Biological nanopore MspA for DNA sequencing

Elizabeth A. Manrao

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Jens H. Gundlach, Chair

Michael Schick

Leslie Rosenberg

Program Authorized to Offer Degree:
Physics

University of Washington

Abstract

Biological nanopore MspA for DNA sequencing

Elizabeth A. Manrao

Chair of the Supervisory Committee:
Dr. Jens H. Gundlach
Physics

Unlocking the information hidden in the human genome provides insight into the inner workings of complex biological systems and can be used to greatly improve healthcare. In order to allow for widespread sequencing, new technologies are required that provide fast and inexpensive readings of DNA. Nanopore sequencing is a third generation DNA sequencing technology that is currently being developed to fulfill this need.

In nanopore sequencing, a voltage is applied across a small pore in an electrolyte solution and the resulting ionic current is recorded. When DNA passes through the channel, the ionic current is partially blocked. If the DNA bases uniquely modulate the ionic current flowing through the channel, the time trace of the current can be related to the sequence of DNA passing through the pore. There are two main challenges to realizing nanopore sequencing: identifying a pore with sensitivity to single nucleotides and controlling the translocation of DNA through the pore so that the small single nucleotide current signatures are distinguishable from background noise.

In this dissertation, I explore the use of *Mycobacterium smegmatis* porin A (MspA) for nanopore sequencing. In order to determine MspA's sensitivity to single nucleotides, DNA strands of various compositions are held in the pore as the resulting

ionic current is measured. DNA is immobilized in MspA by attaching it to a large molecule which acts as an anchor. This technique confirms the single nucleotide resolution of the pore and additionally shows that MspA is sensitive to epigenetic modifications and single nucleotide polymorphisms. The forces from the electric field within MspA, the effective charge of nucleotides, and elasticity of DNA are estimated using a Freely Jointed Chain model of single stranded DNA. These results offer insight into the interactions of DNA within the pore. With the nucleotide sensitivity of MspA confirmed, a method is introduced to controllably pass DNA through the pore. Using a DNA polymerase, DNA strands are stepped through MspA one nucleotide at a time. The steps are observable as distinct levels on the ionic-current time-trace and are related to the DNA sequence. These experiments overcome the two fundamental challenges to realizing MspA nanopore sequencing and pave the way to the development of a commercial technology.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables vii

Chapter 1: Introduction 1
1.1 Implications for Widespread Sequencing 1
1.2 Evolution of Sequencing Techniques 3
1.3 Nanopore DNA Sequencing 6

Chapter 2: Using the MspA Nanopore as a Single Molecule Sensor 10
2.1 Experimental Procedure 10
2.2 Data Acquisition and Analysis 25
2.3 Experiments Highlighted in this Dissertation 27

Chapter 3: Biological Porin MspA for Nanopore Sequencing 30
3.1 Preliminary Results 30
3.2 Duplex Interrupted Translocation of DNA through MspA 32
3.3 Conclusions 40

Chapter 4: Nucleotide Discrimination with DNA Immobilized in MspA . . 42
4.1 Introduction 42
4.2 Results 43
4.3 Discussion 49

Chapter 5: Enzyme Controlled Translocation of DNA through MspA using Phi29 DNA Polymerase 54
5.1 Introduction 54
5.2 Results 58

5.3 Discussion 68

Chapter 6: Force Spectroscopy of Single-Stranded DNA by Immobilization in MspA 70
6.1 Introduction 70
6.2 Results 72
6.3 Analysis 78
6.4 Discussion 83

Chapter 7: Conclusions / Outlook 86

Bibliography 88

Appendix A: Supplemental Information for DNA Immobilization Experiments 95
A.1 Supplemental Figures and Tables 95

Appendix B: Supplemental Information for Enzyme Controlled Translocation Experiments 100
B.1 Level Detection and Comparison Methods 100
B.2 Supplemental Figures and Tables 101

Appendix C: Supplemental Information for Force Spectroscopy Experiments . 123
C.1 Force Calculation Details 123
C.2 Homopolymer Thymine Strand 126
C.3 Single Nucleotide Polymorphism (SNP) Strand 126
C.4 Brownian Motion Calculation Details 132

LIST OF FIGURES

Figure Number		Page
1.1	Implications of Widespread DNA Sequencing	2
1.2	Cost to Sequence the Human Genome	5
1.3	Principle of Nanopore Sequencing	7
1.4	Comparison of Nanopore Channels	8
2.1	Schematic Diagram of Experimental Setup	12
2.2	Photo of Experimental Setup	13
2.3	Lipid Bilayer Assembly	14
2.4	MspA Dimensions and Hydrophobicity	15
2.5	MspA Charge Profile and Mutations	16
2.6	Conductances of MspA Mutants	18
2.7	Nucleotide Structure	19
2.8	Structure of DNA	20
2.9	Structure of G-Tetrad	21
2.10	DNA Polymerase Activity	24
2.11	Power Spectral Density for Experimental Setup	27
3.1	Experimental Results for the Confirmation of DNA Translocation	31
3.2	Experimental Results for DNA Interaction with MspA Mutants	33
3.3	Duplex Interrupted DNA Translocation Through MspA	34
3.4	Experimental Results for Hairpin DNA with Homopolymer Tails	36
3.5	Determining the Region of Sensitivity of MspA with Hairpin DNA	37
3.6	Experimental Results for Single Nucleotide Substitutions in Hairpin DNA with a poly-dA Tail	38
3.7	Effect of Hairpin Composition on Residual Ionic Current	39
3.8	Demonstration of Duplex Interrupted (DI) Nanopore Sequencing	41
4.1	Schematic Diagram of DNA Immobilization with NeutrAvidin	44

4.2 Experimental Results for Homopolymer Strands Immobilized in MspA 46
4.3 Experimental Results for Methylated Cytosine Immobilized in MspA 47
4.4 Experimental Results for Single Heteromeric Substitutions in Homopolymer ssDNA Immobilized in MspA 50
4.5 Experimental Results for Immobilized Strands Containing a Single Nucleotide Polymorphism (rs889312) 53

5.1 Cover of Nature Biotechnology, April 2012 55
5.2 Event Structure for Enzyme Controlled DNA Translocation Experiments 56
5.3 Example Current Trace for Enzyme Controlled Translocation of Block Homopolymer DNA During Polymerase Synthesis 62
5.4 Example Current Trace for Enzyme Controlled Translocation of Repeating DNA Template Sequence During Polymerase Synthesis 64
5.5 Experimental Data for Enzyme Controlled Translocation of 'Heteromer DNA 1' 66

6.1 Schematic Diagram of DNA Immobilized in MspA 72
6.2 Experimental Results for Homopolymer DNA Immobilized in MspA at Varying Voltages 73
6.3 Experimental Results for Single Nucleotide Substitutions in a Poly-dA Strand Immobilized in MspA at Varying Voltages 75
6.4 Fluctuations of Inter-Event Resistance for DNA Strands Immobilized in MspA at Varying Voltages 76
6.5 Positioning of DNA within MspA's Recognition Site when Immobilized at Varying Voltages 77
6.6 Voltage Dependent Force on DNA Immobilized in MspA 81
6.7 Voltage Dependent Elasticity of DNA Immobilized in MspA 82

A.1 Example Voltage and Current Trace for DNA Immobilization Experiments 96
A.2 Experimental Results to Determine the Region of Sensitivity of DNA Immobilized in MspA 97
A.3 Experimental Results for Immobilized Strands Containing a Single Nucleotide Polymorphism (rs1447295) 98

B.1 Experimental Results for Control Experiments of Enzyme Controlled DNA Translocation 102

B.2 Similarity of 'Block Homopolymer' DNA Current Pattern for Enzyme Controlled DNA Translocation 104
B.3 Reproducibility of 'Block Homopolymer' DNA Current Pattern for Enzyme Controlled DNA Translocation 106
B.4 Exponential Behavior of Level Durations for Enzyme Controlled DNA Translocation 107
B.5 Location of Synthesis Site Determined from Enzyme Controlled DNA Translocation Experiments 108
B.6 Example Event for Enzyme Controlled DNA Translocation of 'Heteromer DNA 1' 110
B.7 Similarity of 'Heteromer DNA 1' Current Pattern for Enzyme Controlled DNA Translocation 112
B.8 Example Event for Enzyme Controlled DNA Translocation of 'Heteromer DNA 2' 113
B.9 Similarity of 'Heteromer DNA 2' Current Pattern for Enzyme Controlled DNA Translocation 114
B.10 Reproducibility of 'Heteromer DNA 2' Current Pattern for Enzyme Controlled DNA Translocation 115
B.11 Example Event for Enzyme Controlled DNA Translocation of 'Heteromer DNA 3' Sequence 116
B.12 Similarity of 'Heteromer DNA 3' Current Pattern for Enzyme Controlled DNA Translocation 117
B.13 Reproducibility of 'Heteromer DNA 3' Current Pattern for Enzyme Controlled DNA Translocation 118
B.14 Example Event for Enzyme Controlled DNA Translocation of 'Heteromer DNA 4' Sequence 119
B.15 Similarity of 'Heteromer DNA 4' Current Pattern for Enzyme Controlled DNA Translocation 120
B.16 Reproducibility of 'Heteromer DNA 4' Current Pattern for Enzyme Controlled DNA Translocation 121

C.1 Schematic Diagram of the Freely Jointed Chain Model Applied to DNA Immobilized in MspA 124
C.2 Force-Extension Curves for Immobilized DNA Modeled as a Freely Jointed Chain 127

C.3 Experimental Results for Single Nucleotide Substitutions in Poly-dT strands Immobilized in MspA 134
C.4 Experimental Results for Single Nucleotide Polymorphism Strands . . 135
C.5 Width of MspA's Recognition Site Compared to Brownian Motion Calculations of DNA Immobilized in MspA 136

LIST OF TABLES

Table Number Page

A.1 DNA Sequences used in DNA Immobilization Experiments 99

B.1 DNA Sequences used in Enzyme Controlled DNA Translocation Experiments 122

C.1 FJC Calculations for Single Nucleotide Substitutions in Poly-dA . . . 128

C.2 FJC Calculations for Single Nucleotide Substitutions in Poly-dT . . . 130

C.3 FJC Calculations for SNP Strands 131

C.4 DNA Sequences used in Force Spectroscopy Experiments 137

ACKNOWLEDGMENTS

Many people have helped me throughout my graduate career. Jens Gundlach, was a great advisor who provided inspiration and always kept the lab motivated with his excitement about our results and the future direction of our research. My predecessor, Ian Derrington, recruited me to lab and provided much mentoring and assistance throughout my research. Kyle Langford trained me on how to form lipid bilayers and run a nanopore experiment. Tom Butler laid the groundwork for our lab. This work could not have been completed without the hard work and dedication of my many nanopore colleagues, including Joshua Bartlett, Marcus Collins, David Feldman, Nathaniel Gillgren, Matthew Hopper, Andrew Laszlo, Graham Nayler, and Jennifer Wierman. Finally, my friends and family have been a source of joy and have kept me grounded and sane. I particularly want to thank my partner, Michael Hotz, for his continued love and support.

DEDICATION

This work is dedicated to my father, Kenneth Manrao, who taught me the importance of education.

Chapter 1

INTRODUCTION

Scientific advances are uncovering new information about life every day. We are coming closer to understanding how living organisms form, function, and respond to stimuli. The key to this scientific revolution lies with understanding the information held in an organism's genome. A genome encodes genetic information as a DNA sequence. This sequence contains instructions for the development and functioning of all cell components, including proteins and RNA. Understanding this programing will aid not only biological research but also the practice of medicine.

In this chapter, I discuss the importance of fast, inexpensive DNA sequencing. I then review the current state of sequencing technology and explain the need for a new system able to read DNA with increased speed and at reduced cost. Finally, I introduce nanopore DNA sequencing and outline the progress that has been made with developing this technology.

1.1 Implications for Widespread Sequencing

Studies of genetic variation among humans as well as other organisms have already provided us with greater understanding of biological functions and the progression of disease. However, this research has been limited by the high cost of DNA sequencing. An inexpensive technology would allow for widespread sequencing and would revolutionize both life science research as well as health-care (Figure 1.1). The greatest benefit from inexpensive sequencing would be the ability to catalog many sequences and their variations [30]. Amassing genetic information for numerous individuals across species would provide a wealth of information, and allow for studies of genetic

markers for various processes and diseases.

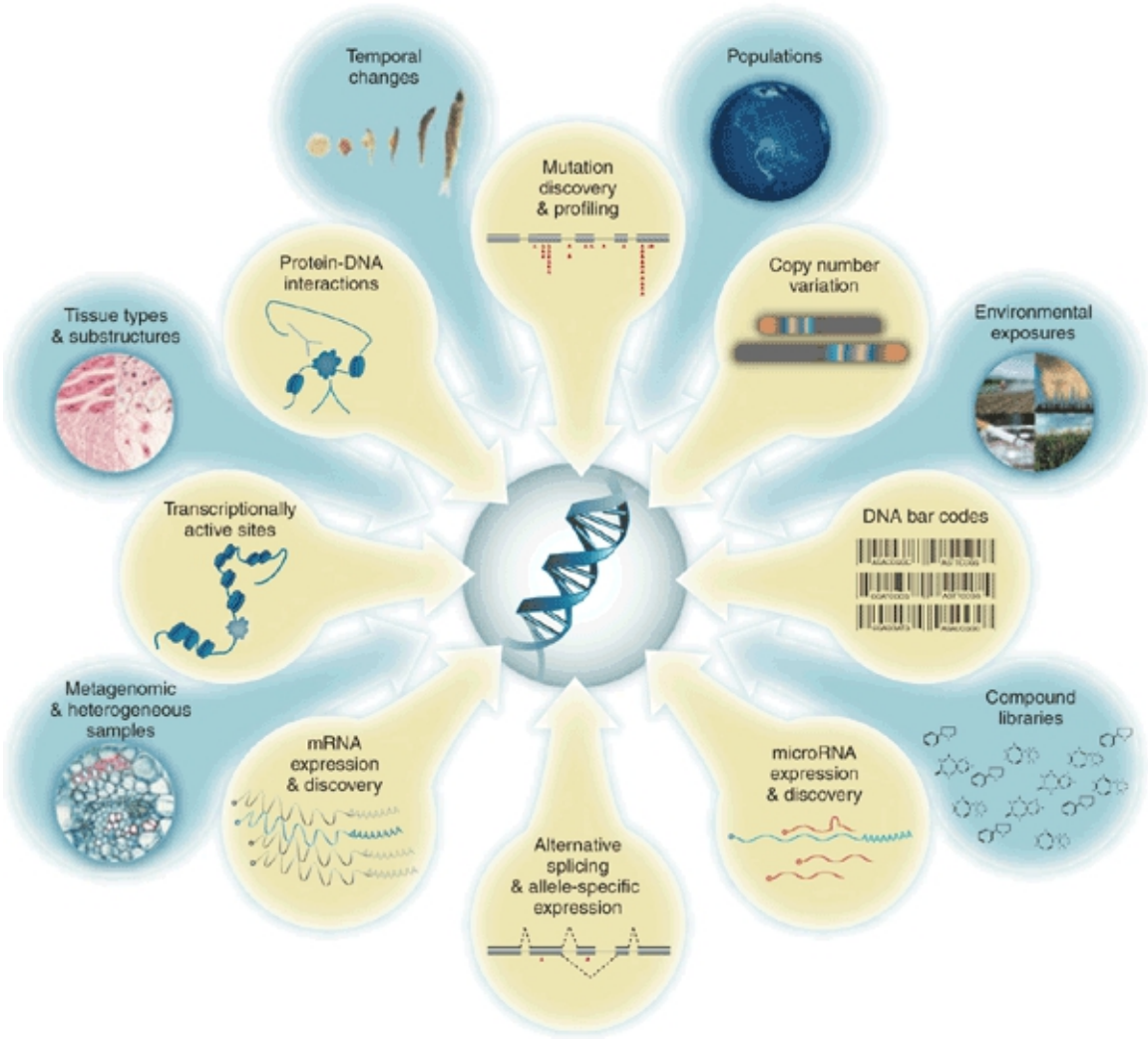


Figure 1.1: **Implications of Widespread DNA Sequencing.** The breadth of information that can be generated from DNA sequencing (yellow) and the variety of sample sources available (blue) are illustrated. Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology [30] ©2008

Discovering new genes and studying those already known will provide insight into their functions and interactions. This will give us understanding of complicated processes including the development of embryos from a single cell, how various biological

components work together to create life, and how the genome influences cellular physiology. Observing genetic characteristics between species will benefit phylogenetics in understanding evolutionary relationships and how various organisms developed.

Widespread sequencing will also provide unprecedented advances to the understanding of human disease. By cataloging the genomic sequences of persons with disease and comparing them to healthy individuals, we can determine genetic markers and identify how disease predisposition occurs. There are several known genetic markers for diseases including single nucleotide polymorphisms (SNP), copy number variations (CNV), and epigenetic modifications. In order to determine the role that these variations play in the development and progression of disease, we must refine our genomic maps and study many more individuals.

The advanced understanding of disease will bring new and more powerful diagnostic tools and treatments. By understanding biochemical processes, we can develop computer models to accurately simulate how cells respond to external stimuli thus assisting in the development of new treatments. Multiple resequencing will allow for studying and monitoring of genomic changes including the growth of tumor cells, evolution of disease, and development of drug resistance. These changes will have a profound effect on the practice of medicine. Sequencing could become a routine part of health-care where doctors may use it as a tool in screening and diagnosis. Treatments would be tailored to an individual's genetic make-up allowing for more effective care.

1.2 Evolution of Sequencing Techniques

In order to allow for widespread sequencing, the cost must be reduced. A goal of $1,000 has been set by the National Human Genome Research Institute (NHGRI) for the cost to sequence a human genome. A multitude of sequencing technologies have been introduced, but none have been able to meet this goal yet. Figure 1.2 shows the cost of sequencing the human genome over the years.

1.2.1 Early Sequencing

The first DNA sequencing methods were developed in the mid-1970s including the Sanger Sequencing method [50], which was the primary sequencing technique used for nearly three decades. In this method, DNA replication was selectively halted after the incorporation of a specific nucleotide type. The lengths of the synthesized strands, then, indicate the positions of that nucleotide type. A complete sequence could be determined from multiple reactions probing the positions of each nucleotide type.

The first draft of the human genome was the result of the Human Genome Project coordinated by the National Institutes of Health and the U.S. Department of Energy. This approximately $3 billion initiative began in 1990 and had contributors including universities from across the United States as well as international partners in the United Kingdom, France, Germany, Japan, and China. When the project was completed in April 2003, sequencing the human genome using the Sanger Sequencing method would take roughly 3-4 months and cost approximately $20 million.

1.2.2 Next Generation Technologies

In 2004, the NHGRI awarded its first grants as part of the Advanced DNA Sequencing Technology Program. The goal of this program was to develop second, or next generation (NextGen) sequencing technologies that could reduce the cost of sequencing the human genome to under $100,000. This program was wildly successful and resulted in multiple technologies that were smaller and less expensive than Sanger Sequencing. The cost to sequence the human genome was reduced to approximately $20,000 and could be completed in under a week (See Figure 1.2). The methods employed in the NextGen sequencers were largely based on sequencing by synthesis where DNA replication was observed to determine the identity of bases being incorporated into a new strand. These methods included observing the incorporation of fluorescently labeled nucleotides or detecting byproducts of DNA synthesis.

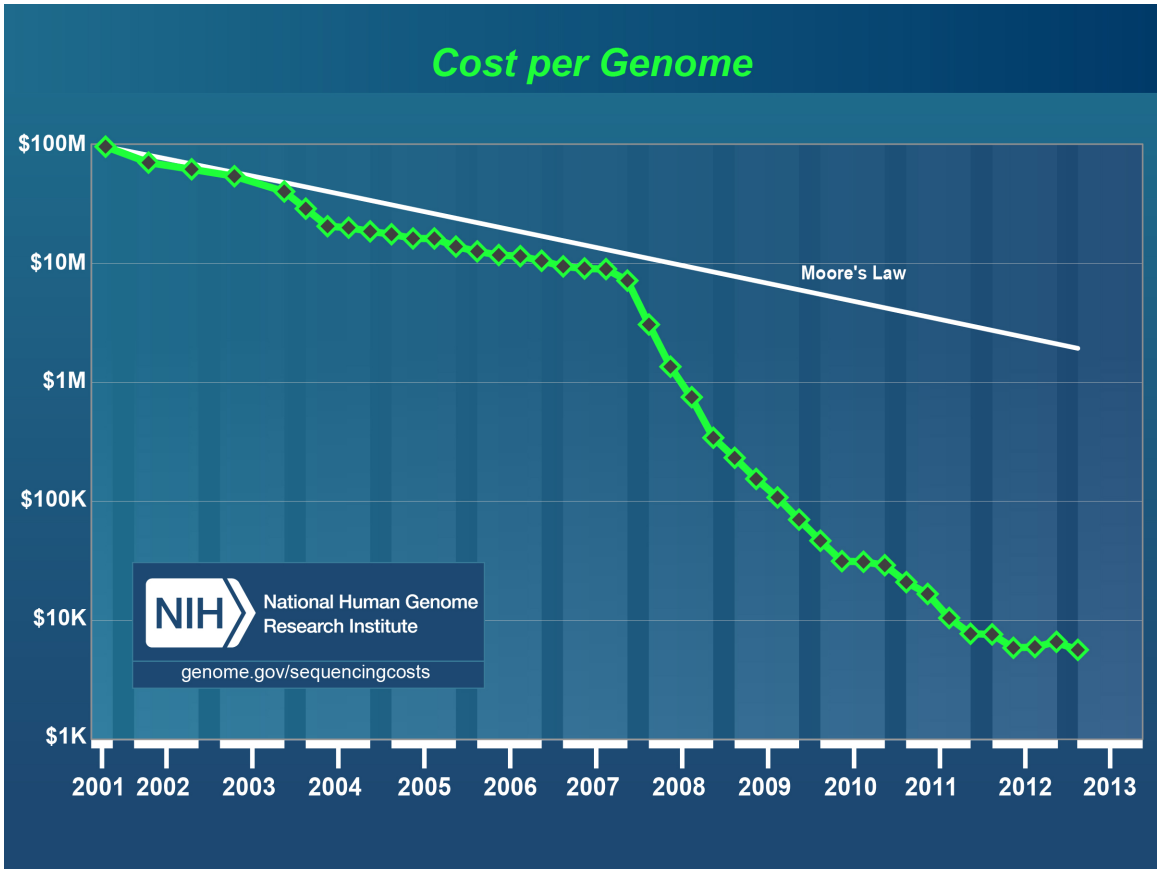


Figure 1.2: **Cost to Sequence the Human Genome.** The cost to sequence a human genome has been reduced dramatically over time. Prior to 2008, sequencing was primarily performed by the Sanger method. In January 2008, sequencing centers transitioned to NextGen technologies, drastically reducing the cost of sequencing. It is expected that third generation sequencers will reduce the cost to less than $1,000.

Beginning in 2009, the NHGRI began supporting research for third generation technologies (3Gen) which will allow sequencing of the genome for less than $1,000. This ultimate goal of $1,000 would allow DNA sequencing to be accessible to a wider population and sequencers would be affordable for individual labs and medical facilities. Much 3Gen research is based on refinement of NextGen sequencers. However, several challenges imped their further reduction in cost [20]. These methods include

significant sample preparation and surface chemistry which can be difficult and require multiple reagents. Additionally, the use of fluorescent labels requires expensive optical instrumentation. Finally, computational power is needed since many techniques require breaking up the genome into smaller sections for sequencing and then reassembling the data. In order to overcome these limitations, a new technology has been introduced that is simple, reagent free, and allows long read-lengths. This new 3Gen technology is called nanopore sequencing.

1.3 Nanopore DNA Sequencing

Nanopore sequencing is based on a technique introduced by Wallace H. Coulter in the late 1940s for sizing and counting particles suspended in an electrolyte solution [13]. In this method, the conductivity across an orifice is measured as a function of time. Particles moving through the channel cause resistive pulses due to their transient blockages. The frequency and amplitude of these pulses can be related to the concentration and size of the particles in the sample. This principle was proposed for use in DNA sequencing in 1996 using a biological porin called α-hemolysin as the orifice [31].

In this method, an external voltage is applied across the nanometer-scale, electrolyte filled pore inducing an electric field. Negatively-charged single stranded DNA (ssDNA) molecules are drawn into the constriction and modulate the ionic current through the pore (see Figure 1.3). It was hypothesized that if each base-type, adenine (dA), cytosine (dC), thymine (dT), and guanine (dG), produces a unique residual current level, the time record of the ionic current could be used to identify nucleotides as they passed through the channel. It was the hope that non-standard bases, such as methylations, would also produce unique current signatures that could be read as part of the sequence. The simplicity of this technique promises inexpensive, fast, direct, and nearly 'reagent-free' DNA sequencing [8]. Additionally, there is no intrinsic limit on the read length of strands. This will overcome a challenge faced by other tech-

nologies which require breaking up the genome into several segments for sequencing and then reassembling the data.

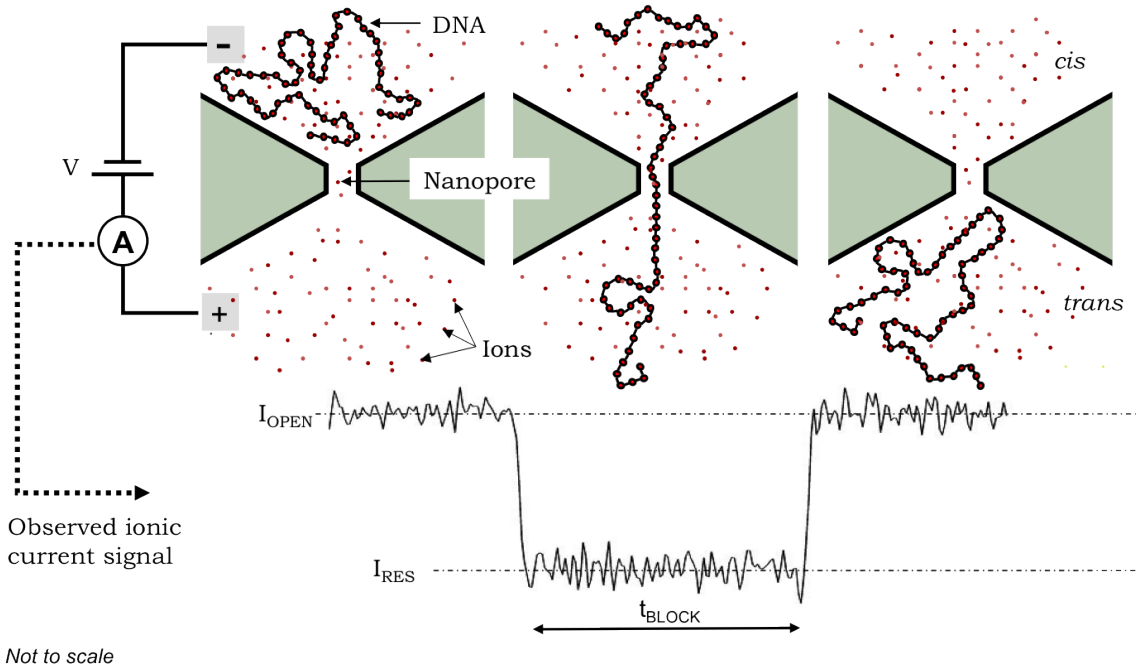


Figure 1.3: **Principle of Nanopore Sequencing.** A voltage is applied across a nanopore in an electrolyte solution resulting in an ionic current, I_{OPEN}. While DNA passes through the pore, the current is reduced. The residual ionic current, I_{RES}, and blockage duration, t_{BLOCK}, are dependent on the composition of the DNA.

