
## 2  86.29356     802.45      nome
## 3  89.52220     206.75      nome
## 4  55.81189      14.75      nome
## 5  79.11031       6.35      nome
## 6  60.56889       3.95      nome

dim(bothend.ml)

## [1] 1151    3

dwelltime$Condition <- as.factor(dwelltime$Condition)

df_ML <- bothend.ml
df_ML <- na.omit(df_ML)


df_ML.f <- df_ML%>%filter( df_ML$Amplitude > 75 & df_ML$Dwell.Time > 100 & df_ML
$Dwell.Time <10000) #Filtering data: only choose data with dwell time > 50ms.
dim(df_ML.f)

## [1] 360   3

df_ML.f$Amplitude_scaled <-scale(df_ML.f$Amplitude)
df_ML.f$Dwell.Time_scaled <-scale(df_ML.f$Dwell.Time)

set.seed(35)
train_indx <- createDataPartition(df_ML.f$Condition, p = 0.80, list = FALSE)

train_set <- df_ML.f[train_indx,]
test_set <- df_ML.f[-train_indx,]

nrow(train_set)

## [1] 289

dim(train_set)

## [1] 289   5

nrow(test_set)
```

```
## [1] 71

###################### Train the algorithm ############################
fitControl <- trainControl(method="boot",
                           number = 10,
                           preProcOptions = list(thresh = 0.99), # threshold for pca preprocess
                           classProbs = TRUE,
                           savePredictions = TRUE,
                           summaryFunction = twoClassSummary)

##################### Random Forest ##################################

model_rf <- train(Condition~ Amplitude_scaled+Dwell.Time_scaled,data = train_set, method="r
f", metric="ROC",
                  preProcess = c('center', 'scale'), trControl=fitControl)

## note: only 1 unique complexity parameters in default grid. Truncating the grid to 1 .

pred_rf <- predict(model_rf, test_set)
cm_rf <- confusionMatrix(pred_rf, test_set$Condition, positive = "nome")
cm_rf

## Confusion Matrix and Statistics
##
##          Reference
## Prediction me_MBD2 nome
##    me_MBD2      28    7
##    nome          7   29
##
##                Accuracy : 0.8028
##                  95% CI : (0.6914, 0.8878)
##     No Information Rate : 0.507
##     P-Value [Acc > NIR] : 2.446e-07
##
##                   Kappa : 0.6056
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8056
##             Specificity : 0.8000
##          Pos Pred Value : 0.8056
##          Neg Pred Value : 0.8000
##              Prevalence : 0.5070
##          Detection Rate : 0.4085
##    Detection Prevalence : 0.5070
##       Balanced Accuracy : 0.8028
##
##        'Positive' Class : nome
##

################################## Using LDA ###################################
```

```
model_lda_df <- train(Condition ~., data = train_set, method = "lda", metric = "ROC", preProces
s = c("scale", "center"),
                      trControl = fitControl)

prediction_lda_df <- predict(model_lda_df, test_set)
cm_lda_df <- confusionMatrix(prediction_lda_df, test_set$Condition, positive = "nome")
cm_lda_df

## Confusion Matrix and Statistics
##
##          Reference
## Prediction me_MBD2 nome
##    me_MBD2      26   16
##    nome          9   20
##
##                Accuracy : 0.6479
##                  95% CI : (0.5254, 0.7576)
##     No Information Rate : 0.507
##     P-Value [Acc > NIR] : 0.01165
##
##                   Kappa : 0.2976
##
##  Mcnemar's Test P-Value : 0.23014
##
##             Sensitivity : 0.5556
##             Specificity : 0.7429
##          Pos Pred Value : 0.6897
##          Neg Pred Value : 0.6190
##              Prevalence : 0.5070
##          Detection Rate : 0.2817
##    Detection Prevalence : 0.4085
##       Balanced Accuracy : 0.6492
##
##       'Positive' Class : nome
##

################### Logistic regression ####################
model_logreg_df <- train(Condition ~., data = train_set, method = "glm",
                         metric = "ROC", preProcess = c("scale", "center"),
                         trControl = fitControl)

cm_logreg_df <- confusionMatrix(prediction_logreg_df, test_set$Condition, positive = "nome")
cm_logreg_df

## Confusion Matrix and Statistics
##
##          Reference
## Prediction me_MBD2 nome
##    me_MBD2      26   16
##    nome          9   20
##
##                Accuracy : 0.6479
```

```
##                95% CI : (0.5254, 0.7576)
##     No Information Rate : 0.507
##     P-Value [Acc > NIR] : 0.01165
##
##                  Kappa : 0.2976
##
##  Mcnemar's Test P-Value : 0.23014
##
##             Sensitivity : 0.5556
##             Specificity : 0.7429
##          Pos Pred Value : 0.6897
##          Neg Pred Value : 0.6190
##              Prevalence : 0.5070
##          Detection Rate : 0.2817
##    Detection Prevalence : 0.4085
##       Balanced Accuracy : 0.6492
##
##        'Positive' Class : nome
##

################### K-nearest neighbor ##############################
model_knn <- train(Condition~.,
                   data = train_set,
                   method="knn",
                   metric="ROC",
                   preProcess = c('center', 'scale'),
                   tuneLength=10,
                   trControl=fitControl)
pred_knn <- predict(model_knn, test_set)
cm_knn <- confusionMatrix(pred_knn, test_set$Condition, positive = "nome")
cm_knn

## Confusion Matrix and Statistics
##
##          Reference
## Prediction me_MBD2 nome
##    me_MBD2      27   11
##    nome          8   25
##
##               Accuracy : 0.7324
##                 95% CI : (0.6141, 0.8306)
##    No Information Rate : 0.507
##    P-Value [Acc > NIR] : 9.002e-05
##
##                  Kappa : 0.4653
##
##  Mcnemar's Test P-Value : 0.6464
##
##             Sensitivity : 0.6944
##             Specificity : 0.7714
##          Pos Pred Value : 0.7576
```

```
##          Neg Pred Value : 0.7105
##              Prevalence : 0.5070
##          Detection Rate : 0.3521
##    Detection Prevalence : 0.4648
##       Balanced Accuracy : 0.7329
##
##       'Positive' Class : nome

############################ Model Evaluation ##########################
model_list <- list(RF=model_rf, LG=model_logreg_df, LDA = model_lda_df,
                   KNN = model_knn
resamples <- resamples(model_list)
bwplot(resamples, metric = "ROC", main = "Evaluating performance of different algorithms")
```