The key to realizing nanopore sequencing is identifying a pore that is able to clearly resolve individual nucleotides that pass through it. Natural protein pores are attractive for nanopore sequencing because they are reproducible on the atomic scale, have channel diameters well matched to a variety of biologically important molecules, including DNA, and can be modified at the atomic level to tailor their analytical capabilities. Studies of DNA passage though the biological porin α-hemolysin confirmed the existence of distinct ionic current signatures for each nucleotide type [31, 39, 47, 54, 48]. However, α-hemolysin's long, slender channel results in three

separate locations responsible for modulating the ionic current and produces small differences in nucleotide specific current levels (< 4 pA) [47, 54, 48]. See Figure 1.4 for pore geometry. Additionally, translocation times of DNA passage through the pore were very fast (1-7 μs/base) [31, 39] making it impossible to distinguish individual nucleotide signals from the noisy background [42]. The necessary reduction in DNA translocation speed may be achieved by using modified DNA [15] or a molecular motor [12, 35, 3] to slow or halt the strand within the pore. However, α-hemolysin does not provide the resolution necessary to distinguish the signals from individual nucleotides.

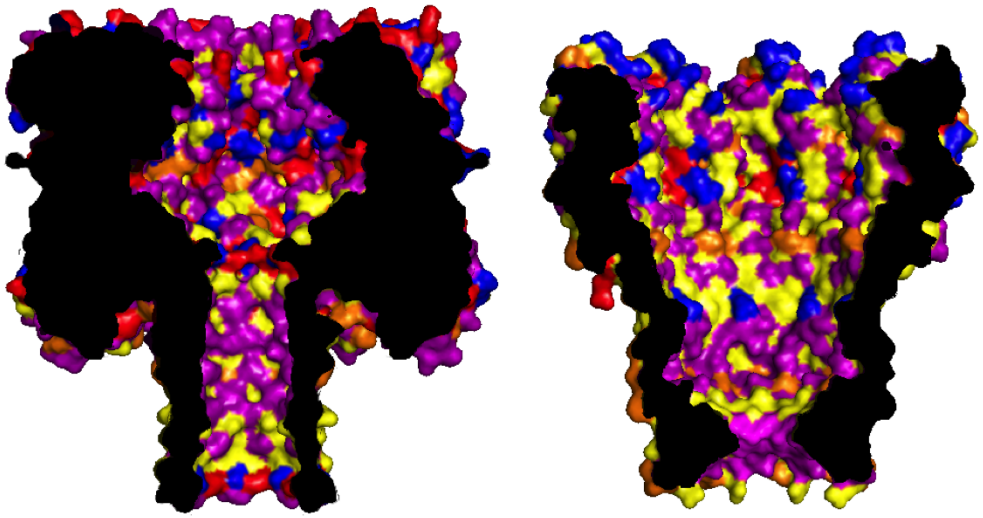


Figure 1.4: **Comparison of Nanopore Channels.** Cross sections of α-hemolysin (Left) and MspA (Right). The long channel at the base of α-hemolysin holds $\sim$12 nucleotides and contains three regions responsible for modulating the ionic current [47, 54, 48]. MspA is shaped like a goblet with a short and narrow constriction.

In 2008, my research group introduced the biological nanopore called *Mycobacterium smegmatis* porin A (MspA) to DNA sequencing. MspA has a short and narrow constriction flanked by much larger regions [10] (Figure 1.4). This geometry was expected to enhance the resolution of nucleotide specific current signatures from DNA passing through the pore.

This dissertation describes the most recent advances towards nanopore DNA sequencing using the MspA protein. In Chapter 2, I outline the procedures used in nanopore experiments and present a brief overview of the experiments described in this dissertation. Chapters $3 - 6$ detail specific studies conducted on MspA. Finally, Chapter 7 discusses the future steps needed to realize nanopore sequencing as a commercial technology.

Chapter 2

USING THE MSPA NANOPORE AS A SINGLE MOLECULE SENSOR

Nanopores may allow sensing of individual molecules passing through their constriction. The nanopore experiments described in this dissertation were conducted on lipid bilayers formed across a horizontal aperture in Teflon. Volumes above and below the bilayer, labeled *cis* and *trans* respectively, contained a buffered salt solution. Voltage was applied across the bilayer and the resulting ionic current was recorded. A single MspA pore was inserted into the bilayer, forming a channel. DNA was added to the *cis* volume and was drawn through MspA to the *trans* side.

In this chapter, I outline the experimental procedures used in the nanopore experiments described in this dissertation including a description of principle biological components. The experiments highlighted in this dissertation are then introduced.

2.1 Experimental Procedure

2.1.1 Experimental Setup

All experiments were conducted on a Teflon 'puck' consisting of two ∼60-100 μl compartments connected by a Teflon tube. See Figures 2.1 and 2.2. One end of the Teflon tube was sealed except for a ∼20 μm diameter horizontal aperture. The volumes above and below the aperture were labeled *cis* and *trans*, respectively. Both compartments and the tube connecting them were filled with experimental buffer. For most experiments, the buffer contained 1.0 M KCl and 10 mM HEPES/KOH buffered at pH 8.0 ± 0.05. A patch clamp amplifier applied a voltage across the aperture and measured the ionic current. A lipid bilayer was formed across the aperture and served

as the substrate for all experiments. Experiments were performed at $23 \pm 1°C$.

2.1.2 Lipid Bilayers

A lipid bilayer is a thin membrane made of two layers of lipid molecules. This membrane creates a barrier for water soluble (hydrophilic) molecules including ions. The lipid molecules forming bilayers consist of a hydrophilic head and two hydrophobic tails. When in water, they may self-assemble into a bilayer with their hydrophobic tails pointing into the center of the bilayer and their hydrophilic heads pointing outward, towards the water. See Figure 2.3.

Forming Bilayers for Experiments

Experimental lipids were purchased from Avanti Polar Lipids. Bilayers were formed with either 1,2-diphytanoyl-sn-glycerol-3-phosphocholine, 1,2-diphytanoyl-sn-glycero-3-phosphate or equal mixtures thereof. They were stored in chloroform at 40-50 μg/ml. The chloroform was allowed to evaporate before use. Prior to experiments, the dry aperture of the Teflon puck was treated with 1-2 μl of a solution of lipid suspended in hexane at about 5 mg/ml. Once the hexane evaporated, experimental buffer was added to the setup. To create a lipid bilayer, a small amount of hexadecene was mixed with dried lipid so that it could be formed into a ball. The ball was placed on the end of a single bristle paintbrush, and lipid was painted around the Teflon aperture. When an air bubble was placed over the aperture, the lipids self-assembled into a monolayer at the air-water boundary. Removal of the air bubble produced a bilayer. Successful bilayer formation separated the *cis* and *trans* volumes, creating seal resistances of >100 GΩ and blocking the ionic current across the aperture. Once a bilayer was in place, MspA was added to the *cis* volume.

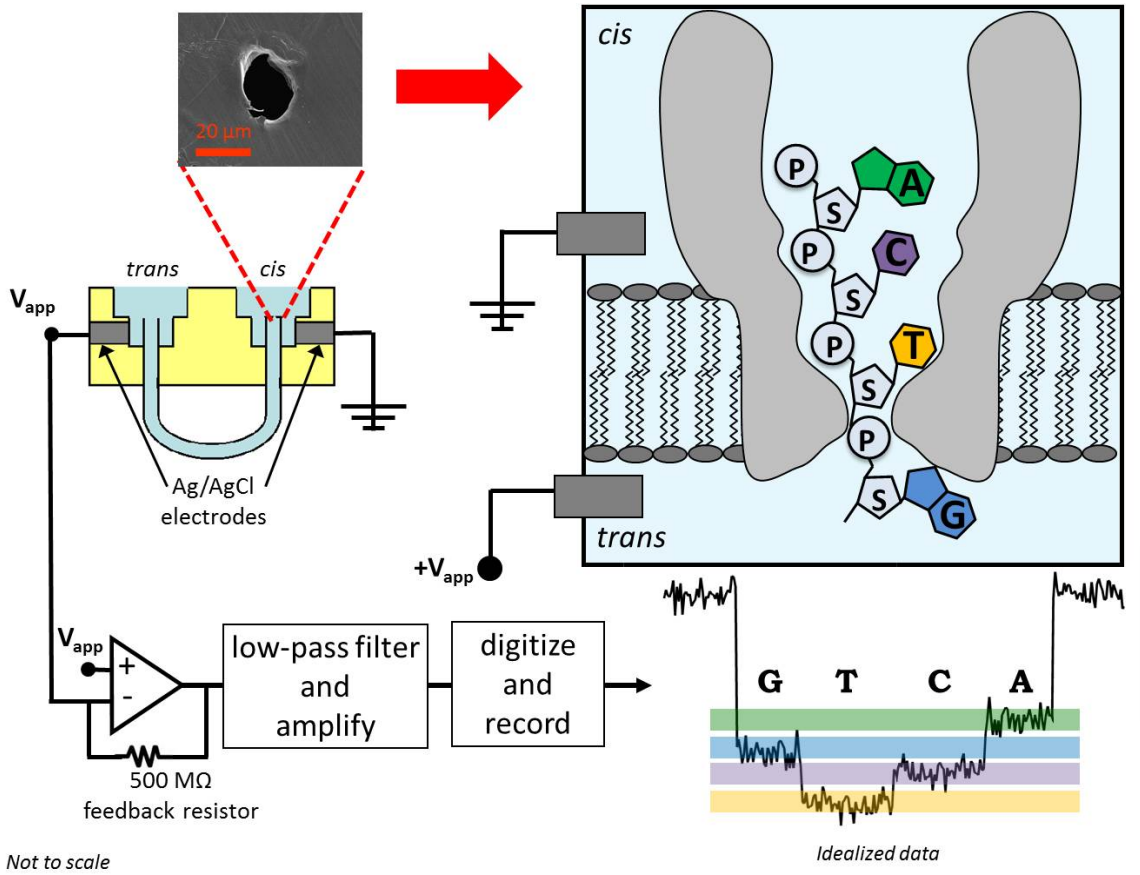


Figure 2.1: **Schematic Diagram of Experimental Setup.** All experiments take place using a Teflon 'puck' (yellow, left). Two wells labeled *cis* and *trans* are connected by a U-shaped Teflon tube. The tube is open on the side extending into the *trans* well but sealed except for a small ($\sim 20\mu$m) aperture in the *cis* well (image of a typical aperture is shown). Both wells and the U-tube are filled with experimental buffer (blue). MspA is set up in a lipid bilayer over the aperture and DNA passes through the pore (right). Electrodes (dark grey) extend into each well and connect the experiment to a patch clamp amplifier, shown as an operational amplifier (lower left). The ionic current signal is low pass filtered and amplified before being digitized and recorded. In the idealized current trace (lower right), each nucleotide type is associated with a different ionic current level.

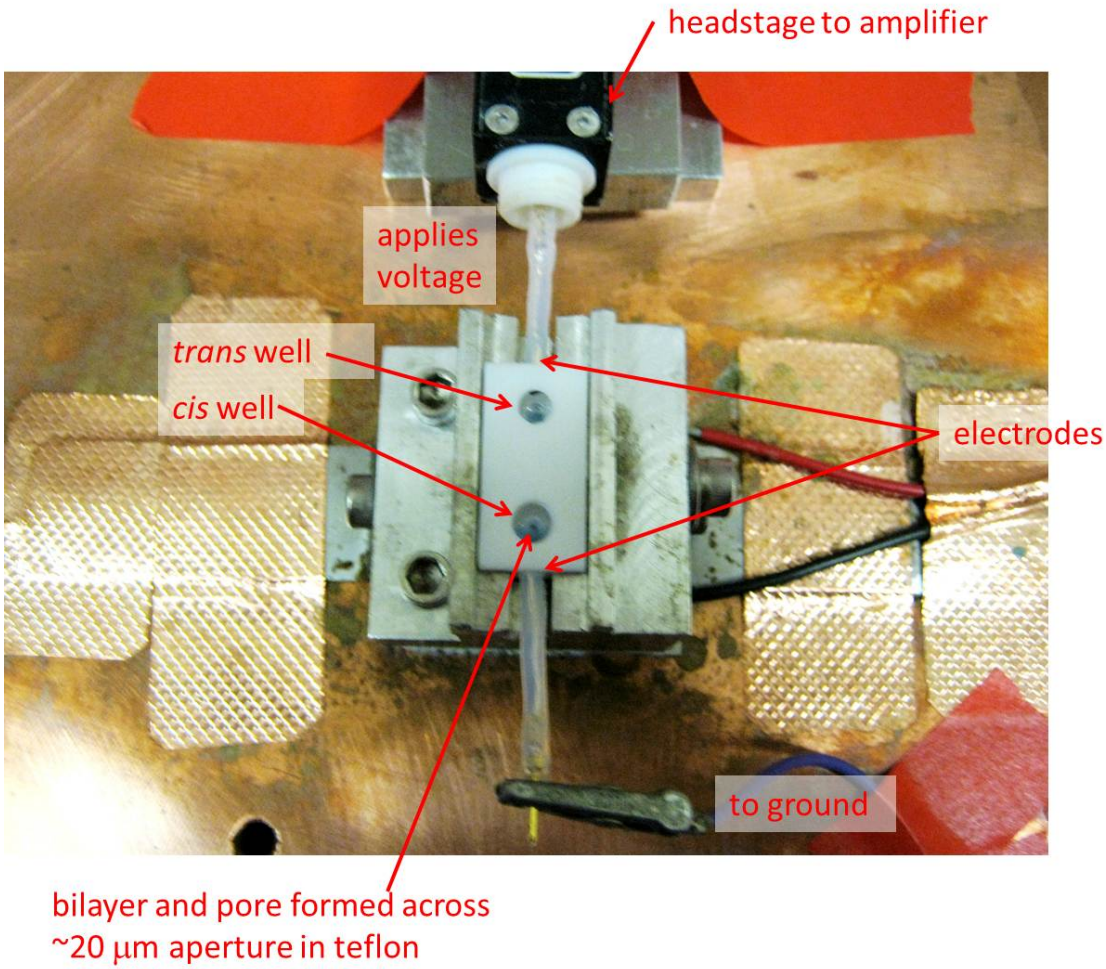


Figure 2.2: **Photo of Experimental Setup.** A Teflon puck is shown (center) with electrodes connected. The *cis* well (lower) is grounded and the *trans* well (upper) is connected to a headstage controlled by the patch clamp amplifier (not pictured).

2.1.3 Mycobacterium smegmatis porin A (MspA)

Mycobacterium smegmatis porin A (MspA) is a channel forming protein derived from *Mycobacterium smegmatis*, a fast growing, non-pathogenic, bacteria that lives in soil, water, and plants. A thick bilayer (∼8 nm [27]) rich with long fatty acids, called mycolic acids, form the bacterium's cell membrane creating a hydrophobic barrier

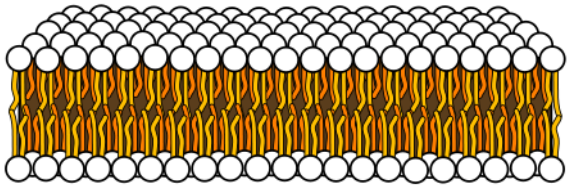


Figure 2.3: **Lipid Bilayer Assembly.** A lipid bilayer consists of 2 layers of lipid molecules arranged such that their hydrophobic tails point inward and their hydrophilic heads point outward.

to the cell interior. Channel forming proteins, such as MspA, are inserted into the bilayer and control passage of hydrophilic molecules into the cell.

MspA is a ∼100 kDa, interconnected, octomeric protein with eightfold rotational symmetry [44, 18]. MspA is ∼10 nm long and shaped like a goblet comprised of three regions. Two hydrophobic β-barrels form the stem and base of the goblet whereas the hydrophilic rim is made of antiparallel β-sheets. The hydrophobic regions promote the insertion of MspA into a lipid bilayer which creates a single, water-filled channel. The central channel varies in diameter between 4.8 and 1.0 nm at the pore constriction [18] (See Figure 2.4).

Each of MspA's eight subunits is made up of 184 amino acids, 134 making up the rim and 50 residues forming the stem and base. The channel constriction is comprised of negatively charged aspartic acids (D) which form a strong electric field reducing the permeability for non-polar, hydrophobic, solutes. The conductance of MspA is influenced by the salt composition of the solution. In 1 M KCl at pH 6.0, MspA has a conductance of ∼4.6 nS [44]. MspA also has selectivity for cations with the ratio of permeability of cations to anions of about 6 [44]. The MspA channel is very robust and is able to maintain channel-forming abilities at a variety of pH values (0-14) and when exposed to heat or denaturing agents like SDS [44, 25].

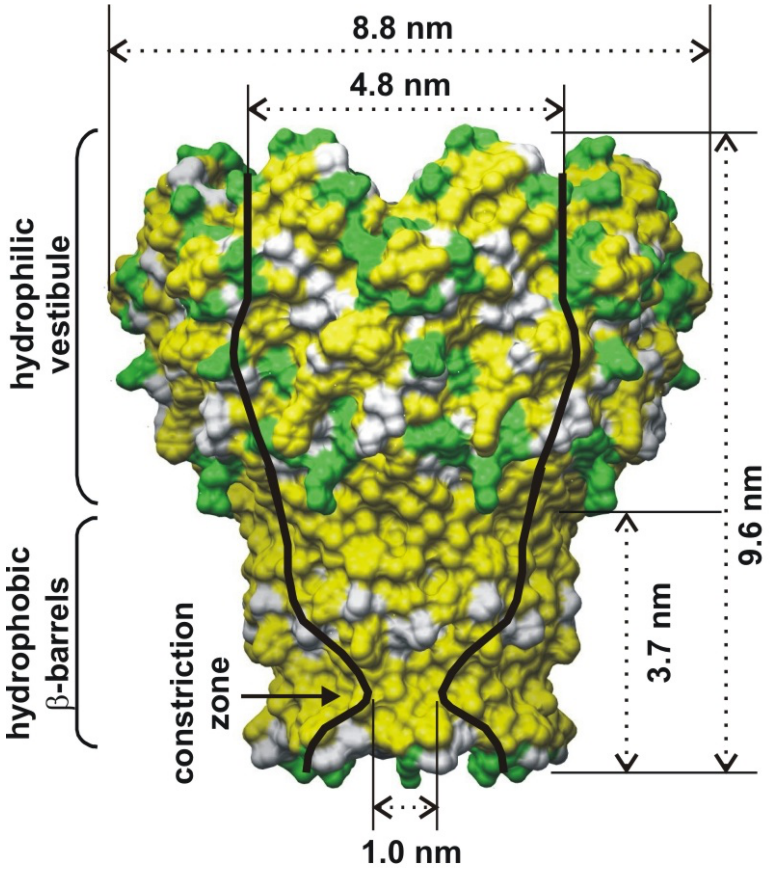


Figure 2.4: **MspA Dimensions and Hydrophobicity.** The dimensions and hydrophobicity of the outer surface of MspA are shown, Green indicates polar, or hydrophilic, regions and yellow indicates non-polar, or hydrophobic, regions. The channel profile is shown in black and varies in diameter between 4.8 and 1.0 nm at the pore constriction.

Engineering MspA for Nanopore Sequencing

The MspA porin used by our lab was prepared by our collaborators, Michael Niederweis and Mikhail Pavlenok, at the University of Alabama in Birmingham. The MspA gene was expressed in *Escherichia coli* as described in ref. [44]. MspA was extracted from *m. smegmatis* cells by boiling them in nonionic detergents at temperatures above

90°C. MspA's superior heat resistance allowed for this selective extraction procedure where most contaminating proteins were removed. MspA was then precipitated by acetone to remove lipids and purified by anion exchange chromatography and gel filtration [26].

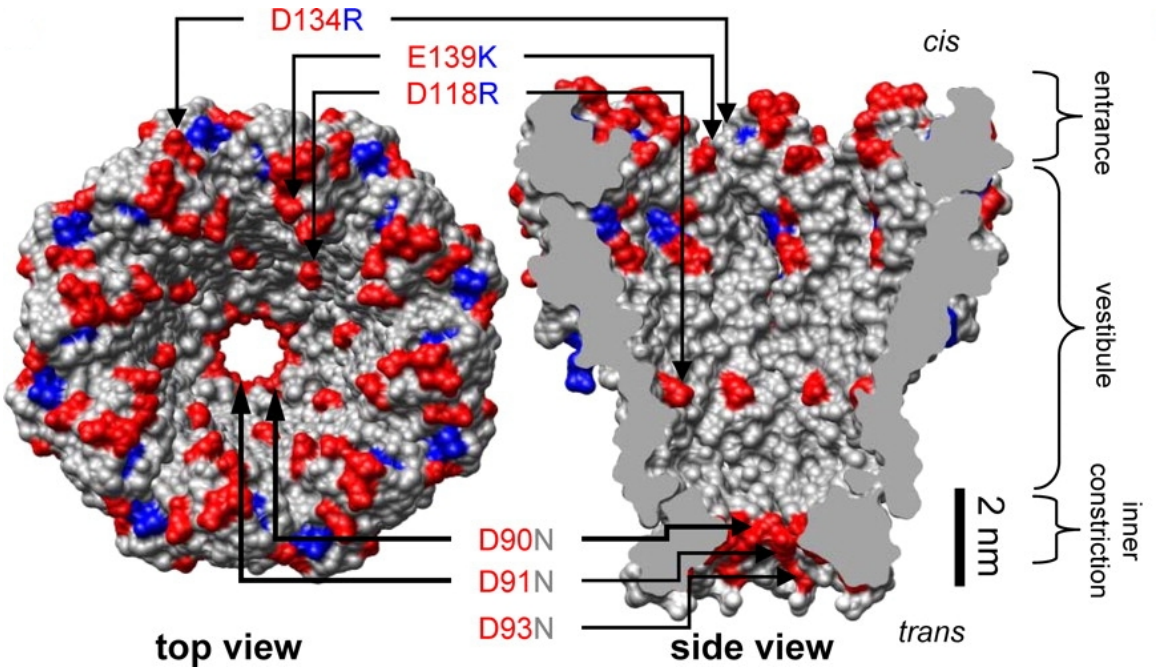


Figure 2.5: **MspA Charge Profile and Mutations.** The structure and charge distribution of wild-type MspA is shown. Negatively charged aspartic acids (D) and glutamic acids (E) are shown in red. Positively charged arginine (R) and lysine (K) residues are shown in blue. Locations of pore mutations are indicated. The M1-MspA pore had mutations only within the pore constriction. The M2-MspA pore contained all mutations indicated in the figure. National Academy of Sciences USA [10] ©2008

When MspA was first used as a channel for nanopore sequencing, there was no observed passage of DNA through the channel [10]. This was attributed to the aspartic acid residues surrounding the constriction which created a barrier for DNA passage. In order for MspA to be used for nanopore sequencing, protein mutations were created using site directed mutagenesis [10]. Two MspA mutants are featured in this

dissertation, M1-MspA and M2-MspA. For both mutants, three aspartic acids (90, 91, 93) were removed and replaced with neutral asparagines (N). The second porin, M2-MspA, has three additional mutations in which negatively charged residues in the rim and stem (134, 139, 118) were removed and replaced with positively charged residues. See Figure 2.5 for mutation sites. Both M1-MspA and M2-MspA allowed DNA translocation but had a reduced conductance, $\sim$2 nS [10] (See Figure 2.6).

Biological porin MspA was added to the *cis* compartment of an experimental setup at a concentration of $\sim$2.5 ng/ml. The M1-MspA mutant was used for all experiments except the enzyme controlled DNA translocation experiments (Chapter 5) where M2-MspA was used. An MspA porin self-assembled into the lipid bilayer causing a $\sim$2 nS increase in measured conductance. Once a single protein inserted, the compartment was flushed with experimental buffer in order to avoid further insertions. The single MspA channel provided the only pathway for DNA molecules to pass from the *cis* to *trans* volumes.

2.1.4 Deoxyribonucleic acid (DNA)

The human genome contains about 3 billion base pairs packaged into 23 chromosomes. DNA normally is in its double-stranded form where two polymer strands are coiled into a helical structure. The backbone of each strand consists of alternating deoxyribose (a sugar) and phosphate. In physiological conditions, each phosphate carries a negative charge of about $1e^-$ causing the strand to be self-repulsive. When in a salt solution, however, this charge is partially shielded allowing for more curvature in the strand. The repeating unit of DNA is a nucleotide (nt) and consists of a phosphate, a sugar and a base. The carbon atoms in the sugar ring are numbered as shown in Figure 2.7. This convention leads to the ends of the DNA being designated as $3'$ or $5'$. Attached to the $1'$ position of the sugars are bases whose identities comprise the DNA sequence. There are four base varieties: adenine, thymine, guanine, and cytosine. Cytosine and thymine are pyrimidines which consist of a single ring structure.

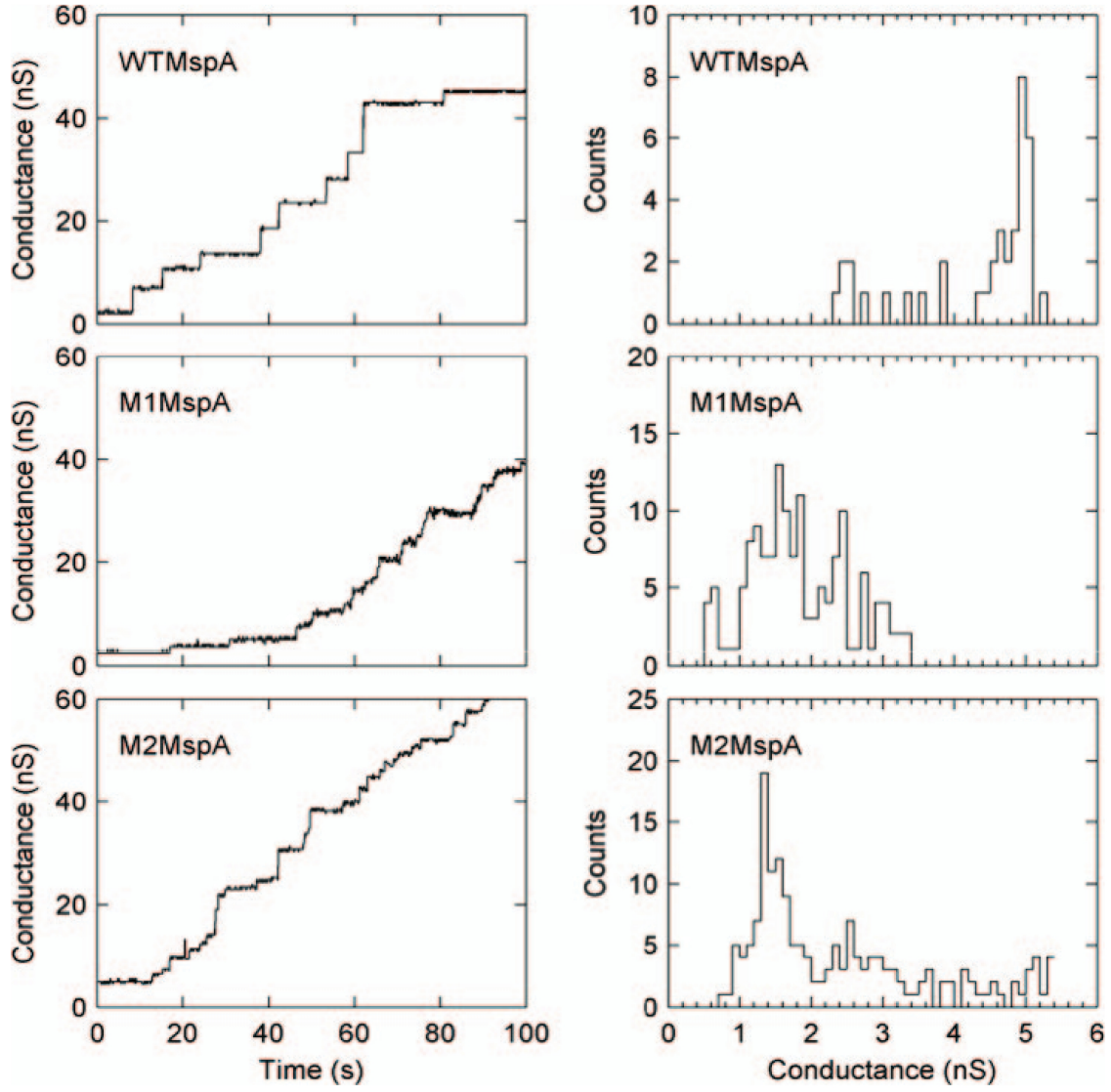


Figure 2.6: **Conductances of MspA Mutants.** Channel-forming activity and single-channel conductances are shown for experiments with Wild Type MspA (WT-MspA), M1-MspA, and M2-MspA (1 M KCl buffer solution). Left: Bilayer conductance over time. Stepwise increases in conductance correspond to insertions of MspA pores into the bilayer. Right: Histograms of conductance step size for WT-MspA (N=3 experiments, n=40 insertions), M1-MspA (N=3, n=144), and M2-MspA (N=5, n=169). National Academy of Sciences USA [10] ©2008

Adenine and guanine are purines which are larger and have a double ring. See Figure 2.8 for structures.

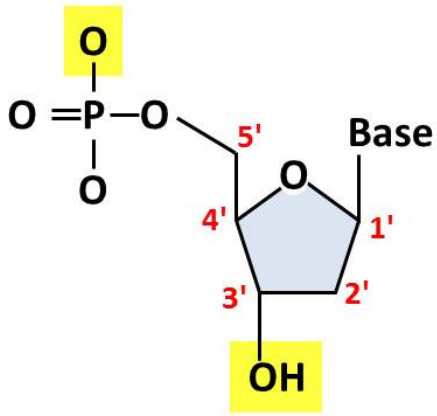


Figure 2.7: **Nucleotide Structure.** The repeating unit of DNA consists of a phosphate, a sugar, and a base. The carbon atoms in the sugar ring are numbered with the DNA base attaching at the $1'$ position. The nucleotides are connected at the highlighted oxygen molecules to form a DNA strand.

The two DNA strands comprising the double helix run in opposite directions where the $3'$ end of one strand is connected to the $5'$ end of the other. They bond via Watson-Crick base pairs (bp) that form between complementary bases on each strand. Adenine and thymine connect with 2 hydrogen bonds, whereas cytosine and guanine attach with 3 hydrogen bonds (See Figure 2.8). The two strands can be dissociated by heating them to approximately 90 $°C$. In addition to Watson-Crick pairings, neighboring bases on the same strand may experience pi-stacking due to hydrophobicity of the chemical rings and four guanines may interact via hydrogen bonds to form a square planar structure called a guanine tetrad (see Figure 2.9). When in its double helix, DNA has a width of approximately 2 nm and rise per base pair of approximately 0.34 nm [58]. A single strand of DNA is approximately 1 nm wide and has an interphosphate distance of approximately 0.6 nm [41].

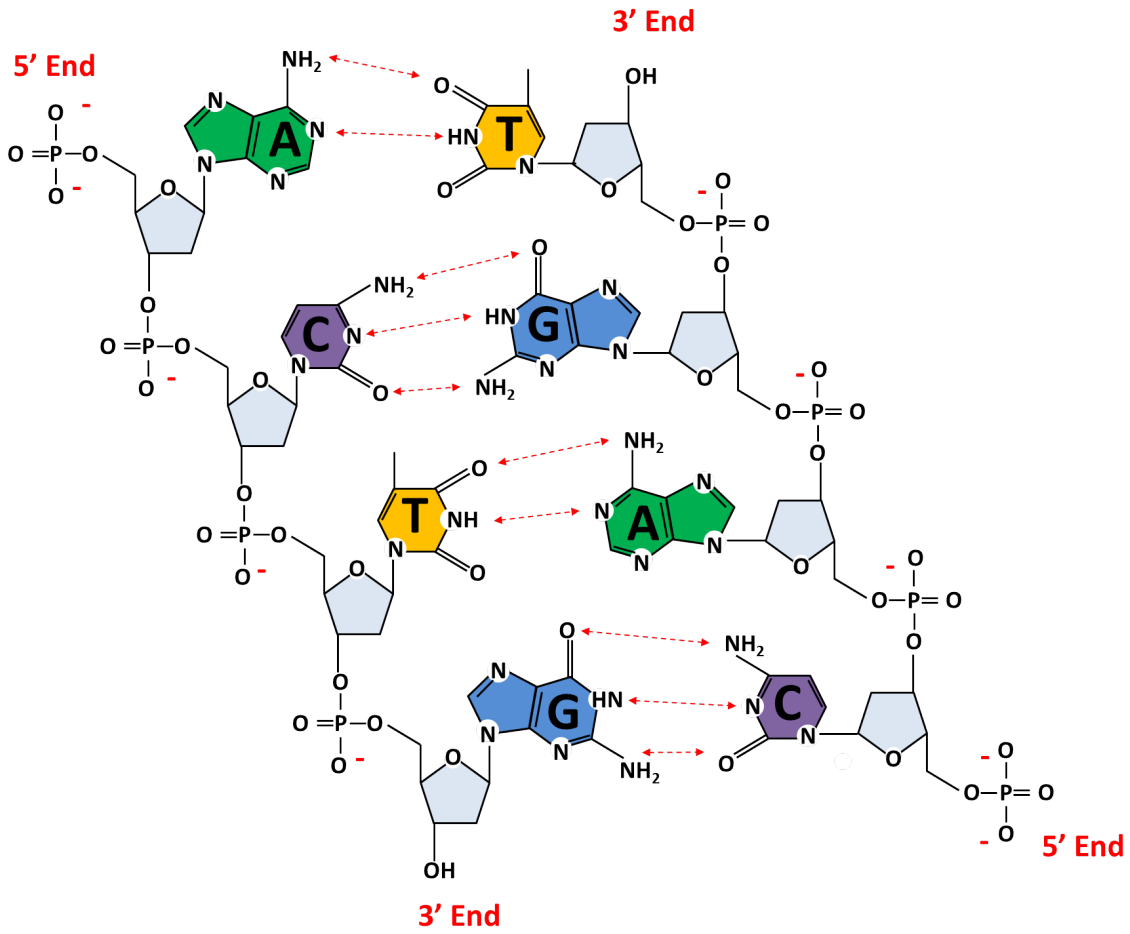


Figure 2.8: **Structure of DNA.** The backbone of each polymer is comprised of alternating sugars (gray) and phosphates. The phosphates are shown with a negative charge and the directionality of the strand (3′ or 5′) is indicated. Each sugar is attached to one of 4 base types, adenine (A, green), cytosine (C, purple), thymine (T, yellow), and guanine (G, blue). The hydrogen bonds between Watson-Crick base pairs are shown as red dashed arrows.

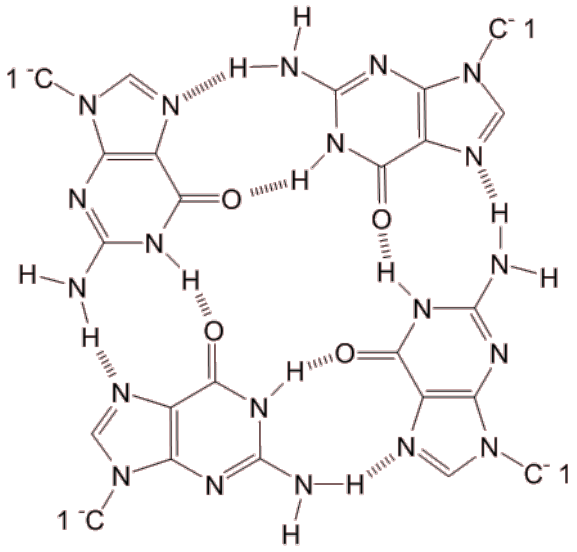


Figure 2.9: **Structure of G-Tetrad.** Four guanine nucleotides may hydrogen bond to from a square planar structure called a guanine tetrad. For this reason, our experimental DNA sequences never contain more than three adjacent guanines.

Experimental DNA

Since MspA's constriction is approximately 1 nm in diameter, it is ideal for sensing single strands of DNA. For most experiments (Chapters 3, 4, and 6), DNA was synthesized by Integrated DNA Technologies (IDT). For enzyme controlled DNA translocation experiments (Chapter 5), DNA was synthesized at the Stanford University Protein and Nucleic Acid Facility and purified at their facility using column purification methods. For all experiments, DNA was stored at $-20°$ C until immediately before use.

Different experiments required modifying a DNA strand or using additional enzymes. Some experiments (Chapters 3 and 5) utilized regions of double stranded DNA. Duplex regions were annealed by incubating complementary strands at $> 90°C$ for $\sim$5 min then gradually cooling them to room temperature. DNA immobilization experiments (Chapters 4 and 6) used a NeutrAvidin molecule to anchor single

stranded DNA in MspA. Enzyme controlled DNA translocation experiments (Chapter 5) involved a DNA polymerase which controlled the passage of DNA through the pore.

2.1.5 NeutrAvidin

NeutrAvidin is synthesized from avidin, a protein found in egg white. NeutrAvidin is a tetramer and has a mass of ∼60 kDa and diameter of 5-6 nm [56, 4]. Each NeutrAvidin subunit has a strong affinity for biotin with a dissociation constant of $K_D \approx 10^{-15}$. Biotin is a small (244.31 g/mol), water-soluble, B-vitamin which may be covalently attached to molecules including proteins and nucleic acids.

Anchoring DNA with NeutrAvidin

For experiments requiring immobilization of DNA within MspA (Chapters 4 and 6), a biotin molecule was attached to the end of single stranded DNA (ssDNA). The biotin was incorporated during DNA synthesis by Integrated DNA Technologies (IDT) and the NeutrAvidin was obtained from Invitrogen. The biotinylated DNA was mixed in equal molar concentrations with NeutrAvidin to create a DNA-NeutrAvidin complex which was stored at -20°C until immediately before use. For experiments, 5 μM of the ssDNA-NeutrAvidin complex was added to the *cis* volume of a setup containing a single MspA pore. The voltage across MspA was held at 180 mV until there was a spontaneous reduction in current to less than 200 pA that lasted longer than 100 ms, signifying DNA immobilized within the pore. Since the diameter of NeutrAvidin is larger than the rim of MspA, it acted as an anchor and prohibited the DNA from fully translocating to the *trans* side. The amplifier then applied 180, 160, 140, 120, 100, and 80 mV for 250 ms per voltage level. The DNA remained immobilized within MspA until -100 mV was applied and the ssDNA was ejected back into the *cis* volume. The voltage then returned to 180 mV. See Figure A.1 of Appendix A for an example event.

2.1.6 DNA Polymerase

A DNA polymerase (DNAP) is the enzyme responsible for replicating DNA. For replication, the DNA molecule is unwound by a helicase enzyme and each strand serves as a template for the production of a complementary strand (Figure 2.10). Generally, a DNA polymerase cannot begin a new strand but it can extend a 'primer' already paired with the template. The polymerase extends the $3'$ end of the primer strand by incorporating nucleotides that are the Watson-Crick match to the template strand (the new strand is formed from $5'$ to $3'$). In order for this reaction to occur, nucleoside triphosphates must be present in solution. A nucleoside triphosphate is a nucleoside (base and sugar) bound to three phosphates and is designated by the identity of its base as dCTP, dATP, dTTP, and dGTP for bases of cytosine, adenine, thymine, and guanine respectively. A new nucleotide is incorporated by cleaving two phosphates from the nucleoside triphosphate and attaching the third to the $3'$ OH group of the preceding nucleotide in the growing primer strand. Additionally, many DNA polymerases have a $3' - 5'$ exonuclease, or proofreading, ability where they can remove a nucleotide from the end of the primer strand. Figure 2.10 shows both extension and proofreading abilities.

DNA Polymerases for Translocation Control

In enzyme controlled DNA translocation experiments, the phi29 DNA polymerase was used to controllably pass DNA through MspA by allowing the polymerase to act on the DNA while it was threaded through the pore (Chapter 5). The phi29 DNA polymerase is a monomeric enzyme ($\sim$66.5 kDa) and has both $5' - 3'$ primer extension and $3' - 5'$ proofreading abilities in the presence of Mg^{2+} [7, 49]. It has long processivity, able to replicate more than 70,000 nt with a single binding event, and can operate against forces of up to 37 pN [5, 29]. Phi29, like other DNA polymerases, has reduced activity in a high salt solution [6, 53].

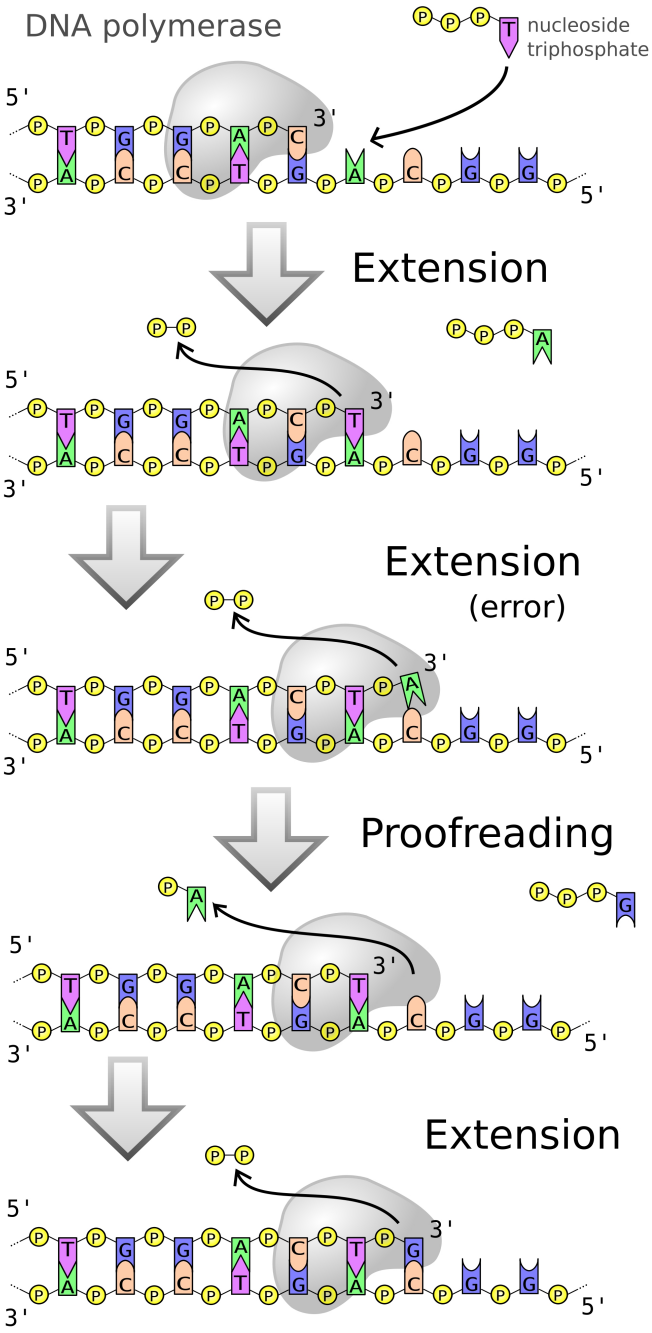


Figure 2.10: **DNA Polymerase Activity.** Extension and proofreading abilities of a DNA polymerase are illustrated. The template strand (lower) is hybridized to a shorter primer strand (upper). Extension: A nucleoside triphosphate that is the Watson-Crick match to the template strand is added to the 3′ end of the primer strand, releasing two phosphates. Extension (error): The DNA polymerase may incorporate an incorrect base. Proofreading: Some DNA polymerases have exonuclease ability where they are able to cleave an incorrect nucleotide from the primer stand before continuing with extension.

For experiments, phi29 DNAP (833,000 U/ml; specific activity 83,000 U/mg) was obtained from Enzymatics and stored at $-20°C$ until immediately before use. Due to its reduced activity at high salt concentrations, experiments involving phi29 DNAP were conducted with experimental buffer of 0.3 M KCl, 1 mM EDTA, 1 mM DTT, and 10 mM HEPES/KOH buffered at pH 8.0 ± 0.05. Phi29 DNAP was added to the *cis* compartment of an experimental setup at a concentration of 0.75 μM. In a standard experiment, 10 mM $MgCl_2$ and 100 μM each of nucleoside triphosphate, dCTP, dATP, dTTP and dGTP, were also added to the *cis* compartment to allow for DNA synthesis.

2.2 Data Acquisition and Analysis

A patch clamp amplifier applied a voltage, V_{app}, across the *cis* and *trans* compartments of an experimental setup and measured the resulting ionic current. Electrical contact was made to the ionic buffer solution in the *cis* and *trans* compartments by Ag/AgCl electrodes. An Axopatch 200B, 1B, or 1C integrating patch clamp amplifier (Axon Instruments) consists of a voltage supply, headstage amplifier, and filter. The headstage of the patch clamp amplifier is a current-to-voltage converter that acts like an operational amplifier with negative feedback. The current from an experimental setup passes though through the amplifier's feedback resistor, R_f, and is given by $I = (V_{app} - V_{out})/R_f$ where V_{out} is the output voltage of the headstage amplifier. This voltage signal from the headstage was low-pass filtered by a 4-pole Bessel filter with a tunable bandwidth of 10-100 kHz. Then the signal was passed to a NI PCI-6259 high-speed multifunction data acquisition device (National Instruments) where it was digitally sampled at a rate of five times the low-pass filter frequency with a precision of 16 bits. The experiment was controlled and the data was collected by custom software written in LabView or LabWindows/CVI (National Instruments).

2.2.1 Noise Sources

There are several sources of noise for this system that range from minor to large enough to obscure a signal. The time resolution depends on the sampling frequency which in turn is dictated by the amplifier's bandwidth. Electronic and experimental noise increase with bandwidth, ultimately limiting our time resolution. The three primary noise sources include Johnson noise, capacitively-coupled voltage noise, and shot noise. The contributions from other noise sources, including Brownian motion of particles, are minor and will not be discussed in this dissertation. Johnson-Nyquist noise is due to thermodynamic properties of the electronics and its one-sided current spectral density is given by $S_I(f) = \frac{4k_BT}{R_f}$. The primary source of Johnson noise is the amplifier's 500 MΩ feedback resistor which results in $S_I(f) = 3.3 * 10^{-5} pA^2/Hz$ at room temperature. The main source of noise for our experiments comes from the capacitance across the lipid bilayer which, at high frequencies, acts as a short circuit for voltage fluctuations. This capacitive coupling produces noise with a spectral density given by $S_I(f) = (2\pi f C_{in})^2 S_V(f)$ [23] where C_{in} is the total capacitance at the amplifier headstage and S_V is the spectral density from voltage fluctuations in the amplifier. Because capacitively-coupled noise increases as f^2, high frequencies are noise dominated. Finally, the noise associated with ions crossing the membrane is shot noise and its spectral density is $S_I(f) = 2Ie$ where e is the electron charge. At $I = 300$ pA, the spectral density for shot noise is $S_I(f) = 9.5 * 10^{-5} pA^2/Hz$. See Figure 2.11 for an example plot of our experimental current spectral density.

2.2.2 Data Analysis

Data was analyzed with custom software written in Matlab (The Mathworks). Blockages were identified by a drop in the ionic current below a threshold of ∼80% of the unblocked current level that lasted longer than a specified duration (dependent on the DNA strand being studied). Minor variations in open-pore current levels were seen

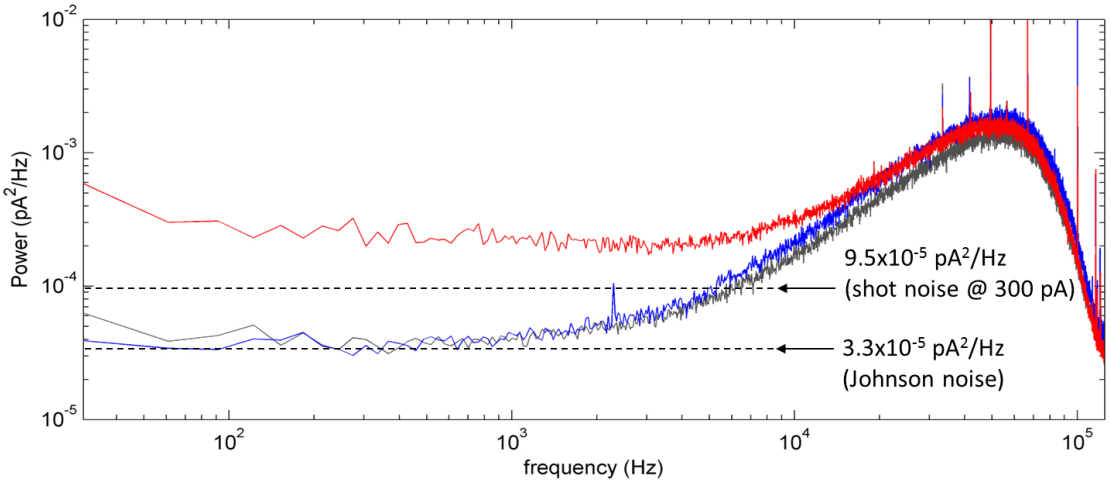


Figure 2.11: **Power Spectral Density for Experimental Setup.** Grey: Open Circuit (no lipid bilayer membrane), Blue: Lipid Bilayer, Red: Single Pore inserted into bilayer. Low frequency noise for electronics and bilayer is consistent with expected Johnson noise from electronics. Above approximately10 kHz, electronic noise is the dominant noise source if no DNA is in the pore.

across a number of experiments and were likely due to minor changes in buffer conditions influencing conductivity. Fluctuations between experiments were minimized by normalizing the residual current for each translocation by the unblocked current level for that experiment. To report values in current, we multiplied the normalized-currents by the average open-pore current for all experiments. To report values in resistance, we divided the applied voltage by the measured ionic current, $R = V/I$. The mean current values and errors reported herein were determined by the peak value and half width at half height of a Gaussian curve fitted to the histogram of mean residual currents for all experiments.

2.3 Experiments Highlighted in this Dissertation

This dissertation highlights nanopore sequencing experiments conducted on the MspA porin. In Chapter 3, I review result from preliminary experiments conducted with

the MspA nanopore. First, the translocation of DNA from the *cis* to *trans* volume is confirmed and unmodified single stranded DNA is passed through the pore. We find that DNA passes too quickly to determine nucleotide specific current levels. Therefore, DNA translocation must be controlled. Duplex regions are then used to interrupt the translocation of DNA through MspA. Since the constriction of MspA is about the width of a single strand of DNA, double stranded sections are not able to translocate. When a double stranded section reaches the pore constriction, the DNA momentarily pauses while the duplex dissociates. Once the duplex region is removed, the single stranded DNA passes to the *trans* side. Using this technique, a variety of DNA strands are probed and the nucleotide resolving ability of MspA is determined. A method of DNA nanopore sequencing is proposed using expanded DNA and multiple duplex regions.

In Chapter 4, DNA is immobilized in MspA using a NeutrAvidin molecule as an anchor, as described above. DNA can be indefinitely held in MspA while the nucleotides residing in the pore constriction are probed. Since the DNA is anchored at the rim of MspA, the ionic current through the constriction is only affected by the target DNA strand. This is in contrast to duplex interrupted translocation where the nucleotides of the second strand also affect the current (Chapter 3). We again determine the nucleotide resolving ability of MspA and the number of nucleotides affecting the ionic current signal. Additionally, we discovered sensitivity to epigenetic modifications and the ability to distinguish single nucleotide polymorphisms in genomic DNA.

In Chapter 5, we use the phi 29 enzyme to control translocation of DNA through MspA. DNA is passed through MspA in both directions (*cis* to *trans* and *trans* to *cis*). To prevent the polymerase from functioning prior to capture on MspA, a blocking oligomer is annealed adjacent to the primer strand shielding its extendable end. When DNA is captured in MspA, it is pulled towards the *trans* side causing the blocking oligomer to dissociate. When the blocking oligomer is completely removed,

the extendable end of the primer strand is exposed and phi29 DNAP is able to begin DNA synthesis. As nucleotides are added to the growing primer stand, the DNA that was threaded through the pore is pulled back towards the *cis* volume. For both directions, we observe steps in the ionic current signal that correspond to single nucleotide movements of the DNA though MspA. These results demonstrate the ability of MspA to be used for nanopore DNA sequencing.

In Chapter 6, we study the elasticity of DNA and the forces applied to molecules within MspA's constriction. We again use the DNA immobilization technique where a NeutrAvidin molecule anchors DNA in MspA's constriction. The voltage across the bilayer is altered and the resistance of the DNA-pore complex is measured. By using a freely jointed chain model, we estimate the elasticity of various DNA stands and determine the forces applied to DNA within MspA's constriction.

This set of experiments demonstrates that MspA is a valuable single molecule tool with the potential for DNA sequencing. Chapter 7 discusses the next steps required for commercialization of a MspA based nanopore sequencer.

Chapter 3

BIOLOGICAL PORIN MSPA FOR NANOPORE SEQUENCING

Mycobacterium smegmatis porin A, MspA, and mutants thereof are biological nanopores currently being investigated for DNA sequencing. It was hypothesized that MspA's geometry, consisting of a single short constriction, would be ideal for nanopore sequencing. In this chapter, I review results from the first experiments conducted on the MspA nanopore. Preliminary results confirm that DNA may thread through MspA reaching the *trans* volume and demonstrate that the translocation speed of unmodified DNA is too fast for single nucleotide resolution. Additional experiments are described where duplex regions were used to temporarily hold a DNA strand in MspA's constriction while current levels were probed.

3.1 Preliminary Results

The first MspA nanopore studies were conducted by Butler et. al. [10] using the experimental procedure described in Chapter 2. To confirm that DNA was able to translocate through the M1-MspA mutant to the *trans* side of the pore, a probe consisting of single stranded DNA (ssDNA) attached to a large NeutrAvidin molecule was added to the *cis* volume and short complementary oligos were in the *trans* volume [10]. When the probe threaded through MspA and reached the *trans* volume, the oligos hybridized to the end of the probe locking it in the pore. This was confirmed by a negative voltage required to dissociate the oligo from the ssDNA probe and clear the blockage. See Figure 3.1.

In additional experiments, both mutants, M1-MspA and M2-MspA, showed in-

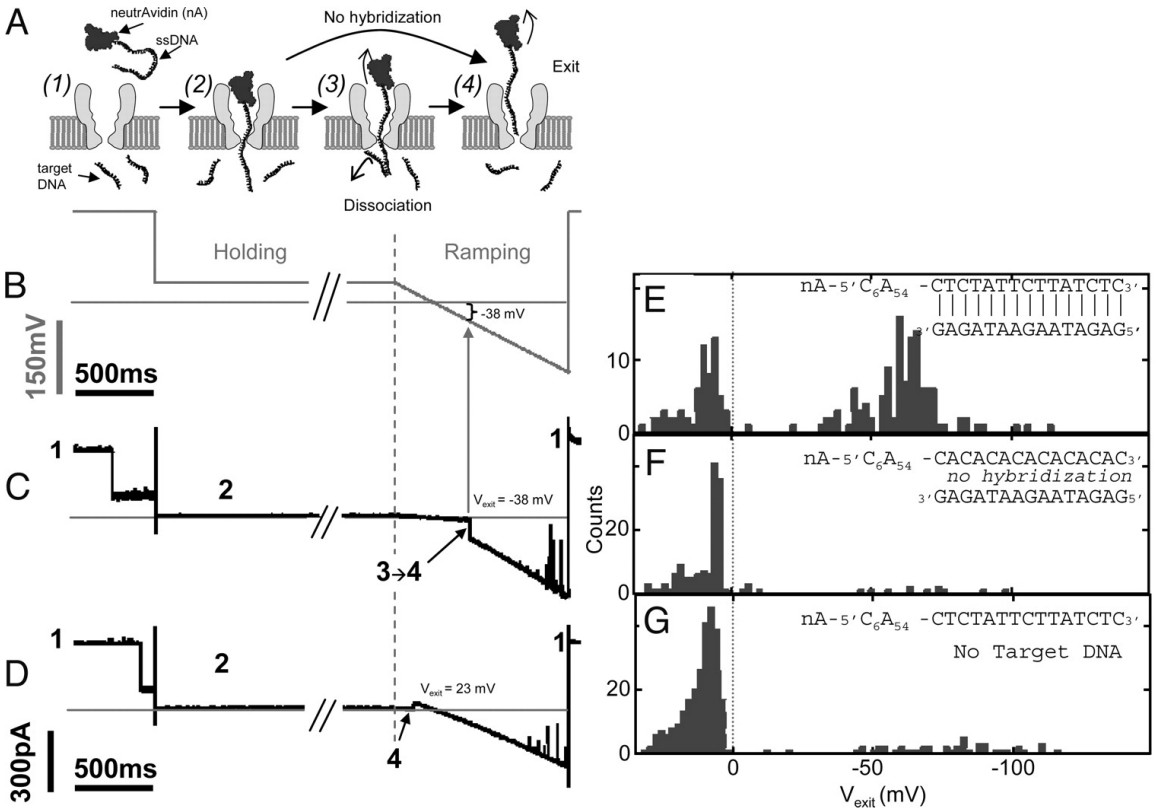


Figure 3.1: **Confirmation of DNA Translocation.** (A) Cartoon of experiment (1) An unblocked pore. (2) ssDNA immobilized in MspA with a NeutrAvidin anchor. (3) Complementary oligos hybridized to the ssDNA probe dissociate with negative voltage. (4) The ssDNA-NeutrAvidin probe exits the pore. (B) Time series of applied voltage. A current blockade triggers a change from the 180 mV capture voltage to a holding voltage of 40 mV. The holding voltage is maintained for 5 s to allow hybridization, and is then ramped negatively. (C and D) Current time series demonstrating DNA exit at negative and positive voltages, respectively. Large current spikes occur because of instantaneous voltage changes and spontaneous pore closure at large negative voltage. (E − G) Exit voltage (V_{exit}) histograms. (E) Oligo is complementary to the ssDNA probe. (F) The oligo is not complementary to the probe. (G) Control experiment with no oligo present in the *trans* volume. A negative V_{exit} peak is only observed in E, where the oligo was complementary to the probe. The control experiments show that the negative V_{exit} peak observed in E is not caused by nonspecific binding. National Academy of Sciences USA [10] ©2008

teraction with homopolymer ssDNA consisting of 50 thymine nucleotides (dT_{50}). At an applied voltage of 180 mV, the addition of 8 μM dT_{50} to the *cis* volume of an M1-MspA experiment resulted in $\sim$5 blockages/sec. At the same voltage, 2μM of dT_{50} added to a M2-MspA experiment led to 25 blockages/sec; a 20x higher interaction rate. The structure of a M1-MspA blockage was a downward spike whereas the translocations through the M2-MspA mutant showed the pattern of a $\sim$100 μs partial blockage followed by a downward spike. See Figure 3.2. Partial blockages were attributed to the additional positive charges in M2-MspA's rim and stem. These likely caused the DNA to pause in the pore vestibule prior to translocation to the *trans* side, which is indicated by the downward spike. For both mutants, the DNA translocation speed was faster than 0.6 μs/nt preventing single nucleotide identification.

3.2 Duplex Interrupted Translocation of DNA through MspA

The experiments described in this section were led by Ian Derrington and were the first MspA nanopore experiments in which I was involved. The work was published in the September 14, 2010 issue of Proceedings of the National Academy of Sciences of the United States of America (PNAS) [15] and was part of Ian Derrington's dissertation [14].

To explore MspA's nucleotide sensitivity, we temporarily halted DNA translocation through the pore using duplex regions. Since MspA's constriction is only wide enough for a single strand of DNA, double-stranded regions cause the translocation to temporarily pause until the duplex dissociates. While the DNA is halted by the duplex, nucleotides residing in MspA's constriction can be studied. Once the duplex region is removed, the single stranded DNA can continue translocating through the pore. Using this technique, we found that the residual ionic current through the M1-MspA mutant, herein designated MspA, was distinguishable for each base type, adenine (dA), cytosine (dC), thymine (dT), and guanine (dG) and individual nucleotides

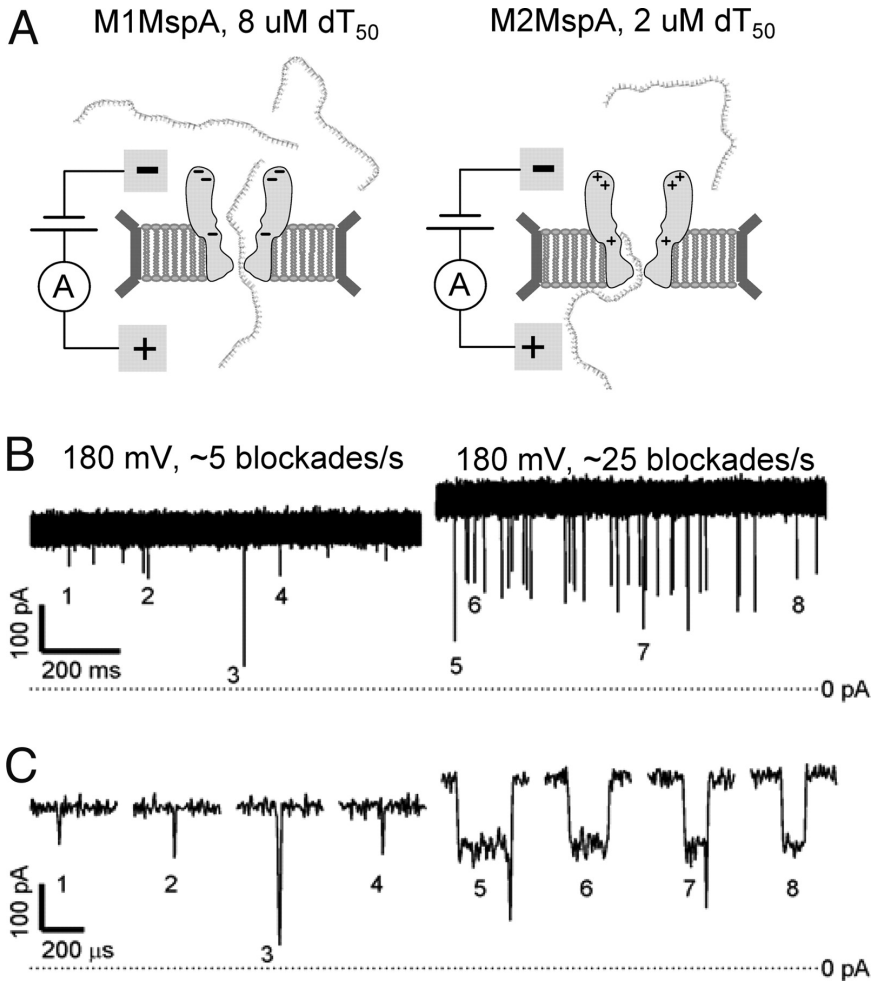


Figure 3.2: **DNA Interaction with MspA Mutants.** Comparison of dT_{50} homopolymer blockades for M1-MspA and M2-MspA. (A) Schematic diagram of experiments. (B) Representative ionic current signals observed for M1-MspA with 8 μM dT_{50} (Left) and M2-MspA with 2 μM dT_{50} (right). (C) Numbered blockades from traces in B shown at expanded time scales. National Academy of Sciences USA [10] ©2008

within a homopolymer strand could be detected. Expanding on this technique, we introduced a novel sequencing method where multiple double-stranded sections were inserted between each nucleotide of the DNA to be sequenced.

3.2.1 Homopolymer DNA

To study duplex interrupted DNA translocation, single stranded DNA was designed such that it formed a 14-base pair (bp) hairpin at one end and had a 50 nucleotide (nt) single stranded DNA (ssDNA) 'tail'. When DNA was captured in MspA under an applied voltage of 180 mV, the single stranded tail threaded through the pore constriction until the duplex region halted further translocation. After ∼10 ms, the hairpin duplex dissociated and the DNA passed to the *trans* volume. While the DNA was held in the pore, the residual ionic current was measured. See Figure 3.3 for a cartoon of the experiment.

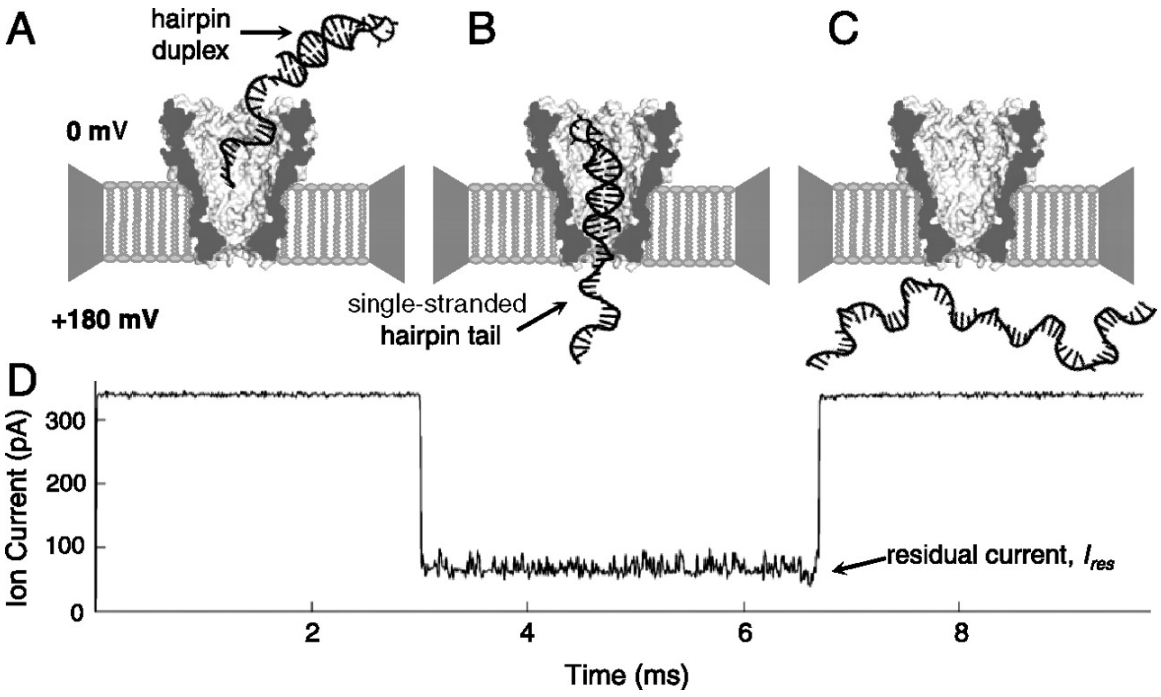


Figure 3.3: **Duplex Interrupted DNA Translocation Through MspA.** A cartoon showing duplex interrupted DNA translocation through MspA and the resulting residual current. (A) Hairpin DNA enters the pore. (B) DNA translocation is temporarily halted by the wide hairpin duplex. (C) The duplex dissociates and the DNA completes translocation to the *trans* volume. (D) The resulting current trace shows the residual current, I_{res}, associated with the hairpin DNA residing in the pore.

The current signatures for each of the four base types were found using 'homopolymer' ssDNA tails, $(dA)_{50}$, $(dC)_{50}$, $(dT)_{50}$, and $(dG)_3(dA)_{47}$. Due to G-tetrad formation (Figure 2.9), $(dG)_3(dA)_{47}$ was used rather than homopolymer guanine. The residual ionic current associated with each of these homopolymer tails was recorded for both DNA orientations, $5'$ and $3'$ ends leading (see Figure 3.4). In both orientations, the current signatures for the homopolymer DNA tails were well resolved with differences in residual currents of up to $\sim$ 24 pA. We also found that the larger purine bases, dA and dG, were associated with a larger residual current than the smaller pyrimidines, dT and dC. This effect has been noticed in α-hemolysin as well [2] and indicates that volume restriction is not solely responsible for the measured residual ionic current through the pore.

3.2.2 Region of Sensitivity

To determine which nucleotide positions most influence the residual ionic current, we used a series of hairpin DNA containing four dC nucleotides substituted at various positions in an otherwise poly-dA tail (see Figure 3.5). When the substitutions were adjacent to the hairpin duplex, the currents were the same as those found above for poly-dC. When the dC substitutions were more than three bases away from the duplex, the residual current resembled that of poly-dA. Since the vestibule of MspA is wider than double stranded DNA (Chapter 2), it is expected that the DNA hairpin sits near the pore constriction. Therefore, only the $\sim$3 nt beside the hairpin reside in MspA's constriction and affect the ionic current through the pore.

3.2.3 Single Nucleotide Recognition

We next explored whether MspA was able to distinguish current levels from single nucleotides. A single dC substitution was made in an otherwise poly-dA ssDNA tail. The dC substitution was positioned at the first three nucleotides after the hairpin (see Figure 3.6). For all three positions, the residual ionic current was intermediate to that

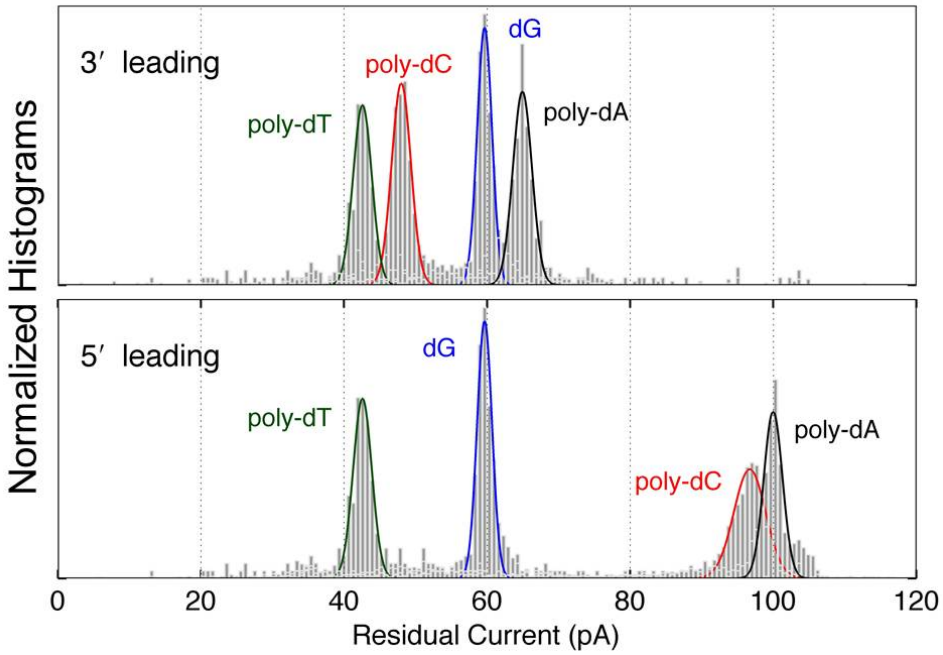


Figure 3.4: **Experimental Results for Hairpin DNA with 'Homopolymer' Tails.** Histograms of the average residual ionic current is shown for 'hompolymer' ssDNA tails with a 14 base pair hairpin (hp). Both (A) 3′ leading and (B) 5′ leading orientations are shown. Experiments were conducted at an applied voltage of 180 mV.

of poly-dA and poly-dC indicating that the single nucleotide substitution produced an observable effect on the residual ionic current. The greatest deviation of the ionic current from that of poly-dA occurred for the first two nucleotides after the hairpin indicating that these nucleotides principally affect the ionic current through the pore.

We also studied the effect the homopolymer background played in the ability to recognize individual nucleotide substitutions. We studied various hairpin DNA strands consisting of a single nucleotide substitution in an otherwise homopolymer tail. The identity of the nucleotides comprising the homopolymer tail was varied. We found that the influence of the single nucleotide substitution on the residual ionic

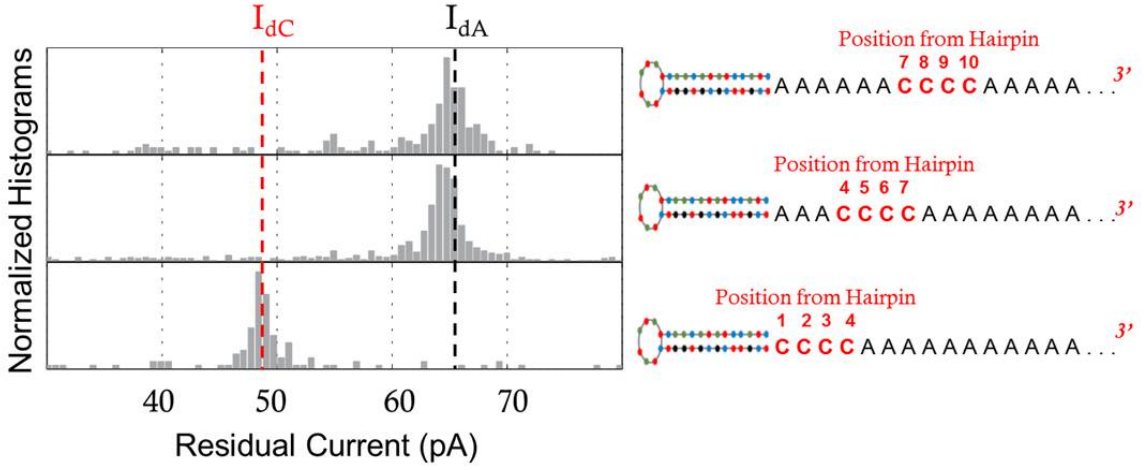


Figure 3.5: **Determining the Region of Sensitivity of MspA with Hairpin DNA.** Histograms of the average residual ionic current is shown for ssDNA tails with a 14 base pair hairpin (hp). Each tail contains poly-dA with four dC nucleotides substituted at various positions along the strand. When the substitutions were located more than 3 nt from the hairpin (top and middle), the residual current was similar to that of poly-dA (vertical black dashed line) indicating that the dC nucleotides do not affect the ionic current. When the dC substitutions were within the first three nucleotides from the hairpin (bottom), the residual ionic current resembled that of poly-dC (vertical red dashed line).

current was strongly dependent on the identity of the surrounding homopolymer nucleotides. A possible reason for this phenomenon is that sequence dependent differences in the elasticity of DNA change the seating of nucleotides in the constriction (see Chapter 6).

3.2.4 Effect of Hairpin Terminus

Because the hairpin duplex rests near MspA's constriction, it is possible that the ionic current is affected by the hairpin in addition to the first three nucleotides of the ssDNA tail. To determine the extent of this contribution, we studied DNA hairpins in which the terminal base pair or the length of the duplex region was altered (see

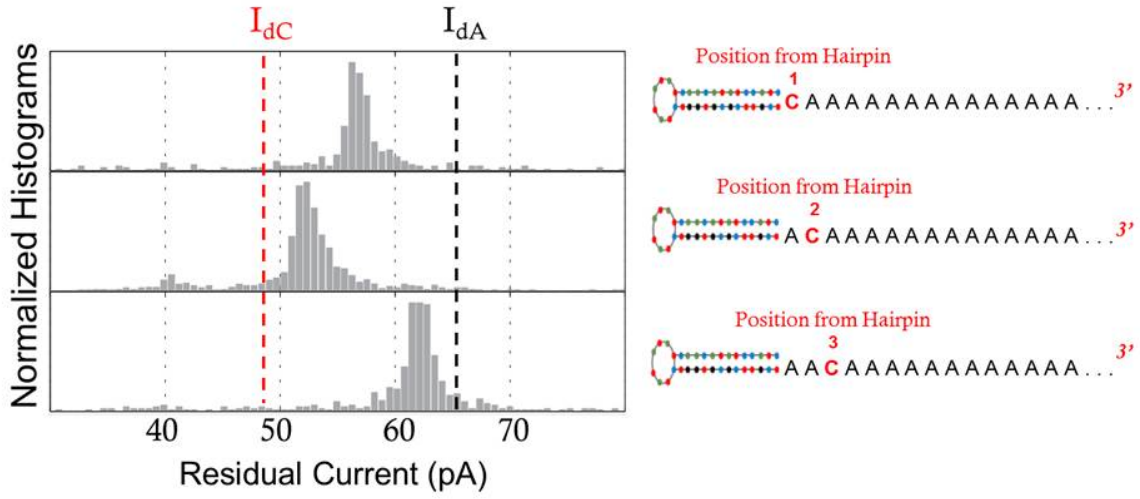


Figure 3.6: **Experimental Results for Single Nucleotide Substitutions in Hairpin DNA with a poly-dA Tail.** Histograms of the residual ionic currents associated with a single dC nucleotide substitution in an otherwise poly-dA hairpin tail. The dC substitution was positioned at the first (top), second (middle), and third (bottom) nucleotide from the DNA hairpin. For comparison, the residual current levels associated with poly-dA and poly-dC are shown as black and red vertical dashed lines, respectively. For all three positions, the residual current was intermediate to that of poly-dA and poly-dC. The first 2 nucleotides after the hairpin had the greatest effect on the ionic current.

Figure 3.7). We found that the residual ionic current was strongly dependent on the identity of the terminal base pair but only weakly dependent on the length of the hairpin duplex. These results indicate that the effect of the hairpin duplex on the residual ionic current cannot be neglected.

3.2.5 Duplex Interrupted (DI) Nanopore Sequencing

In this chapter, we have shown that a duplex region can temporary pause translocation of DNA through MspA such that the nucleotides adjacent to the duplex can be probed. Additionally, we have demonstrated that each nucleotide type blocks at a unique level and that only the ∼3 nucleotides after a hairpin duplex effect the measured ionic

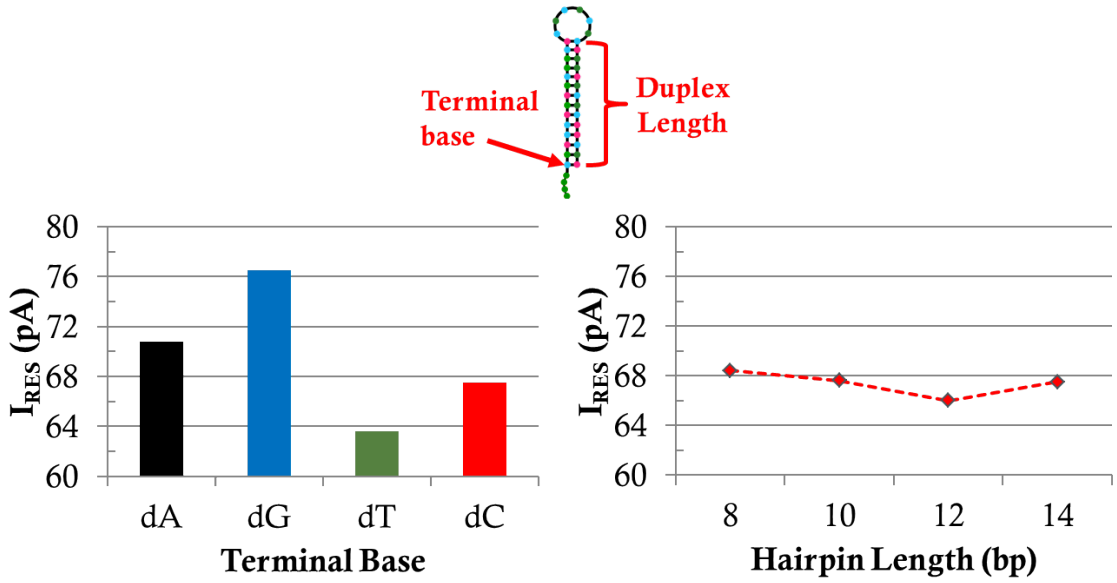


Figure 3.7: **Effect of Hairpin Composition on Residual Ionic Current.** A DNA hairpin consists of several nucleotides which bind to form a duplex region at the end of ssDNA. To determine the effect the hairpin has on the residual ionic current, the identity of the terminal hairpin base (left) and the length of the duplex region (right) were varied as the ssDNA tail remained the same. The residual ionic current showed a strong dependence on the identity of the terminal base whereas the length of the duplex region only had a minor effect.

current. We can expand upon this technique to allow for reads of multiple DNA nucleotides.

Multiple duplex regions can be inserted into a DNA strand to slow the translocation of DNA through the pore. An unknown strand of DNA can be modified to contain short duplex regions between each unknown nucleotide [33, 34]. When the DNA is passed through MspA, it will temporarily halt at each duplex region. While paused, the nucleotide in MspA's constriction can be read. After a short time, the duplex will dissociate and the DNA will continue to translocate until the next duplex region is encountered. This method is called Duplex Interrupted (DI) nanopore

sequencing.

We provide proof of principle of DI sequencing by reading three DNA sequences each consisting of four or five nucleotides, 3′-ATGC-5′, 3′-TACG-5′, and 3′-GTCAC-5′. For each sequence, a 14 bp duplex region was inserted between each nucleotide of interest. To compare the ionic current levels with results found previously for homopolymer strands, the duplex regions were the same as for the hairpin DNA studied above and trinucleotides were used instead of single nucleotides in the sequence of interest. For example, the sequence 3′-ATGC-5′ was converted to 5′ duplex-CCC-duplex-GGA-duplex-TTT-duplex-AAA-dA_{32} 3′.

For each of the three DNA sequences, multiple discrete steps in the ionic current were observed that corresponded to the nucleotides of interest being temporarily held in the pore constriction. Figure 3.8 shows example current traces recorded at 140 mV and histograms of the average current for each level. The levels of these steps were the same as those found previously for homopolymer hairpin DNA.

3.3 Conclusions

These results demonstrate the MspA is a viable pore for nanopore sequencing and has single nucleotide sensitivity. We have also shown that modified DNA can be read using multiple duplex regions to control its translocation through the pore. However, this conversion process is complicated. It is the goal of our reseach to introduce a sequencing method that uses unmodified DNA. Additionally, since the identity of the terminal bases of the duplex region have a measurable effect on the residual ionic current, other methods must be introduced to hold DNA in MspA's constriction. In Chapter 4, we use a large NeutrAvidin molecule to immobilize DNA in MspA and probe individual nucleotides without the interference of nearby duplex regions.

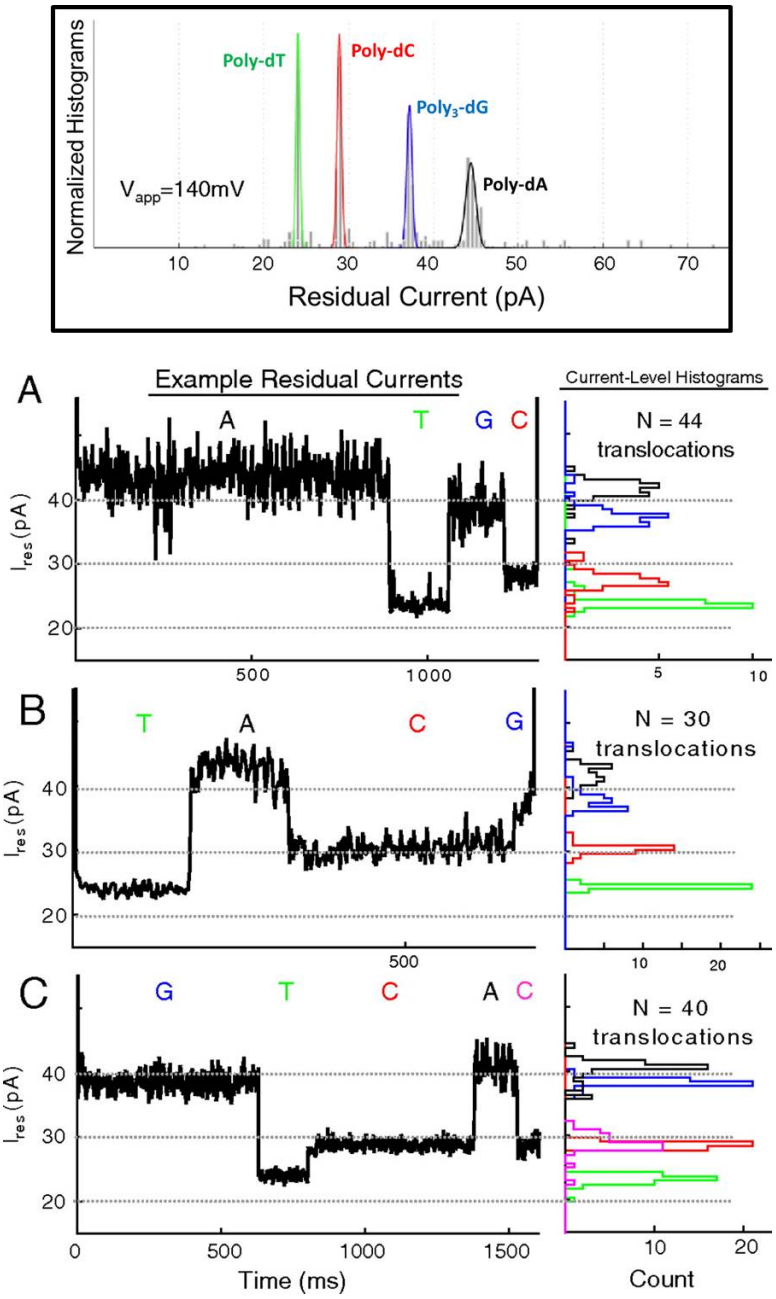


Figure 3.8: **Demonstration of Duplex Interrupted (DI) Nanopore Sequencing.** We use synthesized DNA to have duplex regions between information carrying nucleotides. Each of these duplexes must be sequentially dissociated as the DNA passes through the pore, enabling the residual current to determine the sequence. We provide proof of principle of this technique using DNA chosen to represent different sequences (A) 3′-ATGC-5′, (B) 3′-TACG-5′, and (C) 3′-GTCAC-5′. Example traces are shown (left). Each step in the residual current is representative of three nucleotides within MspA's constriction held by a 14 bp DNA duplex. For each of these steps we generated a histogram of each level (right) for N translocations. These results were generated at 140 mV. For reference, histograms of the residual ionic currents for homopolymer tails at 140 mV are shown (top). At higher voltages, the number of translocations increased, but the level specificity decreased due to reduced time averaging.

Chapter 4

NUCLEOTIDE DISCRIMINATION WITH DNA IMMOBILIZED IN MSPA

This work was published on October 4, 2011 in PLOS ONE [37].

In the previous chapter, we demonstated that current differences from single nucleotides were observable with duplex interrupted translocation through MspA. However, the duplex region of the DNA strand appreciably affected the signal. Here we study MspA's ability to resolve nucleotides when single stranded DNA is held within the pore using a biotin-NeutrAvidin complex. We show that homopolymers of adenine, cytosine, thymine, and guanine in MspA exhibit large current differences and methylated cytosine is distinguishable from unmethylated cytosine. We establish that single nucleotide substitutions within homopolymer single stranded DNA can be detected when held in MspA's constriction. Using genomic single nucleotide polymorphisms, we demonstrate that single nucleotides within random DNA can be identified. Our results indicate that MspA has a high signal-to-noise ratio and the single nucleotide sensitivity desired for nanopore sequencing devices.

4.1 Introduction

In this chapter, we explore the ionic current levels and nucleotide resolution expected in sequencing of unmodified (native) DNA using a technique similar to that first used with α-hemolysin pores (α-HL). When single stranded DNA (ssDNA) was immobilized in α-HL by attachment to streptavidin, a molecule too large to pass through the pore, homopolymer strands of adenine, cytosine, and thymine each produced a distinct, orientation dependent, current blockade [47]. Furthermore, single nucleotide

differences in heteromeric DNA could be resolved [54, 48], and methylated cytosine could be distinguished from unmethylated cytosine [3]. However, because of α-HL's long β-barrel, the current differences were small and several recognition sites contributed to the ionic current [54, 48], making single nucleotide resolution challenging.

In this work, NeutrAvidin was used to hold single stranded DNA (ssDNA) in the constriction of MspA allowing long duration current readings (See Figure 4.1 and the description of NeutrAvidin in Chapter 2). Although the NeutrAvidin molecule is bulky, it does not appreciably affect the measured ionic current [54]. Using NeutrAvidin anchored ssDNAs, we established the nucleotide resolving power of M1-MspA by altering the nucleotides positioned in or near the constriction and recording the characteristic current. First, we determined residual current levels of each base type by using homopolymeric strands and established the current level for methylated cytosine (mC). Next, we characterized the recognition site within the pore using single nucleotide substitutions in an otherwise homopolymeric ssDNA. Finally, using segments of genomic sequence containing a single nucleotide polymorphism (SNP), we demonstrated that the ionic current through M1-MspA could be used to differentiate single nucleotide substitutions in heteromeric ssDNA.

4.2 Results

4.2.1 Nucleotide Current Signatures

When a voltage was applied across a lipid bilayer containing an M1-MspA porin, the ssDNA of a ssDNA-NeutrAvidin complex could thread through M1-MspA, reducing the current through the pore to a nucleotide-dependent level of about 25% of the unblocked value (see Figure A.1 of Appendix A for an example event). The ssDNA remained threaded through the pore until it was ejected back into to the *cis* volume by reversing the voltage. We calculated average residual current levels when the ssDNA was immobilized in M1-MspA under an applied voltage of 180 mV. The current

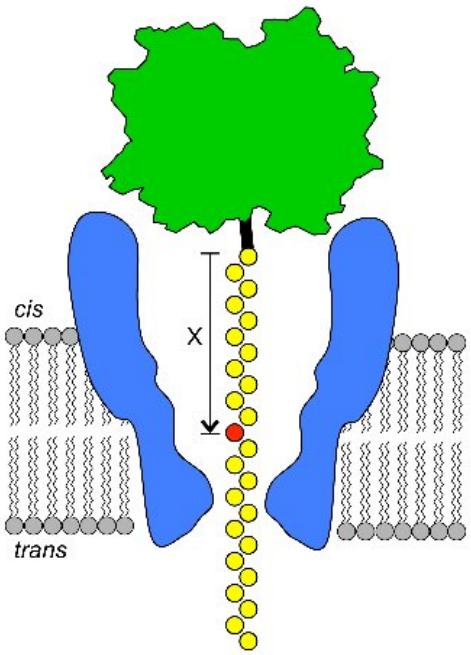


Figure 4.1: **Schematic Diagram of DNA Immobilization with NeutrAvidin.** Schematic diagram of MspA (blue) set up in a lipid bilayer (grey). Single stranded DNA (ssDNA, yellow) was attached to a bulky NeutrAvidin molecule (green) using a biotin linker (black). A specific nucleotide (red) is designated by it's position, X, from the biotin-NeutrAvidin 'anchor' where the nucleotide adjacent to the biotin is X=1. The ssDNA threads through the pore from the *cis* side of the bilayer until the bulky NeutrAvidin prevents it from further translocation. Residual ionic current was recorded as the ssDNA was immobilized within MspA.

distributions for each type of ssDNA were highly repeatable on several different M1-MspA pores (n>3) (see Table A.1 of Appendix A).

We recorded residual currents for ssDNA of homopolymer cytosine, $(dC)_{50}$ (denoted 'poly-dC'), adenine $(dA)_{50}$ ('poly-dA'), and thymine $(dT)_{50}$ ('poly-dT'). Although we could not study homopolymer guanine due to G-tetrad formation (Figure 2.9), we examined 3 guanines in a homopolymer adenine background, $(dA)_{13}(dG)_3(dA)_{34}$, (denoted 'poly$_3$-dG'). The NeutrAvidin molecule was bound to these homopolymers by a biotin linker attached to either the $5'$ or $3'$ end so that the $3'$ or $5'$ end, respec-

tively, threaded through the pore (see Figure 4.1). Histograms of the mean residual current and fitted Gaussians for each homopolymer in the 3′ and 5′ threading orientations are shown in Figure 4.2. The residual current levels are highly sensitive to the identity and orientation of nucleotides within M1-MspA. For both 3′ and 5′ threading, the residual currents for poly-dA, poly-dC, and poly-dT are well resolved with current separations of at least ∼8 pA (0.04 nS) and up to ∼50 pA (0.28 nS). For 3′ threading, poly-dA and poly$_3$-dG overlap by 27% of the area of the two Gaussian distributions. For 5′ threading, the residual currents for poly-dC, poly-dA, and poly$_3$-dG are well resolved with minimal overlap and the current of poly-dT is separated from the other current levels by at least 33 pA (0.18 nS). For poly-dC we find a substantial orientation dependence with a difference in the current of 5′ vs. 3′ threading of 45 pA (0.25 nS). For both orientations, the histograms of residual current for poly$_3$-dG are wider than those of the other homopolymeric strands. This breadth is possibly due to the surrounding adenine nucleotides influencing the ionic current. Neglecting poly$_3$-dG, there is a larger current distinction for 3′ threaded homopolymers. Therefore, we used 3′ threading for the remainder of the experiments described in this chapter.

To determine whether M1-MspA can resolve methylated cytosine, we compared residual current levels of 5′- /biotin/ $(dC)_{12}(mC)_4(dC)_{14}$ -3′, (abbreviated 'poly$_4$-mC'), and poly-dC. To ensure proper calibration between experiments, poly$_4$-mC and poly-dC were added sequentially to each pore. The peak current levels for poly-dC and poly$_4$-mC were separated by ∼1.1 pA (0.006 nS) with an overlap of ∼2% of the area of the two Gaussians (see Figure 4.3).

4.2.2 Region of Sensitivity

To realize nanopore DNA sequencing, the residual current of each nucleotide passing though the pore must be identifiable. This requires a short region over which the pore recognizes nucleotides. Armed with the residual current signatures of each base type, we mapped the region within the anchored ssDNA that is most influential in

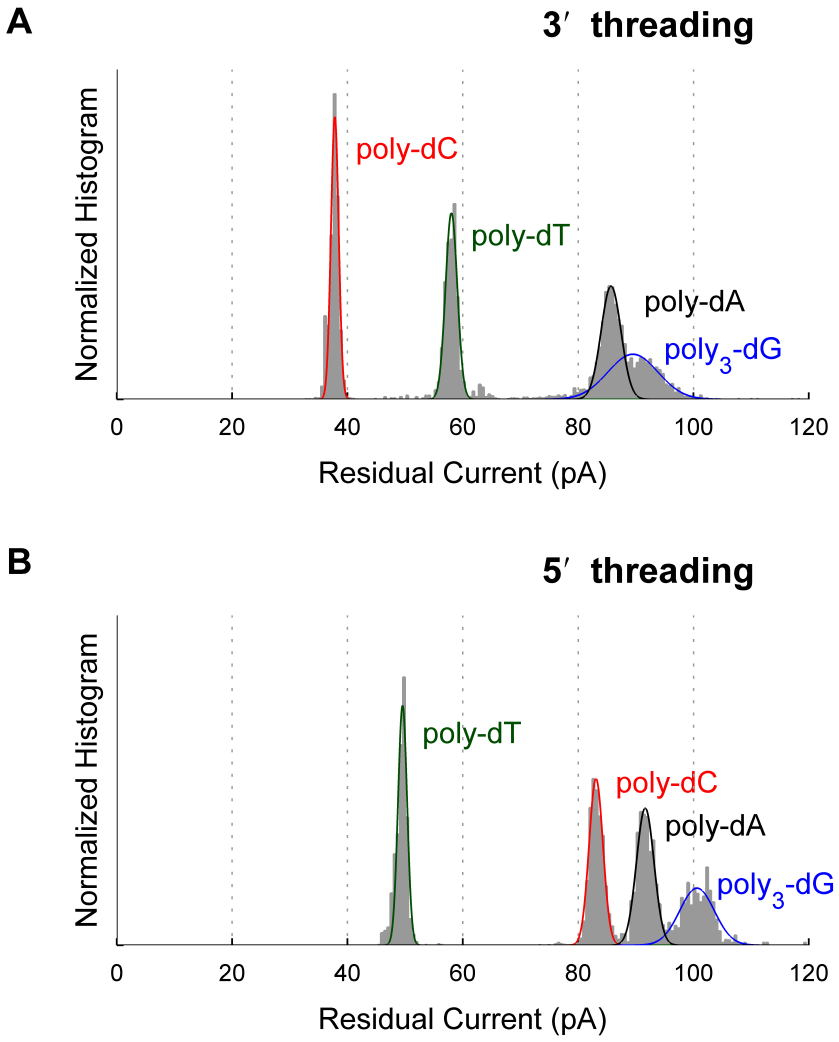


Figure 4.2: **Experimental Results for Homopolymer Strands Immobilized in MspA.** Mean residual ionic current (gray) and fitted Gaussian curves are shown for homopolymer adenine ('poly-dA', black), cytosine ('poly-dC', red), thymine ('poly-dT', green), and guanine ('poly$_3$-dG', blue) for (A) 3′ threading and (B) 5′ threading. With 3′ threading, poly-dA and poly$_3$-dG overlap by 27% of the area of the two Gaussians. With 5′ threading, poly-dA and poly$_3$-dG overlap by 2% and poly-dC and poly-dA overlap by 0.1% of the area of the two Gaussians.

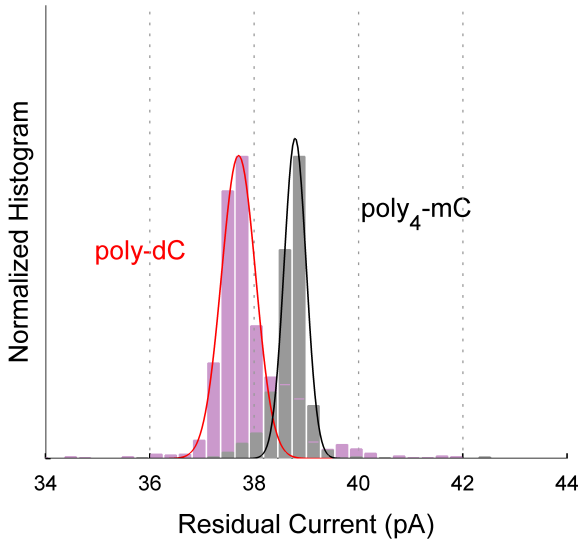


Figure 4.3: **Experimental Results for Methylated Cytosine Immobilized in MspA.** Mean residual ionic currents and the fitted Gaussian curve are shown for ssDNA of homopolymer cytosine ('poly-dC', red) and a strand with 4 methylated cytosines ('poly$_4$-mC', black) substituted at the 13th − 16th nucleotides from the NeutrAvidin anchor in an otherwise homopolymer cytosine strand. Experiments were conducted on three pores with poly$_4$-mC and poly-dC added sequentially to each pore. The peaks of the Gaussians are separated by 1.1 pA and the curves overlap by ∼2% of the area of the two Gaussians.

determining the ionic current through M1-MspA. Based on the crystal structure of MspA [18] (Chapter 2), approximately the first 20 nucleotides reside within its lumen or constriction when the NeutrAvidin molecule is positioned above the rim of MspA. It was expected that longer DNA strands would reach through the constriction of the pore to the *trans* side. We used ssDNA comprised of distinct regions of homopolymeric cytosine, adenine, and thymine to narrow in on the region of nucleotide sensitivity. By comparing the residual ionic current of these strands to the current signatures found above for the homopolymeric strands, the nucleotides most affecting the ionic current were determined to be around the 13th to 15th nucleotides from the NeutrAvidin (see

Figure A.2 of Appendix A).

4.2.3 Single Nucleotide Substitutions in a Homopolymeric Background

To determine which individual nucleotides most affect the ionic current, we introduced a single cytosine, dC, in a homopolymer adenine strand at the 11th through 18th nucleotide from the NeutrAvidin anchor (denoted dC_X where X is the position of substitution). Figure 4.4a shows a plot of the residual ionic currents for each position of dC substitution. The residual current deviates most from that of poly-dA for dC_{14} and dC_{15}, reducing the current to a level intermediate to that of poly-dC and poly-dA. For dC_{13} as well as for dC_{16}, the histograms of the current distributions have two separate peaks that are wider than those for dC_{14} and dC_{15}. The increased widths of the current distribution seen for dC_{13} and dC_{16} and the for above-described poly$_3$-dG, where three guanine nucleotides are surrounded by adenines, are consistent with the nucleotides of interest shifting in and out of the constriction, possibly due to thermal fluctuations, transient bindings and variations in the seating of the NeutrAvidin anchor on the rim of M1-MspA (see Chapter 6 for further discussion). With these data, we conclude that there is a single region in M1-MspA that is sensitive to nucleotides. Assuming a Gaussian distribution, the sensitive region corresponds to X = 14.5 with a half width at half height of 1.5 nucleotides. This indicates that the nucleotides centered in MspA's constriction as well as the neighboring nucleotides just outside the constriction affect the ionic current.

To confirm the location of the sensitive region, we similarly examined a single adenine substitution, dA_X in poly-dC (see Figure 4.4b). The dA_X substitutions in poly-dC influence ionic current much less than the dC_X substitutions in poly-dA found above. To determine whether this effect is the influence of the homopolymer background on the ionic current, we also studied four different dA_X substitutions in an otherwise poly-dT strand (see Figure 4.4b). We see that the current levels for dA_X substitutions in poly-dT are distinct from poly-dT at X = 13, 14, 15, and 16. A

fitted Gaussian gives a peak value at X = 14.1 and half width at half height of 1.5. In total, these results suggest that the ionic current is most influenced by the nucleotides at positions X = 14−15 and that the homopolymer background affects the ability of M1-MspA to recognize single nucleotides. See Chapter 6 for further discussion on the sequence specific characteristics of DNA.

4.2.4 Single Nucleotide Substitutions in a Heteropolymeric Background

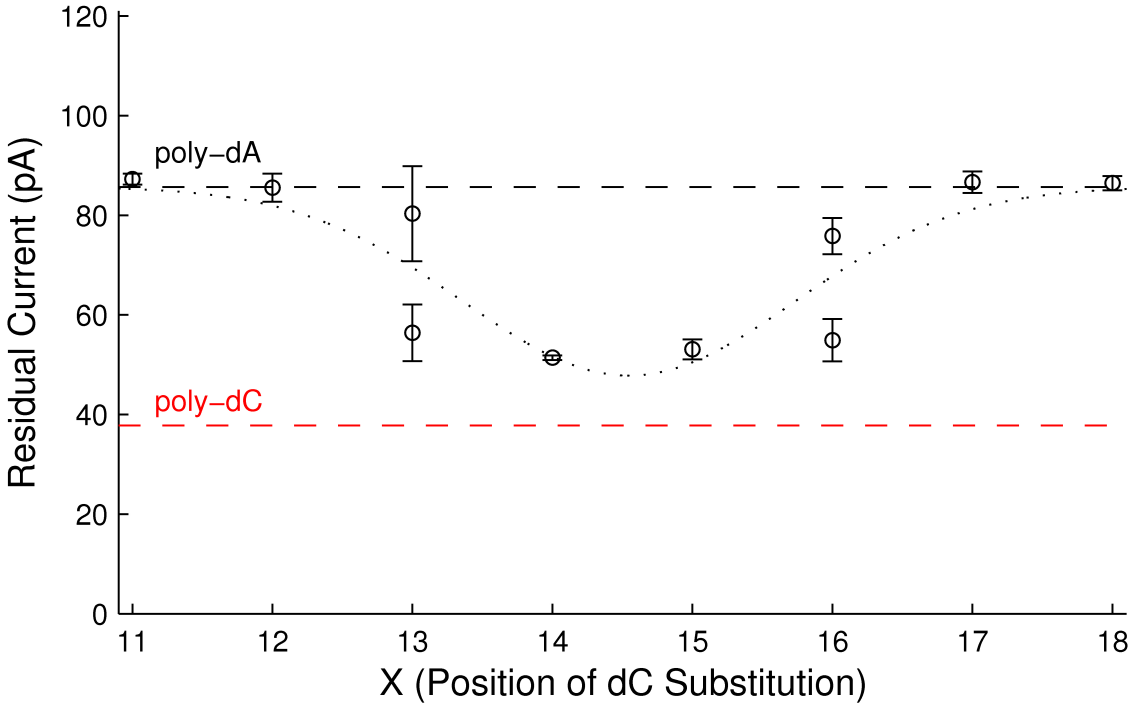
To investigate the prospect of M1-MspA to identify a single nucleotide in a heteromeric background, we prepared genomic sequence DNA with single nucleotide substitutions. Specifically, we analyzed DNA containing the single nucleotide polymorphism (SNP) rs889312, which has been linked to increased risk of breast cancer [17]. The ssDNA was bound to NeutrAvidin such that the SNP (either dC or dA) was at X = 13, 14, 15 or 16 (denoted dC_X or dA_X where X is the position of the SNP). Histograms for the residual ionic currents for each SNP substitution are shown in Figure 4.5. The dA and dC variations are clearly distinguishable for X = 14 and 15 and become complex and poorly resolved for X = 13 and 16. This procedure was repeated for a second SNP, rs1447295, which is associated with an increased risk of prostate cancer [1] [19]. For rs1447295 we also observe two highly resolved current levels distinguishing the two SNP variants (see Figure A.3 of Appendix A). For both SNP experiments, M1-MspA was able to clearly distinguish single nucleotide differences in genomic DNA in its sensitive region.

4.3 Discussion

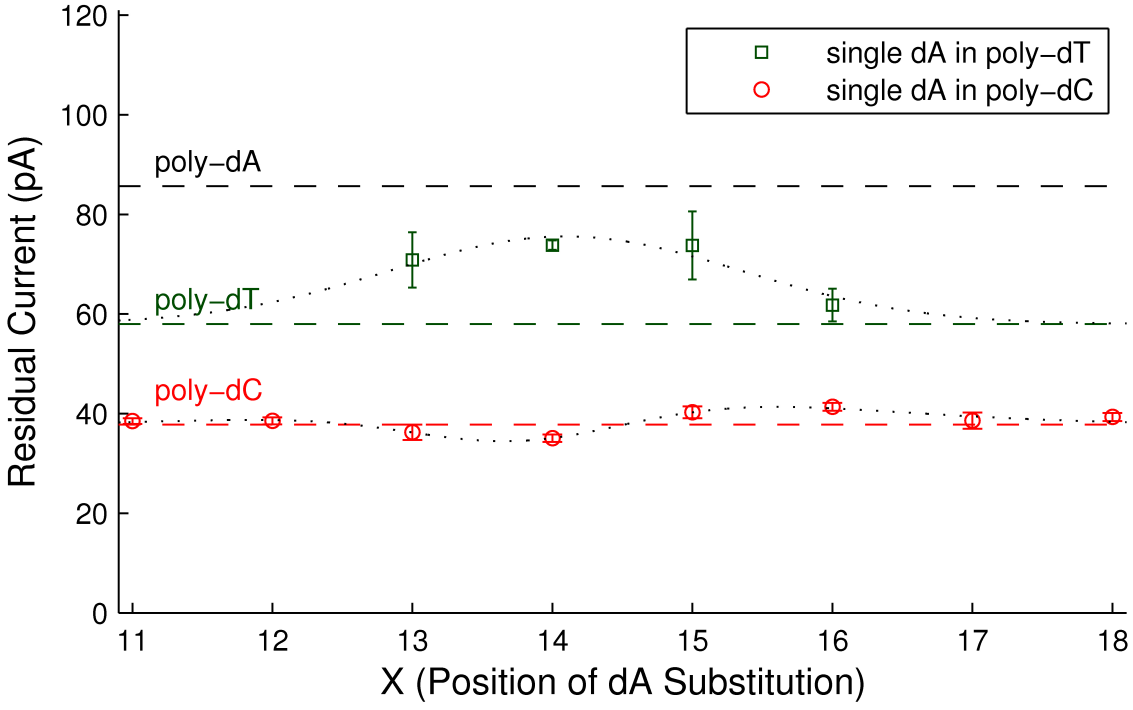
The differences in nucleotide specific current found with M1-MspA are about an order of magnitude larger than those found in similar experiments with α-HL [47, 54, 48]. Additionally, M1-MspA has a single recognition site where nucleotides affect the ionic current whereas α-HL has several such regions [54, 48]. The short M1-MspA constriction results in a small volume of high current density which is likely responsible for

Figure 4.4: **Experimental Results for Single Heteromeric Substitutions in Homopolymer ssDNA Immobilized in MspA.** Peak values and half width half height values, represented as error bars, of fitted Gaussians of mean residual ionic currents are shown for experiments with a single heteromeric substitution in homopolymer ssDNA. For comparison, the mean residual ionic currents for relevant homopolymer strands are shown as horizontal dashed lines. Gaussian curves (dotted) are included to aid the eye. (A) DNA strands containing a single cytosine (dC) substituted at nucleotide position X from the NeutrAvidin anchor in an otherwise poly-dA strand. Large error bars (X = 13) indicate a wide distribution and the two points at X = 13 and X = 16 indicate two current levels for that substitution. The residual current differs most from that of poly-dA (black dashed line), most resembling that of poly-dC (red dashed line) at X = 14 and 15. (B) DNA strands containing a single cytosine (dA) substituted at nucleotide position X from the NeutrAvidin anchor in an otherwise poly-dC (red) or poly-dT (green) strand. A single dA substitution in homopolymer poly-dC yields a residual current similar to that of poly-dC (red dashed). For a single dA substitution in homopolymer poly-dT, the current deviates most from that of poly-dT (green dashed) towards the level for poly-dA (black-dashed) at X = 14 and 15.

A


120
100
80
60
40
20
0
Residual Current (pA)
poly–dA
poly–dC
11
12
13
14
15
16
17
18
X (Position of dC Substitution)

B


single dA in poly–dT
single dA in poly–dC
120
100
80
60
40
20
0
Residual Current (pA)
poly–dA
poly–dT
poly–dC
11
12
13
14
15
16
17
18
X (Position of dA Substitution)

the large open state conductance and the large specificity to various nucleotides in the constriction. The ionic current for poly-dC is significantly higher for the 5′ threading orientation in M1-MspA. This is opposite to α-HL [47], suggesting an orientation dependent interaction with the porin. A simple resistor model wherein the total resistance is determined by the algebraic sum of the resistances posed by each nucleotide and the pore cross section at the position of that nucleotide [48] does not properly describe our data. Our results also do not support a rate-limiting model where the highest impedance in ionic current controls the measured current level [15]. It seems that a more complex model accounting for detailed positioning, hydration, charge distributions and binding affinity of nucleotides to the pore is needed to properly describe this system. Additionally, our data show that nanopore sequencing data analysis of DNA must take into account the effect of neighboring nucleotides not centered in the constriction.

Compared to our own data with DNA held by a duplex section in MspA (Chapter 3) we find larger currents and current differences as well as a slight reordering of the nucleotide specific current levels. An increase in residual current when DNA is anchored with a large molecule is expected [54] and may be due to elongation and straightening of the DNA under the force provided by the electric field of the constriction (see Chapter 6). Additionally, in DNA held by a duplex region the length and sequence of a duplex region near the constriction throttles the current and affects the neighboring single stranded bases occupying the constriction [15].

We have shown that homopolymer ssDNA is well resolved in M1-MspA and methylated cytosine is distinguishable from unmethylated cytosine. M1-MspA is capable of high-resolution single nucleotide discrimination, a prerequisite for DNA nanopore sequencing. In Chapter 5, we employ a technique where a DNA polymerase is used to ratchet ssDNA, one base at a time, through the MspA. The procession time of DNA polymerases such as phi29 is sufficiently slow, >20 ms/nt, that MspA would be able to resolve bases in order as they pass through the pore.

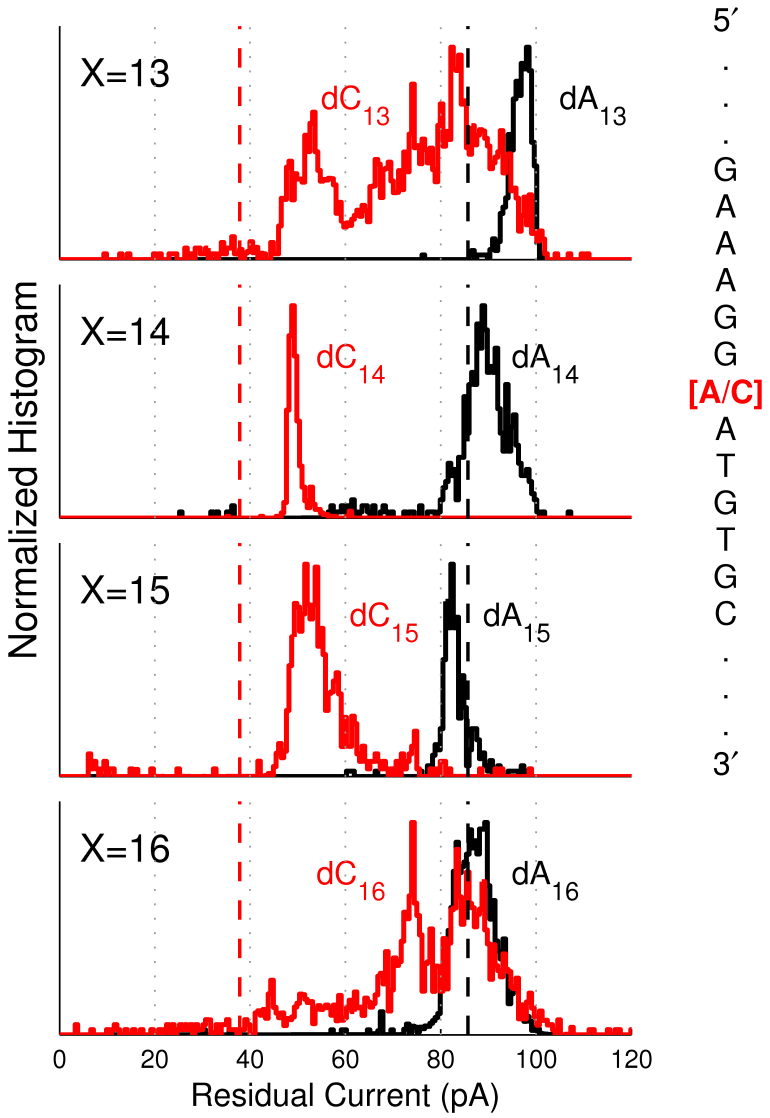


Figure 4.5: **Experimental Results for Immobilized Strands Containing a Single Nucleotide Polymorphism (rs889312).** A segment of DNA containing SNP rs889312 was bound to NeutrAvidin such that the polymorphism is at the 13−16th position from the NeutrAvidin anchor (denoted dC_X or dA_X where X was the position of the SNP). Part of the surrounding sequence is shown with the SNP highlighted in red (see Table A.1 of Appendix A for complete sequence). Histograms of mean residual current levels are shown for each variant and SNP location, X = 13−16. The two variations are most clearly resolved for X = 14 and 15. For reference, the mean residual ionic currents for poly-dC (red dashed) and poly-dA (black dashed) are shown.

Chapter 5

ENZYME CONTROLLED TRANSLOCATION OF DNA THROUGH MSPA USING PHI29 DNA POLYMERASE

This work was published in the April 2012 issue of Nature Biotechnology. [36]. It was featured on the cover(Figure 5.1).

In previous chapters, we found that the protein pore *Mycobacterium smegmatis* porin A (MspA) has high sensitivity to nucleotides held in its constriction. Here we demonstrate the ability to resolve current levels corresponding to a known DNA sequence passing through MspA by using the phi29 DNA polymerase to control the speed of DNA translocation. As phi29 DNAP synthesizes DNA, it functions like a motor pulling a single-stranded DNA template through MspA. We observe well-resolved and reproducible ionic current levels that map to the known DNA sequences. With single-nucleotide resolution and DNA translocation control, this system lays the groundwork for a nanopore DNA sequencing system.

5.1 Introduction

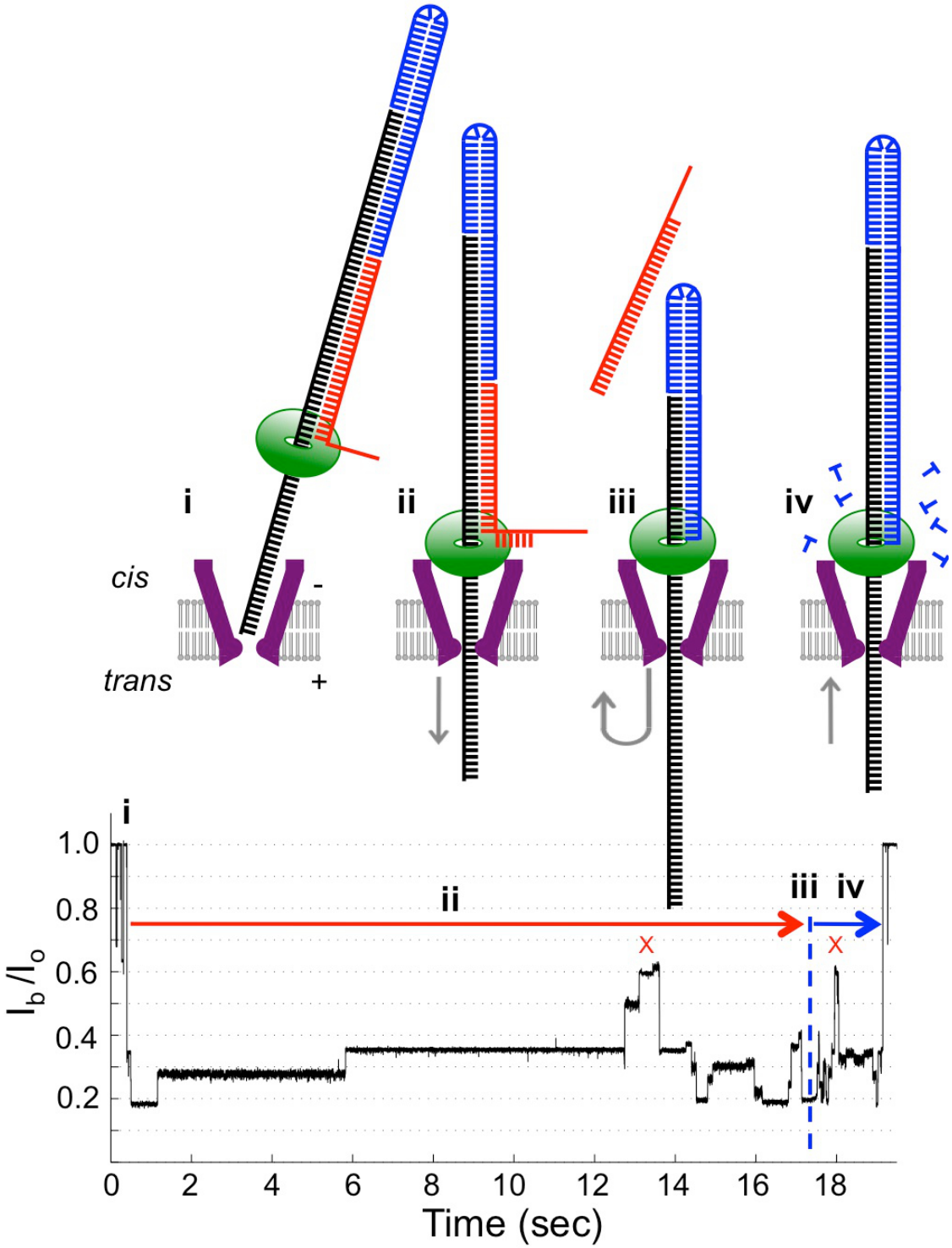
In this work, we read DNA by detecting current levels associated with single-nucleotide movement of the strand through MspA. To our knowledge no system has yet been reported that can read nucleotide-specific current levels as an unmodified strand of DNA passes through a nanopore.

Freely translocating DNA acting under the force of an electric field moves through nanopores at an average rate greater than one nucleotide/μs, $\sim$1,000 times too fast to distinguish nucleotide-specific current changes from noise [31, 8, 10, 39] (Chapter 2). Simple techniques for reducing the velocity of translocating DNA, such as lowering

Figure 5.1: **Cover of Nature Biotechnology, April 2012.** This work was featured on the cover of Nature Biotechnology. An artist's representation of enzyme controlled DNA translocation through MspA. A strand of DNA is released from a 'blocking oligomer' (red) and is poised for ratcheting by phi29 DNA polymerase (coral) through an MspA protein nanopore (blue). Image Credit: ©Kenneth Eward

the experimental temperature or increasing the viscosity of the solution, also reduce the current signal and therefore do not improve the signal-to-noise ratio.

Figure 5.2: **Event Structure for Enzyme Controlled DNA Translocation Experiments.** (Top) Schematic depicting a standard experiment. Roman numerals correspond to positions in the current trace (bottom). (Bottom) The measured blockage current (I_b) as a fraction of the open pore current (I_o) is shown for a sample event. (i) A single MspA pore (purple) in a lipid bilayer (gray). The template strand (black) contains the sequence to be read. A primer strand (blue) is hybridized to the template's $3'$ end. A blocking oligomer (red) with a $3'$ end of several abasic sites is adjacent to the primer. The phi29 DNAP (green) binds to the DNA to form a complex that is driven into MspA. A positive voltage is applied to the *trans* side. The single stranded $5'$ end of the DNA-motor complex threads through MspA and the ionic current drops. (ii) The electric force on the captured strand draws the DNA through the phi29 DNAP, unzipping the blocking oligomer. Arrows show the direction of motion of the DNA template strand. The ionic current exhibits distinct steps while nucleotides pass through the pore. (iii) The blocking oligomer is removed and DNA reverses direction (marked by blue dashed line in the current trace). (iv) The phi29 DNAP incorporates nucleotides into the primer strand, pulling the template toward the *cis* side. The current repeats previously observed levels in reverse time-order. Two abasic sites produce a high current peak ($\sim$0.6 I_o) indicated by red Xs. This marker is first seen during unzipping and then again during synthesis. When synthesis is complete, the DNA and DNAP escape to the *cis* volume, marked by the return to I_o.


i
ii
iii
iv
cis
trans
-
+
i
ii
iii iv
X
X
1.0
0.8
0.6
0.4
0.2
I_b/I_o
0
2
4
6
8
10
12
14
16
18
Time (sec)

Control of DNA translocation through a nanopore has been attempted using interspersed regions of double-stranded DNA [15] (Chapter 3) and DNA polymerases (DNAP) [11, 35, 22, 59, 28, 3]. The DNA polymerases of *Escherichia coli* (Klenow fragment) [22, 59, 28, 3], and of the bacteriophages, T7 (ref. [22]) and phi29 (refs. [11, 35]) have been studied as molecular motors to control the movement of DNA in single-nucleotide steps through α-hemolysin. The monomeric phi29 DNAP ($\sim$66.5 kD) has $5'-3'$ primer strand extension and $3'-5'$ exonuclease functions in the presence of Mg^{2+} as well as $5'-3'$ strand displacement activity [7, 49]. Both the extension and exonuclease functions move DNA in single-nucleotide steps, which can be utilized for controlled DNA translocation through a nanopore. In contrast to Klenow fragment and T7 DNAP, phi29 DNAP is able to synthesize very long stretches of DNA (>70 kb) along a single template strand, and it can synthesize under loads of up to $\sim$37 pN [29, 5], enough to counteract the electrophoretic force needed to drive DNA through the pore (see Chapters 2 and 6).

In a recently developed technique [11], phi29 DNAP is bound to a DNA template and a 'blocking oligomer' prevents extension and excision of the primer strand annealed at the template's $3'$ end. The single-stranded end of this phi29 DNAP-DNA complex is drawn into the pore until the DNAP rests on the pore's rim. The force on the template strand, estimated to be $\sim$10 pN, causes the blocking oligomer to 'unzip' $3'$ to $5'$, exposing the extendable $3'$ end of the primer and allowing synthesis to start. Previous experiments using this strategy with α-hemolysin showed that the DNA translocation speed was slowed to $2.5-40$ nt/sec (ref. [11]), but α-hemolysin lacks the ability to resolve nucleotide-specific current.

5.2 Results

Here we combine our previously engineered M2-MspA nanopore, herein designated MspA, with the phi29 DNAP blocking oligomer technique to read well-resolved and distinguishable current levels as DNA is drawn through the pore. An illustration of

the experiment and a typical current trace are shown in Figure 5.2. The current through an open MspA pore in 0.3 M KCl was $I_o = 110 \pm 6$ pA (mean $\pm$ s.d., N = 25). In all experiments, 80−91 nucleotide (nt) DNA template strands containing the section to be read were annealed to a 23-nt primer complementary to the template's $3'$ end. The primer had a hairpin on its $5'$ end to prevent the phi29 DNAP from acting on the double-stranded end of the DNA. Adjacent to the primer we annealed the blocking oligomer, a 15 or 29 nt complementary DNA oligomer with a $3'$ end consisting of seven abasic residues and a three-carbon spacer. Most DNA template strands also contained two adjacent abasic sites, which generated a high ionic current level of $\sim$0.6−0.7 I_o as they passed through the pore (see Table B.1 of Appendix B for sequences).

To study an easily resolvable DNA sequence, we used a 'block homopolymer' DNA template containing all four bases in short homopolymer sections of adenine (dA_3), guanine (dG_3) and cytosine (dC_5), each separated by thymine (dT_3). In control experiments without phi29 DNAP, the DNA translocation was too fast (>1 nt/μs) to read individual nucleotides (Figure B.1a of Appendix B). Next, we added phi29 DNAP to the *cis* volume, but omitted the divalent cations and deoxyribonucleotide triphosphates (dNTPs) necessary for synthesis. We recorded translocation events with current patterns consistent with force-activated unzipping of the blocking oligomer. These events exhibited distinct current levels including a high peak ($\sim$0.6 I_o) when the two abasic sites passed through the constriction (Figure B.1b,c of Appendix B). As expected, after removal of the blocking oligomer, the phi29 DNAP was unable to extend the primer strand owing to the absence of divalent cations and dNTPs. The phi29 DNAP eventually either fell off, allowing cooperative dissociation of the primer in MspA, or continued onwards, unzipping the primer strand (Figure B.1b of Appendix B). The duration of these events was generally >20 s.

To allow DNA synthesis to proceed, we added all four standard dNTPs (100 μM each) and 10 mM $MgCl_2$. In five separate experiments using the block homopolymer

DNA, we found 33 events with the same distinct current pattern. A typical current trace (Figure 5.2) is characterized by a rapid succession of current steps (levels) with various durations and current values ranging between 0.18 and 0.4 of I_o and two peaks with current levels above 0.6 I_o. Closer inspection reveals that many levels at the beginning of an event were repeated later in the event, but in the opposite order. These results are consistent with part of each DNA molecule being read twice; once, during stepwise motion of the DNA template toward the *cis* side, while unzipping the blocking oligomer, and again, during stepwise motion back toward the *trans* side, while phi29 DNAP was synthesizing. The levels observed during unzipping were symmetrical in time with levels observed during synthesis (Figures B.2 and B.3 of Appendix B). Once phi29 DNAP completed synthesis of the DNA template strand, both the DNA and phi29 DNAP exited to the *cis* side. The average duration of events containing phi29 DNAP synthesis ($\sim$14 s) was shorter than events recorded in the absence of Mg^{2+} and dNTPs because synthesis is faster than the time for phi29 DNAP to release from the DNA template or to unzip the primer strand. In our experiments with $MgCl_2$ and dNTPs, $> 10\%$ of the total data acquisition time was spent with a DNA-motor complex threaded through MspA.

The observed duration of each current step was generally shorter during DNA synthesis than during unzipping of the blocking oligomer. Examining levels associated with unzipping and synthesis revealed that both processes appear to be stochastic. Generally, level durations were exponentially distributed (Figure B.4 of Appendix B) with time constants of $\tau_{unzip} = 100 \pm 12$ ms and $\tau_{syn} = 40 \pm 10$ ms. In many events, a few much longer levels ($\sim$1 s) were observed during unzipping or synthesis. In many events, one or more of the expected steps were missing because the duration was too short to be observed or adjacent current levels were indistinguishable.

Occasionally during synthesis, the current returned briefly to its previous current level before continuing on. These easily identifiable 'toggle' patterns are consistent with phi29 DNAP back-stepping, owing to its $3'$ to $5'$ proofreading activity, or failing

to incorporate a nucleotide (Figure B.5 of Appendix B). In many cases, toggles occur numerous times in succession. In rare cases, we observed back-stepping of two sequential nucleotides.

We investigated the effect of dNTP concentration on synthesis by reducing the concentration of dCTP. In these 'starvation' experiments we found very long pauses in synthesis as well as toggling between distinct current levels. Furthermore, these experiments allowed us to determine that the distance between the phi29 DNAP's nucleotide incorporation site and MspA's constriction is 14−15 nucleotides (Figure B.5 of Appendix B).

We analyzed the current traces from the MspA-DNA-motor complex in its synthesis phase to filter out noise and extract current levels (Figure 5.3). We observed that the temporal ordering of the current levels was preserved across all traces and that the levels could be aligned with the known DNA sequence. The ionic current was lowest ($\sim$0.20 I_o) when dT_3 was centered in MspA's constriction, highest ($\sim$0.40 I_o) for dA_3 and moderate ($\sim$0.37 I_o and $\sim$0.30 I_o) for dC_4 and dG_3, respectively. These results are qualitatively consistent with our previous work [37] where homopolymer DNA was held statically in MspA by a NeutrAvidin molecule (Chapter 4). Those experiments at 1.0 M [KCl] resulted in $\sim$0.15 I_o for poly-dT, $\sim$0.25 I_o for poly-dC and $\sim$0.28 I_o for poly-dA; blockage current for poly-dG could not be reliably determined.

We qualitatively aligned the sequence of the DNA to the levels (Figure 5.3c) by assuming equal spacing of nucleotides and matching the position of the abasic peak, the turnaround point and the low currents typically found with multiple thymines. The nucleotides positioned near the center of the constriction dominate the total current of a level, whereas the nucleotides to either side control the current to a lesser extent. This is consistent with previous MspA experiments where each current level was affected by approximately four nucleotides in and around the constriction [37] (Chapters 3 and 4). For example, when DNA is positioned so that the four nucleotides centered in the constriction are TTAA (Level 3 of Figure 5.3c), the current

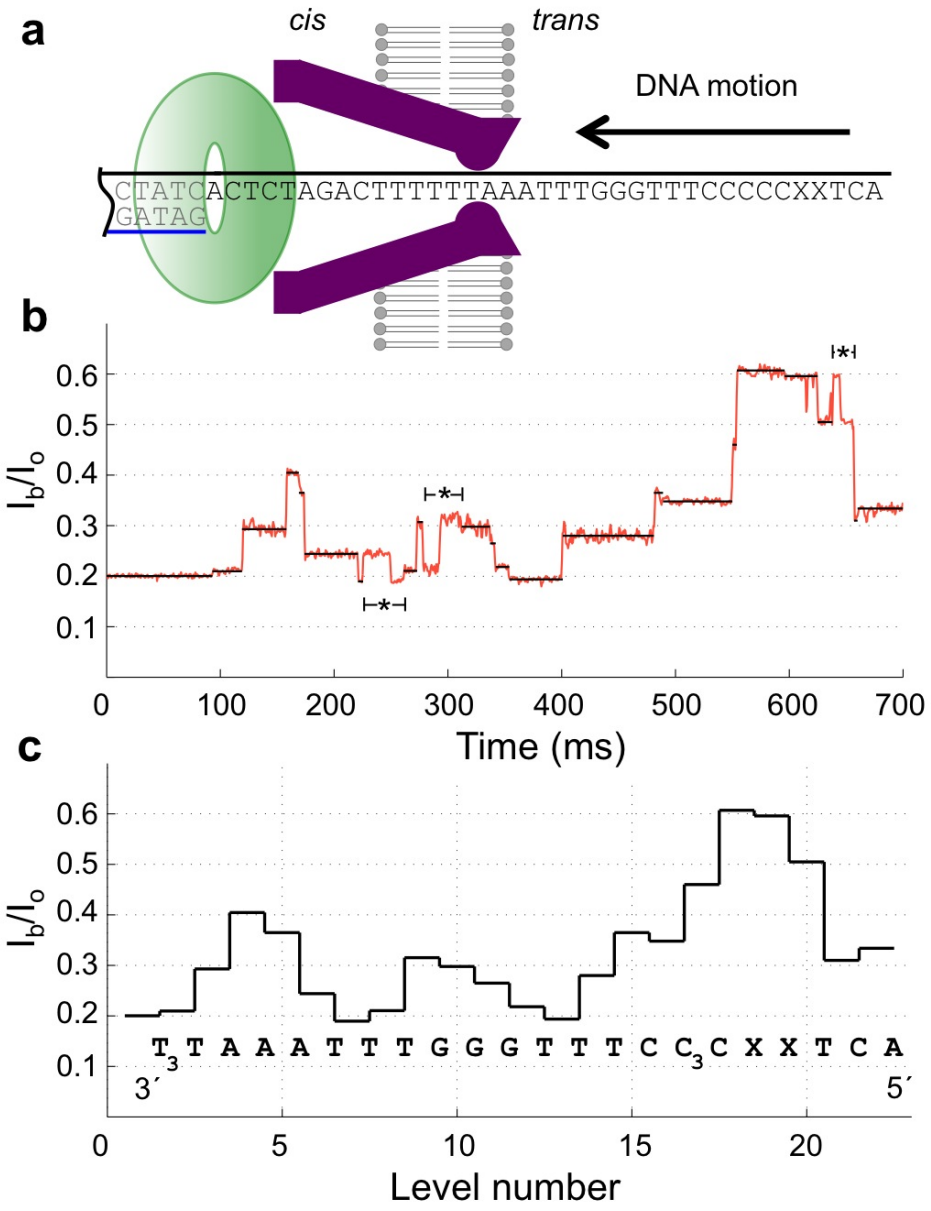


Figure 5.3: **Example Current Trace for Enzyme Controlled Translocation of Block Homopolymer DNA During Polymerase Synthesis.** (a) Illustration of phi29 DNAP during synthesis. (b) Typical current trace for synthesis of 'block homopolymer' DNA. At the beginning of the read, multiple thymines (dT_4) produce a current level of $I_b/I_o \sim 0.20$. At the end of the read, two abasic residues (XX) produced a high current ($I_b/I_o \sim 0.61$).Toggles are indicated by *. (c) Mean currents of levels extracted from b are plotted with the associated DNA sequence.

is intermediate to the current levels for TTTT (level 1) and AAAT (level 5). However, small shifts in the position of nucleotides within the pore constriction may play an important role in the observed current (see Chapter 6).

To demonstrate that our system could achieve single-nucleotide resolution, we used an 82-nt DNA template containing repeated instances of the trinucleotide sequence CAT except for one triplet in the center of the sequence that had a single dT substituted with a dG. The majority of events recorded (n = 475 events, N = 5 pores) with this DNA template consisted of a repeated stair-step current motif, except at one location where it changed as a result of the dG substitution (Figure 5.4). Based on the results with the block homopolymer DNA template (Figure 5.3), the 'CAT' stair-step motif during synthesis is consistent with cyclic rotations of $TACT \rightarrow ACTA \rightarrow CTAC$ centered about the constriction, where the principal contribution is from the middle two nucleotides (note the DNA is written from $3'$ to $5'$ from left to right in the traces). The single-nucleotide substitution causes the pattern to be perturbed for about four steps before returning to the CAT motif (Figure 5.4). Surprisingly, when ACGA is centered in the constriction, the current in the pore has a relatively high value, even though the adenines are only peripherally involved. This suggests that nucleotide type is not the only determinant of blockage current. Other factors such as secondary structure of the DNA, nucleotide-nucleotide interactions, nucleotide interactions with the pore and exact positioning within the constriction must be considered (see Chapter 6). However, a single-nucleotide substitution passing through MspA's constriction alters the current level pattern significantly, indicating that a nucleotide's identity and position in the strand is encoded in the observed current. This is consistent with our previously reported finding [37] using a single-nucleotide polymorphism in NeutrAvidin-anchored DNA (see Chapter 4).

Next, we demonstrated the feasibility of reading heterogeneous DNA with MspA and phi29 DNAP. We tested four DNA heteromeric base sequences (Table B.1 of

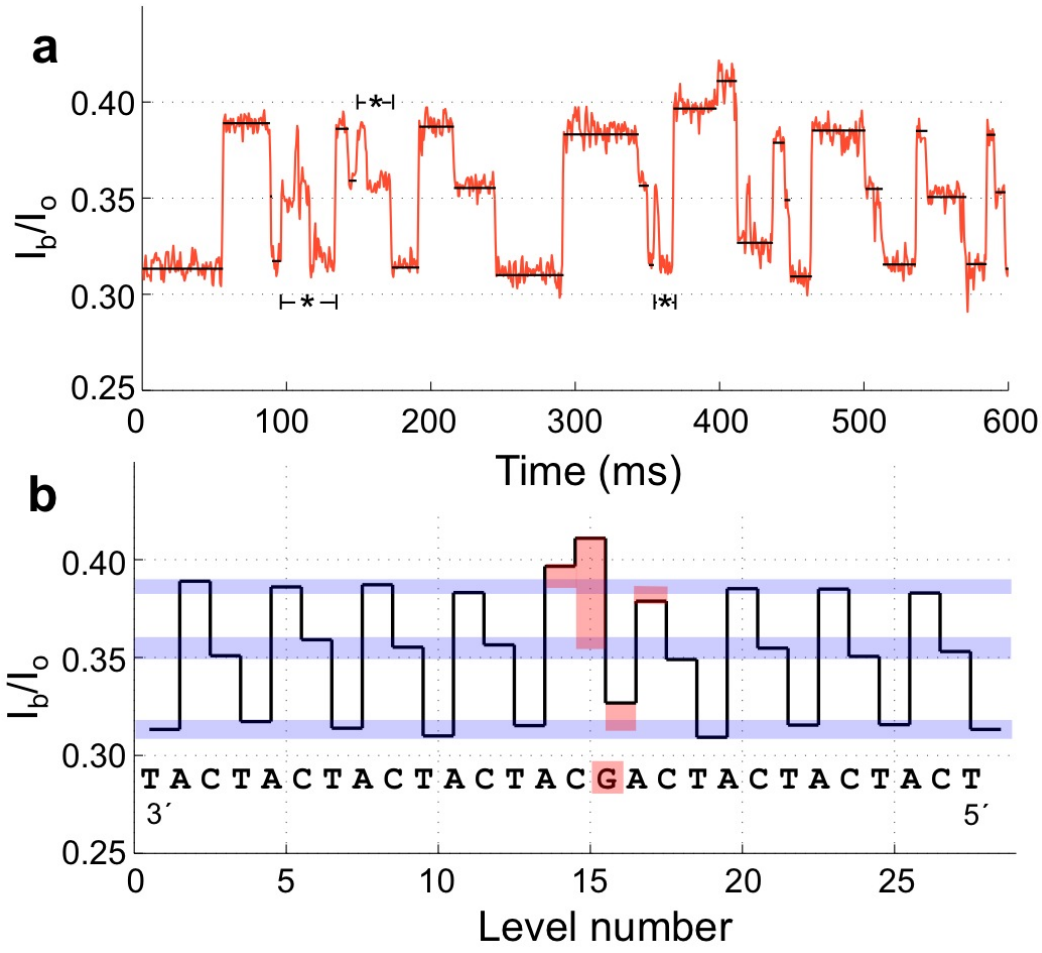
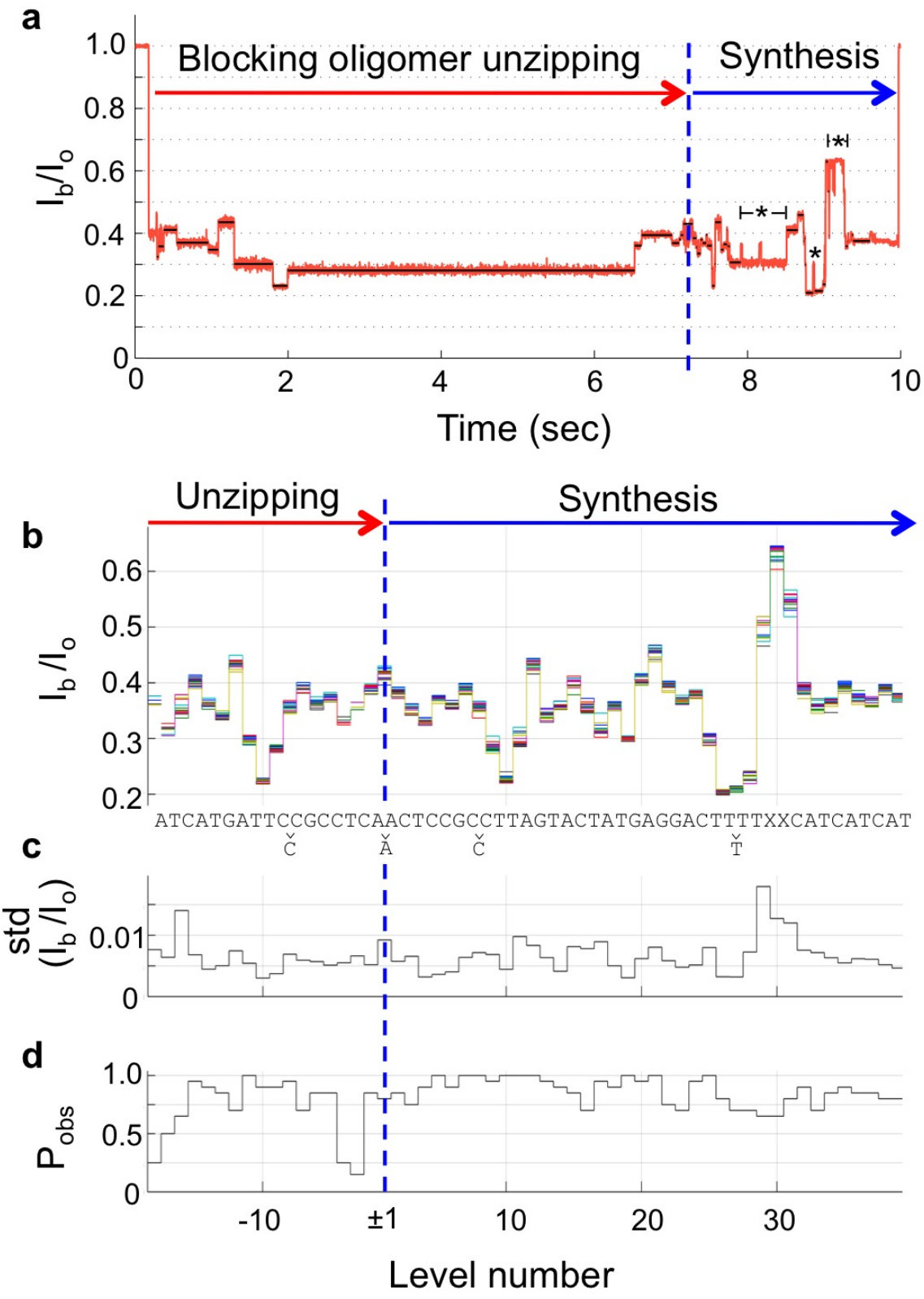


Figure 5.4: **Example Current Trace for Enzyme Controlled Translocation of Repeating DNA Template Sequence During Polymerase Synthesis.** (a) Example trace for a DNA template composed of repeated 'CAT' trinucleotides, with the exception of one 'CAG' triplet in the middle of the sequence. (b) Mean currents of levels extracted from a with the associated DNA sequence are shown. The current trace exhibits a repeating pattern of three levels (blue bars) interrupted by the single dG substitution (highlighted in orange). Four levels are affected by the single dG with the largest deviations closest to the substitution. This indicates that the residual current is principally influenced by one or two nucleotides with a lesser influence from neighboring nucleotides. Note the DNA is written from 3′ to 5′ from left to right in the trace.

Appendix B). Analogous to the block homopolymer DNA, these strands also had two abasic residues near the end of the template sequence to produce a clear marker signifying successful completion of synthesis. For each experimental sequence, we obtained consistent current patterns with distinct features corresponding to individual nucleotide steps as DNA passed through MspA's constriction. Figure 5.5a shows a sample current trace from one of several experiments (n = 47, N = 4) recorded with a 91-nt DNA template strand called 'heteromer DNA 1'. We extracted current levels from 20 events (obtained from two different pores) that exhibited unzipping and synthesis. We formed a consensus current level sequence from multiple events. Next we aligned the events to the consensus using a Needleman-Wunsch algorithm and overlayed the aligned levels (Figure 5.5b). Consistent with previous experiments [11], the current trace showed a symmetry about a level (dotted blue line in Figure 5.5b) that corresponded to the nucleotides in MspA's constriction when the blocking oligomer is completely removed. This demonstrates that current levels observed during unzipping are repeated in opposite time-order during synthesis. As the blocking oligomer used in this experiment was shorter (15 nt) than the length of the template to be read, the number of levels observed during unzipping was fewer than those observed during synthesis, and the abasic peak marking the end of the read was not observed during unzipping.

At the beginning of synthesis, adenines within MspA's constriction caused a current level near $\sim$0.4 I_o, consistent with the current due to dA_3 in experiments with the block homopolymer DNA. Near the end of the sequence, just before the abasic marker, a thymine repeat section resulted in a distinct poly-dT level, $\sim$0.2 I_o. The sequence of the heteromer DNA 1 could be aligned to the trace to match expected features (Figure 5.5b). For example, the local minimum at level 10 and the peak at level 30 are matched to CTTA and TXXC, respectively. We expect the nucleotides listed to the right and left of each level to contribute most to the current value. The total number of observed current steps was less than the length of the sequence be-

Figure 5.5: **Experimental Data for Enzyme Controlled Translocation of 'Heteromer DNA 1'.** (a) Example event for 'heteromer DNA 1'. Unzipping and synthesis phases are indicated. (b) Demonstration of the repeatability and similarity of current levels found in multiple events. Time-ordered levels for n = 20 events collected on N = 2 pores are found using a level detection algorithm. A consensus current level sequence was created from multiple events. The time-ordered levels consistent with the consensus sequence are overlaid (one color for each event). The symmetry of the plot about level ~ 1 shows that the same levels occur during both unzipping and synthesis (in reverse order). The correlated sequence is shown below. Levels not detected, owing to repeat nucleotides and/or similar currents, are indicated with the unobserved nucleotide beneath the sequence. (c) Standard deviations (std) derived from the scatter of each overlaid level. The average 1-sigma fluctuation in levels is below 1% of I_o. (d) The probability, P_{obs}, of finding a given level in the proper order. Some levels, such as level 10 or level 21 are found in all events (n = 20). Levels immediately before the transition from unzipping to synthesis are often too fast to be detected, resulting in low P_{obs}.


a
Blocking oligomer unzipping
Synthesis
I_b/I_o
1.0
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10
Time (sec)
b
Unzipping
Synthesis
I_b/I_o
0.6
0.5
0.4
0.3
0.2
ATCATGATTCCGCCTCAACTCCGCCTTAGTACTATGAGGACTTTTXXCATCATCAT
C
A
C
T
c
std (I_b/I_o)
0.01
0
d
P_{obs}
1.0
0.5
0
-10
±1
10
20
30
Level number

tween the turnaround and the abasic peak, indicating that two current steps were unobserved. The suspected locations of these missing levels are indicated (Figure 5.5b). These missing levels are thought to be indistinguishable from neighboring current levels. For example, level 27 is associated with the dT_5-homopolymer region and likely consists of two steps with indistinguishable currents. The scatter of current values for each level reflects its reproducibility from event to event and from pore to pore. We observed that the average scatter, quantified by the standard deviation of current levels, was below 1% of I_o (Figure 5.5c). This indicates that the positioning of the DNA in the constriction is highly reproducible. The average probability that our level-finding algorithm detected a given level in the proper order between the turn-around point and the abasic peak was 90% (Figure 5.5d). The unzipping process was more likely to skip levels, particularly those near the beginning or end of unzipping. The scatter and probability of observing a level are dependent on the level detection and alignment algorithms (see Appendix B.1). Three additional heteromer DNA sequences, 'heteromer DNA 2' (n = 121, N = 3), 'heteromer DNA 3' (n = 94, N = 3) and 'heteromer DNA 4' (n = 142, N = 5) produced similar results (Table B.1 and Figures B.6 to B.16 of Appendix B). Differences in nucleotide-specific current levels differed by as much as ∼0.36% I_o or 40 pA.

5.3 Discussion

In summary, we demonstrated that individual single-stranded DNA molecules traversing through the short and narrow constriction of MspA under the control of phi29 DNAP yielded distinct current levels that were related to the sequence of the nucleotides in MspA's constriction. We observed current patterns associated with two distinct processes: the 5′ leading motion of the DNA template, which is consistent with the nearly monotonic unzipping of the blocking oligomer, and the 3′ leading motion, which is consistent with synthesis by phi29 DNAP. Motion during synthesis is faster than unzipping, and the levels associated with most single-nucleotide advances can

be identified. The DNA nucleotides that passed twice through MspA's constriction, once during unzipping of the blocking oligomer and once during synthesis, yielded redundant current readings.

The patterns of current levels can be related to a known DNA sequence. Automatic sequence extraction algorithms must be developed to sequence unknown DNA. Because a model predicting the current pattern will be complex, deconvolution of a current trace to extract the underlying sequence will likely involve measuring a library of current levels for all possible combinations of 4-nt sequences. Furthermore, a method to confidently identify extended homopolymer regions must be developed. The path to commercial nanopore sequencing will involve other developments, including a robust platform and parallelization. However, we have shown that the high nucleotide sensitivity of MspA combined with the translocation control of phi29 DNAP enables single-nucleotide discrimination of DNA passing through a nanopore.

In Chapter 6, we observe the response of ssDNA immobilized in MspA as it experiences a variable applied voltage. By modeling DNA as a freely jointed chain and estimating the Brownian motion of the nucleotides, we estimate the effective charge of nucleotides, elasticity of DNA, movement of nucleotides within MspA's constriction, and forces applied to DNA while in MspA. These results will be valuable for the further development of a nanopore sequencing system since they provide insight into the interactions between DNA and MspA and can assist in the development of base calling algorithms.

Chapter 6

FORCE SPECTROSCOPY OF SINGLE-STRANDED DNA BY IMMOBILIZATION IN MSPA

This work is part of a manuscript in preparation to be published.

In previous chapters, I introduced *Mycobacterium smegmatis* porin A (MspA), and mutants thereof, as biological nanopore sensors with high signal-to-noise resolution of DNA nucleotides. In this chapter, I describe experiments in which a biological nanopore was used as a single-molecule sensor to stretch DNA strands. For this work, we performed force spectroscopy with an MspA mutant on single-stranded DNA attached to a large NeutrAvidin molecule. As in the experiments described in Chapter 4, a single DNA strand was driven into MspA until the NeutrAvidin came to rest on the rim of the pore, prohibiting further translocation. The nucleotides within the pore's constriction experienced a force proportional to the applied voltage. By varying the voltage applied across the pore, we detected stretching of the DNA with single nucleotide precision. Using a freely jointed chain model, we extracted the force applied to the DNA strand as well as the charge density and elasticity of DNA within MspA. These experiments demonstrate MspA as a single molecule tool to quantitatively study mechanical properties of DNA and other molecules.

6.1 Introduction

Biological nanopores are frequently used as a single molecule tool to detect individual DNA nucleotides [57, 36, 15] as well as to study protein folding [46, 40, 45] and peptide structure [55] by observing modulations in ionic current while the molecule interacts with the pore [31].

We previously introduced mutants of *Mycobacterium smegmatis* porin A (MspA) which demonstrated sensitivity to all four standard DNA bases as well as methylations [37] (Chapters 3 and 4), allowing DNA sequencing [36, 15] (Chapters 3 and 5). Additionally, when a polymerase was used to control the passage of DNA through the nanopore, we observed sequence dependent steps in ionic current [36] (Chapter 5). In this work, we use the M1-MspA mutant, hereafter referred to as simply 'MspA', in which three negatively charged aspartic acids were replaced with asparagines, creating a neutral constriction [10] (Chapter 2). Here we demonstrate that the sensitivity of MspA to individual DNA nucleotides can be used as a tool to study individual single-stranded DNA (ssDNA) molecules.

We performed force spectroscopy on ssDNA held in an MspA nanopore in a 1M KCl electrolyte solution at pH=8.0. Biotinylated ssDNA was attached to a NeutrAvidin molecule, which has a diameter larger than the entrance of MspA. The DNA was drawn into MspA by an applied voltage of 180 mV. The DNA threaded through the pore's constriction until the NeutrAvidin anchor stopped at the rim of MspA and prevented further translocation (See Figure 6.1). While the DNA was immobilized in the pore, the applied voltage was decreased in 20 mV steps from 180 mV to 80 mV with each voltage applied for 250 ms. The resulting ionic current through the DNA-pore complex was recorded.

Altering the applied voltage across the nanopore modulated the electric field in the pore constriction. The single, narrow, and short constriction of MspA focused the electric field onto a few nucleotides. The force, F, on DNA within an electric field, E, is $F = qE$ where q is the total effective charge of DNA nucleotides in the nanopore. The charge of a DNA nucleotide at physiological pH is $\sim 1e^-$ per phosphate of the DNA backbone. However, at high salt conditions the effective charge is reduced due to charge shielding. By observing the ionic current response of the DNA-pore system as the voltage is modulated, we estimated the force on the DNA within the pore constriction as well as the effective charge and elasticity of DNA within the MspA

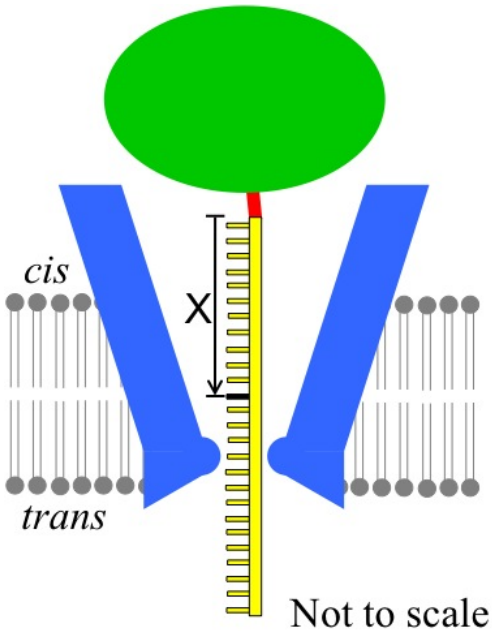


Figure 6.1: **Schematic Diagram of DNA Immobilized in MspA.** Cartoon of MspA (blue) set up in a lipid bilayer (grey) in an electrolyte solution (not shown). Single stranded DNA (ssDNA, yellow) is attached to a bulky NeutrAvidin molecule (green) with a biotin linker (red). A specific nucleotide (black) is designated by the position, X, measured from the biotin-NeutrAvidin 'anchor' where the first nucleotide after the biotin is X=1. Under an applied voltage (*trans* +), the ssDNA threads through MspA from the *cis* side and remains immobilized in the pore until the applied voltage is reversed (*trans* -). The nucleotides residing in the narrowest part of MspA, the constriction, modulate the ionic current through the pore.

nanopore.

6.2 Results

The 'Open Pore' resistance of MspA at each applied voltage was determined without DNA present to be $\approx 0.56 \pm 0.01 G\Omega$ (Figure 6.2, blue line), in agreement with previous results [10, 15, 37, 36].

First, homopolymeric adenine (poly-dA), cytosine (poly-dC), and thymine (poly-dT) were held within the pore in both orientations ($3'$ and $5'$ ends threaded through

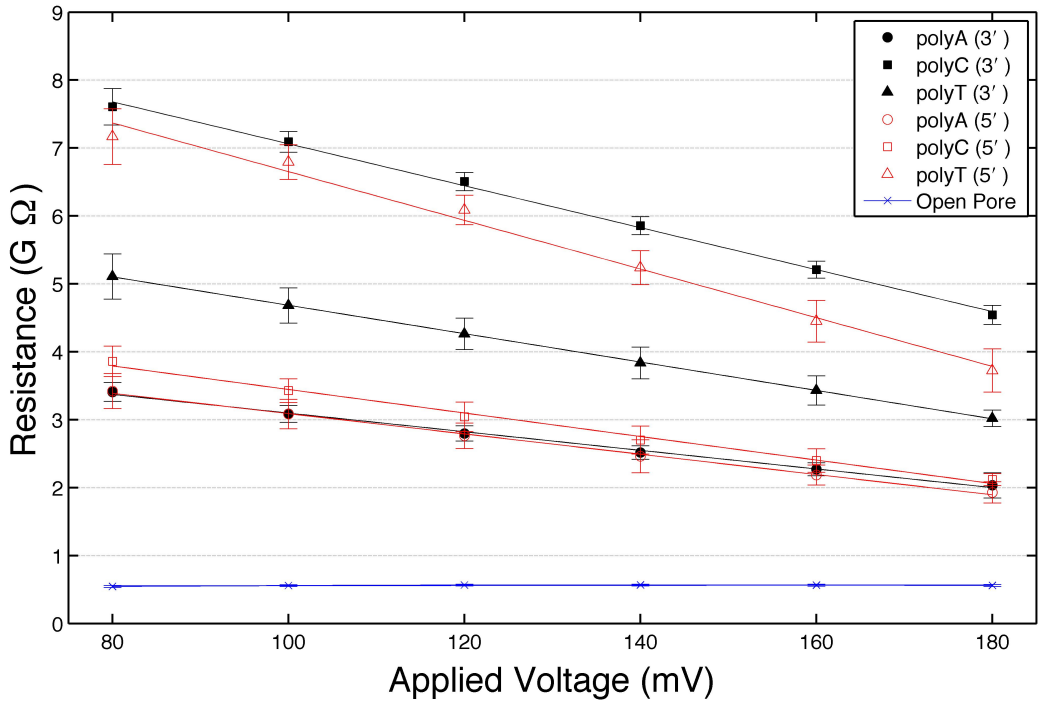


Figure 6.2: **Experimental Results for Homopolymer DNA Immobilized in MspA at Varying Voltages.** The voltage dependent resistance of the DNA-pore system is plotted for homopolymer adenine ('poly-dA', circles), cytosine ('poly-dC', squares), and thymine ('poly-dT', triangles) for both 3′ threading (black) and 5′ threading (red) orientations. The resistance of MspA without DNA present is also plotted ('Open Pore', blue x's). Fitted curves are shown to aid the eye.

the pore). Homopolymer guanine was not available due to G-tetrad structure formation [24]. The voltage dependent resistance for each DNA-pore complex is shown in Figure 6.2. Homopolymer thymine and cytosine in the 3′ leading orientation produced the highest resistance. Both orientations of homopolymer adenine produced similar resistances, which were found to be the lowest of all strands. Homopolymer cytosine with 5′ leading had a resistance similar to poly-dA. The resistance of homopolymer cytosine was strongly orientation dependent with the 3′ leading direction having a much higher resistance than that for 5′ leading.

Next, we studied the elasticity of ssDNA using homopolymer adenine strands in the $3'$ leading orientation. Each strand had a single dC nucleotide substitution at a position that was varied between the 11-18 nucleotide (X) from the biotin anchor. The voltage dependent resistance of each strand was compared to that of poly-dA, found above (see Figure 6.3). The plotted histograms show the contribution of each nucleotide position on the measured DNA-pore resistance. We define MspA's recognition site as the location on a DNA strand near MspA's constriction where nucleotides most affect the DNA-pore resistance. The central nucleotide and the width of the recognition site are taken as the peak position and full width at half max (FWHM) of the fitted Gaussian curve, respectively (see Figure 6.5). At lower voltages, the principle nucleotide affecting the current is positioned at $X \sim 16$. Nucleotides located before the 14th position had little to no effect on the resistance, as they were located above MspA's recognition site. At higher voltages, the nucleotide most affecting the current changed to $X \sim 14$ indicating that the DNA had shifted further towards the *trans* side of the bilayer. In addition to shifting the central nucleotide within the constriction, there is a slight reduction in the width of the recognition site with higher voltages (see Figure 6.5).

The average fluctuation of the DNA-pore resistance within an event was also examined (Figure 6.4). As expected, the inter-event fluctuations in resistance were highest when the dC substitution was within the recognition site since the DNA was not stationary but instead stochastically drifted during the course of the experiment. At lower voltages, the fluctuation of the resistance within an event was larger than at higher voltages, suggesting that the strand is more taught at higher voltages.

Next, we repeated the single nucleotide substitution experiments with a single dA nucleotide substituted at positions X=13-16 in an otherwise poly-dT strand (See Figure C.3 of Appendix C). A similar trend was observed where the central nucleotide in the constriction varied from $X \sim 14$ for 180 mV to $X \sim 16$ at 80 mV (see Figure 6.5). However, the number of nucleotides affecting the current did not change

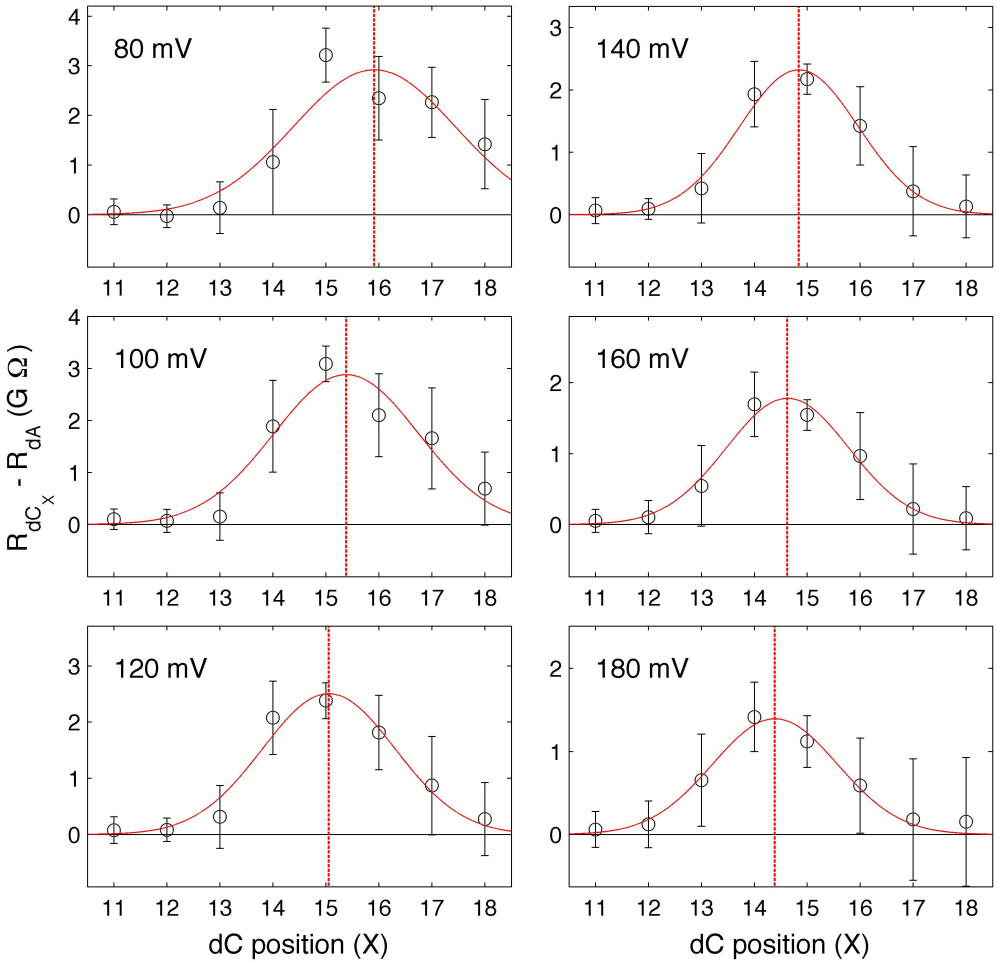


Figure 6.3: **Experimental Results for Single Nucleotide Substitutions in a Poly-dA Strand Immobilized in MspA at Varying Voltages.** Comparison of resistance for DNA with a single cytosine (dC) substitution at position X in an otherwise poly-dA strand. The deviation of resistance from that of poly-dA is shown. The top of each plot corresponds to the resistance for a poly-dC strand. The red vertical dashed line indicates the peak of the fitted Gaussian curve.

appreciably with voltage.

Finally, we studied a genomic DNA sequence known to contain a single nucleotide polymorphism (SNP). This heteromeric DNA sequence had a particular nucleotide

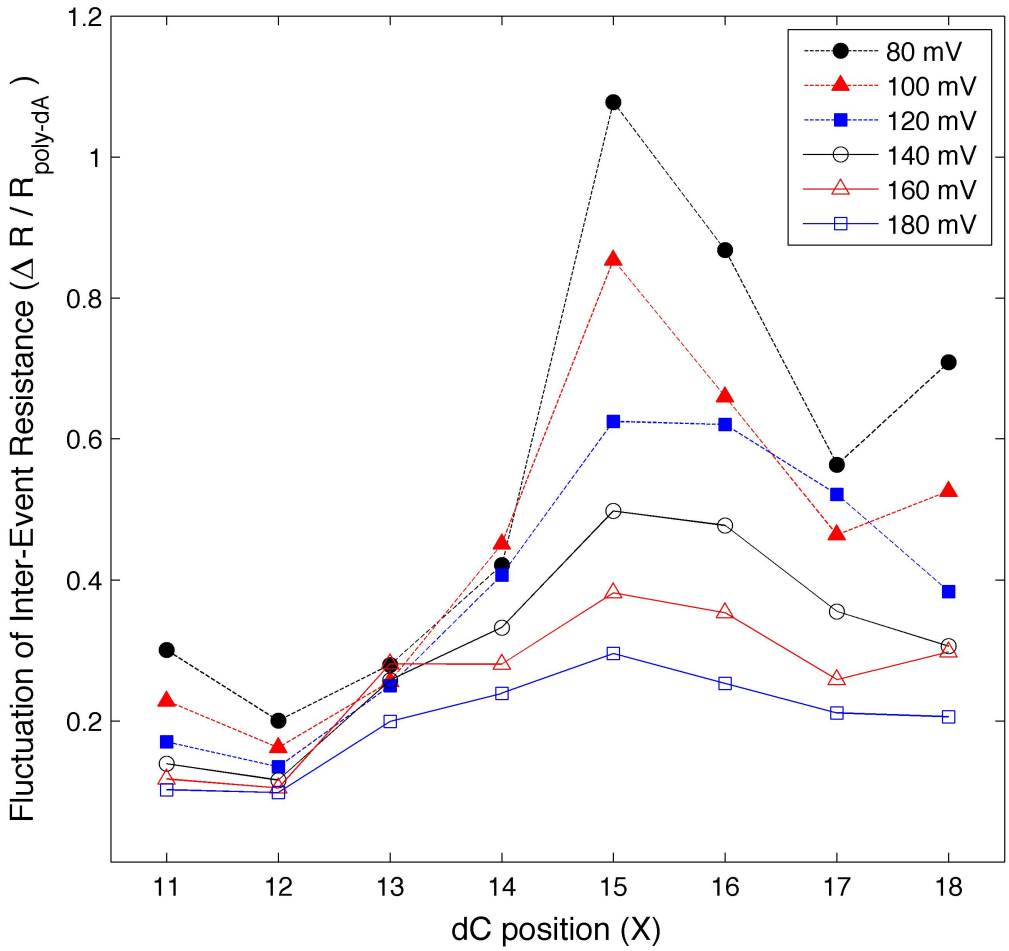


Figure 6.4: **Fluctuations of Inter-Event Resistance for DNA Strands Immobilized in MspA at Varying Voltages.** The average fluctuations of the DNA-pore resistance within an event are shown above for a strands containing a single dC substitution at position X within a poly-dA strand. The fluctuations are taken as the standard deviations of the resistance for each event and normalized by the resistance for homopolymer adenine.

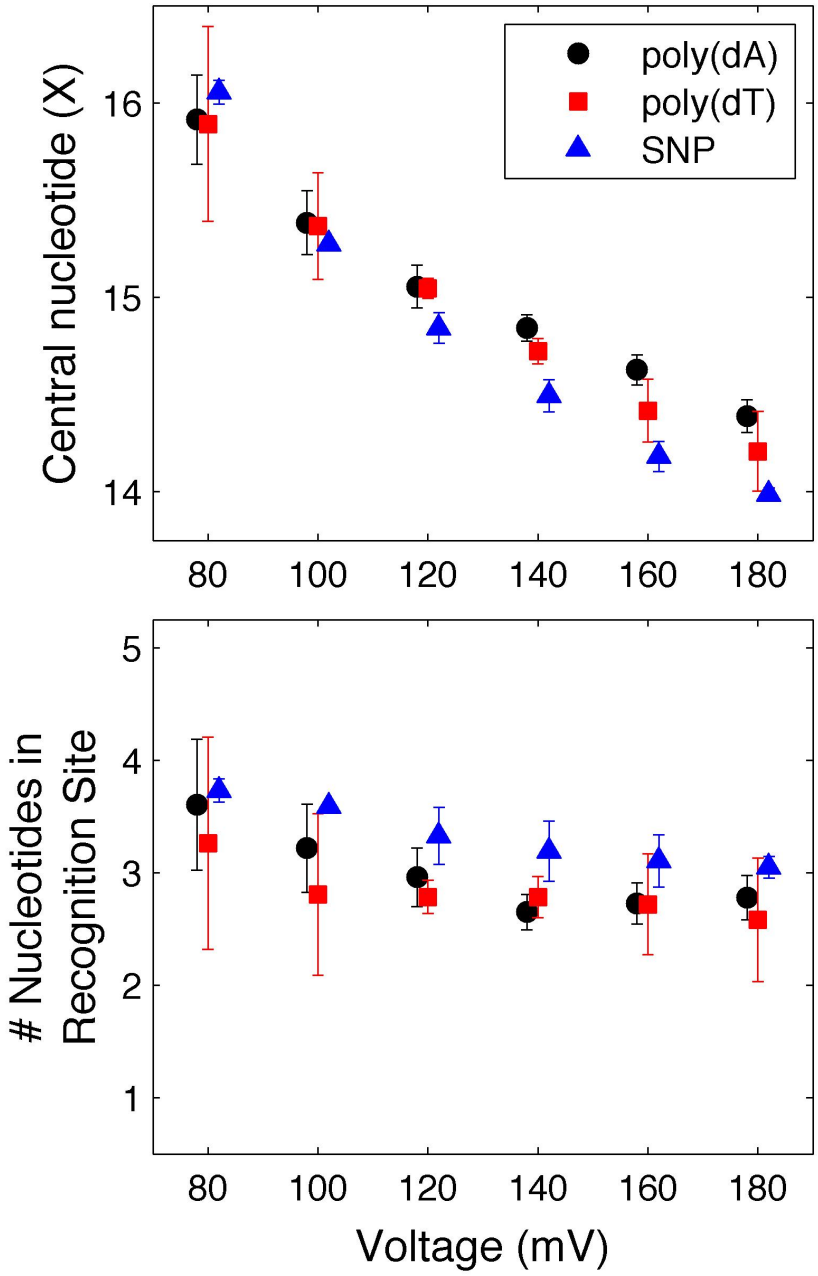


Figure 6.5: **Positioning of DNA within MspA's Recognition Site when Immobilized at Varying Voltages.** The nucleotides residing in MspA's recognition site are shown at each voltage for poly-dA (black circles), poly-dT (red squares), and SNP strands (blue triangles). The central nucleotide (upper) is taken from the peak of the fitted Gaussian curve as shown in Figure 6.3. The number of nucleotides in MspA's recognition site (lower) is the FWHM of the Gaussian curve. The error bars indicate the standard errors in the Gaussian fit for each parameter. Voltage dependent stretching of the DNA within MspA is observed by the voltage dependent shifting of the nucleotides in MspA's constriction.

that was either a dC or a dA (See Figure C.4 of Appendix C for sequence). By incrementally adding nucleotides at the NeutrAvidin end, the SNP was positioned at X=13-16. As with the other strands, the central nucleotide in MspA's recognition site shifted from $X \sim 16$ at low voltages to $X \sim 14$ at higher voltages (Figure C.4 of Appendix C). Figure 6.5 compares the nucleotides within MspA's recognition site for the three different DNA strands.

6.3 Analysis

Our results indicate that the longitudinal positioning of DNA held within MspA is voltage dependent. At 180 mV, the nucleotide most effecting the DNA-pore resistance is located at $X \sim 14$ from the biotin anchor. As the voltage decreases, the DNA nucleotides shift upward so that the $X \sim 16$ is centered in the recognition site and the number of nucleotides involved in controlling the ionic current slightly increases (Figure 6.5).

The observed shifting of nucleotides in MspA's recognition site with varying voltages is due to either elongation of the DNA strand or changes in positioning of the NeutrAvidin anchor with respect to MspA's constriction. Elongation of the ssDNA under force may be caused by undoing of base stacking or straightening of the strand. In either case, the elasticity of the ssDNA would be strongly voltage dependent. Alternately, the position of the NeutrAvidin anchor may change with varying voltages resulting in the entire DNA strand shifting while the conformation remained unchanged. This shifting may be due to changes in the positioning of the NeutrAvidin on the pore or deformation of the rim of MspA. In either case, we would expect the elasticity of the ssDNA to be voltage independent. The fluctuations observed in DNA-pore resistance within a capture event show a strong voltage dependence (ie force dependence) (Figure 6.4). This suggests that the shifting of nucleotides within MspA's recognition site was primarily due to elongation of the strand rather than repositioning of the biotin.

Three different DNA strands were examined and each showed the voltage dependent shifting of nucleotides in MspA's recognition site (Figure 6.5). Since poly-dA is known to form secondary structure, where poly-dT and heterogeneous strands do not, base unstacking cannot explain the entirety of the observed effect. Therefore, the observed shifting of the nucleotides with varying voltage is mostly due to straightening of the strand. At low voltages, the central nucleotide in the recognition site was comparable for each poly-dA, poly-dT and SNP strands indicating that they all have similar conformations. With increased force, however, poly-dA was elongated to a lesser extent than poly-dT or SNP strands indicating that poly-dA was stiffer than either of the other two strands. This is expected due to the secondary structure of poly-dA.

The electric field in MspA's constriction can be estimated to be $E = \frac{V}{a}$, where V is the electrostatic potential drop and a is the length over which the potential changes. The force on the DNA within the electric field is then:

$$F = \frac{\lambda s V}{a} \tag{6.1}$$

where λ is the linear charge density and s is the contour length of the strand.

Assuming that the FWHM value from experiments, n, is proportional to the contour length, s, of DNA within the electric field, and taking the linear charge density, λ, and length of the electric field, a, as voltage independent, forces at two voltages, V_1 and V_2, can be related by:

$$\frac{F_1}{F_2} = \frac{n_1}{n_2} * \frac{V_1}{V_2} \tag{6.2}$$

The extension, x, of ssDNA under an applied force, F, can be modeled as a freely jointed chain (FJC) with an added stretch modulus [52]:

$$x = L * [coth(\frac{Fb}{k_b T}) - (\frac{k_b T}{Fb})](1 + \frac{F}{S}) \tag{6.3}$$

where x is the end-to-end distance of the strand, L is the contour length, k_b is the Boltzmann constant, T is the temperature, S is the stretch modulus, and b is the Kuhn length. At room temperature, the product k_bT is 4.11 pN-nm. At a 1 M salt concentration, the Kuhn length, b, of ssDNA is approximately 3 nm [43] and the stretch modulus, S, is taken as 800 pN [52]. The contour length, L, is extracted from experiment as the length of DNA between the NeutrAvidin anchor and the recognition site, assuming an interphosphate distance of 0.56 nm.

As discussed above, the seating of NeutrAvidin on the rim of MspA is not expected to significantly change as the experimental voltage is altered. Therefore, the end-to-end distance of the DNA within MspA's vestibule, x, was constant whereas the contour length of the strand, L, within this region changed with varying voltage. The applied force and elongation of DNA, x, was found with the FJC model by comparing the force-extension curves (Equation 6.3) at two voltages using the force relationship described above (Equation 6.2). See Appendinx C for details. The determined forces for each DNA strand are shown in Figure 6.6. Poly-dT and SNP produced similar forces over the range of applied voltages. Poly-dA, however, produced higher forces over the same range.

The linear charge density of DNA within the field was then determined from Equation 6.1, $\lambda = \frac{Fa}{sV}$, assuming the curvature of DNA within the constriction is similar to that within the vestibule. The quantity $\frac{a}{s}$ was therefore estimated to be $\frac{x}{L}$, which was determined from the FJC model above. See Appendix C for details. We found that the linear charge density of poly-dA was $\sim 0.29\frac{e^-}{nt}$ and the charge density of poly-dT and SNP were both $\sim 0.22\frac{e^-}{nt}$. The larger charge density of poly-dA accounts for the discrepancy in forces observed in Figure 6.6.

As previously discussed, the FWHM of a Gaussian curve fitted to experimental results was used to determine how many nucleotides redided in MspA's recognition site. The width of this Gaussian curve had contributions from both MspA's constriction size and Brownian motion of nucleotides. We assumed that the applied voltage

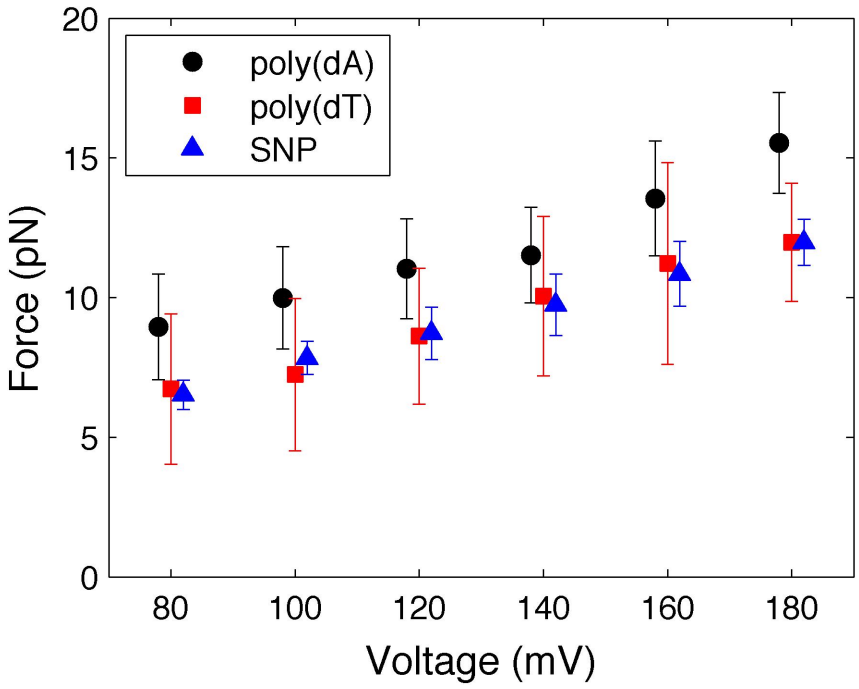


Figure 6.6: **Voltage Dependent Force on DNA Immobilized in MspA.** The force applied to DNA in MspA's constriction is calculated at each applied voltage using the FJC model and force relationship described in the text (Equations 6.2 - 6.3) and Appendix C.

dropped entirely over MspA's constriction so that only nucleotides within this region experienced a force. However, Brownian motion caused nucleotides around the constriction to move in and out of the electric field making the number of nucleotides affecting the DNA-pore resistance to be larger than expected for a stationary system. A DNA strand immobilized in MspA can be modeled as a spring with a linear restoring force such that $\kappa = \frac{F}{x}$, where F is the applied force, x is the resulting displacement, and κ is the spring constant. Brownian motion fluctuations caused the DNA to move around its equilibrium position. From the equipartition theorem, the variance in the displacement from equilibrium can be related to the spring constant, κ, of the DNA by $\frac{1}{2}k_bT = \frac{1}{2}\kappa < x^2 >$ where $< x^2 >$ is the variance in the displacement

from equilibrium, k_b is the Boltzmann constant, and T is temperature. The variance of the DNA due to Brownian motion is then:

$$< x^2 >= \frac{k_b T}{\kappa} \tag{6.4}$$

The spring constant, κ, is found from the force extension relation (Equation 6.3) and is plotted for each strand in Figure 6.7.

The Brownian motion contribution can be compared to the width of the recognition site determined from experiment. See Appendix C for details. The results for experiments using poly-dA, poly-dT and SNP are shown in Figure C.5 of Appendix C. For all strands, the contribution from Brownian motion account for the majority

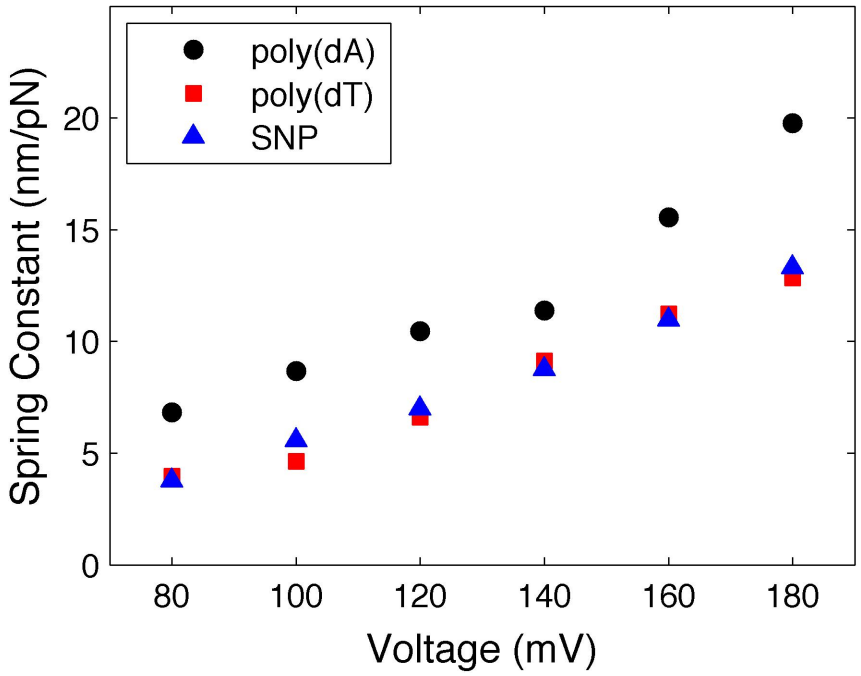


Figure 6.7: **Voltage Dependent Elasticity of DNA Immobilized in MspA.** The spring constant is determined from the FJC model (Equation 6.3) using the experimentally derived contour length of the ssDNA, L, as described in the text.

of the width of the recognition site. For the poly-dA strand, we find that the mean effective length of MspA's constriction was $\approx 0.79\pm 0.23$ nm. This agrees favorably with that expected from geometry. For poly-dT and SNP strands under low voltage, the recognition site was smaller than the expected contribution from Brownian motion suggesting that either unstacked strands cannot be modeled in this way or the nucleotides interact with the pore constraining their motion.

6.4 Discussion

Adenine is the largest nucleotide and forms a tight helical structure [24]. An overstretching transition occurs at an applied force of 23 pN [32, 9] corresponding to the removal of secondary structure. The experimental forces applied with the MspA nanopore are well below this threshold, so we expect that the secondary structure of poly-dA was intact. This was supported by the absence of a voltage dependent step in the observed resistance for poly-dA (Figure 6.2), which would be expected if the overstretching transition were crossed in the course of our experiments. The low observed resistance of poly-dA in both $3'$ and $5'$ leading directions seen in Figure 6.2 suggests that the tight helical structure caused minimal impedance of the ions through the nanopore. Thymine nucleotides, on the other hand, form no observable secondary structure [32]. We found that the number of nucleotides within the constriction with poly-dT was slightly lower than that found with poly-dA (Figure 6.5). The high resistance of poly-dT (Figure 6.2), then, was not a result of more nucleotides within the constriction. This suggests that the primary impedance for ions through a nanopore is governed by the nucleotide structure where strands without secondary structure produce a larger hindrance for the moving ions.

The two unstacked strands, poly-dT and SNP, behaved similarly under the FJC model. The strands experienced similar forces at a given applied voltage. Additionally, they had the same linear charge density and elasticity. Compared to unstacked strands, the poly-dA strand provided a higher concentration of charge in the constric-

tion resulting in a larger force applied to the DNA strand for a given voltage (Figure 6.6). As expected, the stacked poly-dA strand was stiffer, resulting in a higher spring constant (Figure 6.7).

Comparing our values for the effective charge within MspA to previous studies on the α-hemolysin nanopore, which was found to have an effective charge of $0.1e^-$ [51, 38, 60] per nucleotide (1M salt), we find a larger charge density. Assuming an interphosphate distance of 0.56 nm, the charge density of DNA in MspA is $\approx 0.2 - 0.3\frac{e^-}{nt}$ depending on the stacking on the strand. This value is larger than that found with α-hemolysin [51, 38, 60] and is likely due to geometrical differences between the pores reducing the charge shielding of DNA in MspA. The recognition site of α-hemolysin is less localized due to its long β-barrel stem, resulting in a lower electric field for any given voltage when compared to MspA. Additionally, the diameter of α-hemolysin's β-barrel is larger than MspA's constriction. This increased width is likely responsible for the greater charge shielding in α-hemolysin since water molecules can more easily surround the DNA strand.

Using our force estimations and experimental results, we found that the number of bases affecting the ionic current through MspA is dictated by Brownian motion of the nucleotides and not the length on MspA's constriction. This suggests that further mutations to shorten MspA's constriction will not significantly decrease the length of the recognition site. To decrease the number of nucleotides affecting the ionic current signal, Brownian motion must be reduced. Therefore, further modifications to MspA will likely include the addition of DNA-pore interactions.

By varying the applied voltage on a DNA molecule held in an MspA nanopore, we have estimated the experimental force applied to ssDNA by MspA as well as the effective charge and elasticity of ssDNA in the nanopore. This system gave us the unique opportunity to explore DNA properties while spatially confined. The MspA nanopore is a precise, versatile and simple new single molecule technique that can provide a wide range of information on single stranded DNA and as well as other

polymers.

Chapter 7
CONCLUSIONS / OUTLOOK

The biological porin MspA has proven to be a valuable single-molecule tool able to read DNA. Using both duplex and NeutrAvidin anchored DNA, we found that the nucleotide specific current levels were well resolved and approximately 4 nucleotides effected the measured ionic current. Additionally, methylation sensitivity was discovered and single nucleotide polymorphisms were detectable. We also introduced a DNA sequencing technique using a molecular motor to controllably pass DNA through MspA's constriction. We demonstrated that nucleotide specific steps in the ionic current signal were detectable when DNA was processively drawn through the pore by a polymerase. These steps could be related back to the sequence. This body of work demonstrates MspA's potential as a next generation sequencing technology.

In order to realize nanopore sequencing with MspA as an industrial technique, several more issues must be addressed.

- Base-calling algorithms must to be developed that are able to translate ionic current reads into a DNA sequence. We have shown that enzyme controlled DNA translocation through MspA produces an ionic current signature mappable to the known sequence. However, an algorithm to identify an unknown sequence from a current trace is still in development. Roughly 4 nucleotides within MspA's constriction contribute to the measured ionic current. Additionally, there are sequence specific interactions between the DNA and pore that are not well understood. However, this research has laid the groundwork for the development of an algorithm to identify DNA sequence from a given current trace.

- Additional epigenetic modifications must be read as part of a longer sequence. While immobilized in MspA, strands containing methylated cytosine were distinguishable from unmethylated cytosine. These experiments must be expanded to explore additional epigenetic modifications and to ensure that they can be read as part of a larger sequence moving through MspA.

- Longer strands of DNA must be examined. One important benefit of nanopore sequencing is its potential for long DNA reads. The experiments described in this dissertation contained DNA templates of fewer than 100 nucleotides. In order to compete with other upcoming technologies, longer DNA strands must be studied. Longer DNA strands will likely require additional enzymes, such as histones, to organize the DNA within the *cis* volume and assist in DNA threading through MspA.

- Robust setups must be able to run in parallel so that multiple pores can be examined at once. Although a single DNA translocation may provide sequence information, analysis of multiple reads is required to enhance precision. Parallelization of the system will allow simultaneous DNA reads and increase sequencing throughput. Additionally, present MspA nanopore setups utilize a fragile lipid bilayer. In order to provide reliable DNA sequencing, modifications must to made to the membrane to ensure robust bilayers are maintained throughout the read.

Nanopore sequencing using MspA is very promising. With successful implementation of the above recommendations, a nanopore sequencing device will likely be commercially available in the next couple years.

BIBLIOGRAPHY

[1] L.T. Amundadottir, P. Sulem, J. Gudmundsson, A. Helgason, A. Baker, B.A. Agnarsson, A. Sigurdsson, K.R. Benediktsdottir, J.B. Cazier, J. Sainz, M. Jakobsdottir, J. Kostic, D.N. Magnusdottir, S. Ghosh, K. Agnarsson, B. Birgisdottir, L. Le Roux, A. Olafsdottir, T. Blondal, M. Andresdottir, O.S. Gretarsdottir, J.T. Bergthorsson, D. Gudbjartsson, A. Gylfason, G. Thorleifsson, A. Manolescu, K. Kristjansson, G. Geirsson, H. Isaksson, J. Douglas, J.E. Johansson, K. Bälter, F. Wiklund, J.E. Montie, X. Yu, B.K. Suarez, C. Ober, K.A. Cooney, H. Gronberg, W.J. Catalona, G.V. Einarsson, R.B. Barkardottir, J.R. Gulcher, A. Kong, U. Thorsteinsdottir, and K. Stefansson. A common variant associated with prostate cancer in european and african populations. *Nature genetics*, 38:652–658, 2006.

[2] N. Ashkenasy, J. Sanchez-Quesada, H. Bayley, and M. R. Ghadiri. Recognizing a single base in an individual dna strand: a step toward dna sequencing in nanopores. *Angew Chem Int Ed Engl*, 44(9):1401–4, 2005.

[3] S. Benner, R. J. A. Chen, N. A. Wilson, R. Abu-Shumays, N. Hurt, K. R. Lieberman, D. W. Deamer, W. B. Dunbar, and M. Akeson. Sequence-specific detection of individual dna polymerase complexes in real time using a nanopore. *Nature Nanotechnology*, 2(11):718–724, 2007.

[4] M. Berndt, M. Lorenz, J. Enderlein, and S. Diez. Axial nanometer distances measured by fluorenscence lifetime imaging microscopy. *Nano Letters*, 10:1497–1500, 2010.

[5] L. Blanco, A. Bernad, J. M. Lazaro, G. Martin, C. Garmendia, and M. Salas. Highly efficient dna synthesis by phage phi29 dna polymerase. *J. Biol. Chem.*, 264(15):8935–8940, 1989.

[6] L. Blanco, A. Bernad, and M. Salas. Transition from initiation to elongation in protein-primed phi-29 dna replication: Salt-dependent stimulation by the viral protein p6. *Journal of Virology*, 62(11):4167–4172, 1988.

[7] L. Blanco and M. Salas. Relating structure to function in phi29 dna polymerase. *Journal of Biological Chemistry*, 271(15):8509–8512, 1996.

[8] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. H. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss. The potential and challenges of nanopore sequencing. *Nature Biotechnology*, 26(10):1146–1153, 2008.

[9] A. Buhot and A. Halperin. Effects of stacking on the configurations and elasticity of single-stranded nucleic acids. *Physical Review E*, 70(020902):1–4, 2004.

[10] T. Z. Butler, M. Pavlenok, I. M. Derrington, M. Niederweis, and J. H. Gundlach. Single-molecule dna detection with an engineered mspa protein nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, 105(52):20647–20652, 2008.

[11] G. M. Cherf, K. R. Lieberman, H. Rashid, C. E. Lam, K. Karplus, and M. Akeson. Automated forward and reverse ratcheting of dna in a nanopore at five angstrom precision. *Nat Biotech*, this volume, 2011.

[12] S. L. Cockroft, J. Chu, M. Amorin, and M. R. Ghadiri. A single-molecule nanopore device detects dna polymerase activity with single-nucleotide resolution. *J Am Chem Soc*, 130(3):818–20, 2008.

[13] W. Coulter. Means for counting particles suspended in a fluid. US Patent 2656508, 1953.

[14] I. Derrington. *There's a Hole in the Bucket: DNA Sequencing with the Nanopore MspA*. University of Washington, 2011.

[15] I. M. Derrington, T. Z. Butler, M. D. Collins, E. Manrao, M. Pavlenok, M. Niederweis, and J. H. Gundlach. Nanopore dna sequencing with mspa. *Proc Natl Acad Sci U S A*, 107(37):16060–16065, 2010.

[16] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, 11th edition, 2006.

[17] D.F. Easton, K.A. Pooley, A.M. Dunning, P.D. Pharoah, D. Thompson, D.G. Ballinger, J.P. Struewing, J. Morrison, H. Field, R. Luben, N. Wareham, S. Ahmed, C.S. Healey, R. Bowman, SEARCH collaborators, K.B. Meyer, C.A. Haiman, L.K. Kolonel, B.E. Henderson, L. Le Marchand, P. Brennan, S. Sangrajrang, V. Gaborieau, F. Odefrey, C.Y. Shen, P.E. Wu, H.C. Wang, D. Eccles, D.G. Evans, J. Peto, O. Fletcher, N. Johnson, S. Seal, M.R. Stratton,

N. Rahman, G. Chenevix-Trench, S.E. Bojesen, B.G. Nordestgaard, C.K. Axelsson, M. Garcia-Closas, L. Brinton, S. Chanock, J. Lissowska, B. Peplonska, H. Nevanlinna, R. Fagerholm, H. Eerola, D. Kang, K.Y. Yoo, D.Y. Noh, S.H. Ahn, D.J. Hunter, S.E. Hankinson, D.G. Cox, P. Hall, S. Wedren, J. Liu, Y.L. Low, N. Bogdanova, P. Schürmann, T. Dörk, R.A. Tollenaar, C.E. Jacobi, P. Devilee, J.G. Klijn, A.J. Sigurdson, M.M. Doody, B.H. Alexander, J. Zhang, A. Cox, I.W. Brock, G. MacPherson, M.W. Reed, F.J. Couch, E.L. Goode, J.E. Olson, H. Meijers-Heijboer, A. van den Ouweland, A. Uitterlinden, F. Rivadeneira, R.L. Milne, G. Ribas, A. Gonzalez-Neira, J. Benitez, J.L. Hopper, M. McCredie, M. Southey, G.G. Giles, C. Schroen, C. Justenhoven, H. Brauch, U. Hamann, Y.D. Ko, A.B. Spurdle, J. Beesley, X. Chen, kConFab, AOCS Management Group, A. Mannermaa, V.M. Kosma, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447:1087–1093, 2007.

[18] M. Faller, M. Niederweis, and G. E. Schulz. The structure of a mycobacterial outer-membrane channel. *Science*, 303(5661):1189–92, 2004.

[19] M.L. Freedman, C.A. Haiman, N. Patterson, G.J. McDonald, A. Tandon, A. Waliszewska, K. Penney, R.G. Steen, K. Ardlie, E.M. John, I. Oakley-Girvan, A.S. Whittemore, K.A. Cooney, A.S. Inglesd, D. Altshuler, B.E. Hendersond, and D. Reich. Admixture mapping identifies 8q24 as a prostate cancer risk locus in african-american men. *Proceedings of the National Academy of Sciences of the United States of America*, 103(38):14068–14073, 2006.

[20] Carl W. Fuller, Lyle R. Middendorf, Steven A. Benner, George M. Church, Timothy Harris, Xiaohua Huang, Stevan B. Jovanovich, John R. Nelson, Jeffery A. Schloss, David C. Schwartz, and Dmitri V. Vezenov. The challenges of sequencing by synthesis. *Nat Biotech*, 27(11):1013–1023, 2009.

[21] Adrian J. Gibbs and George A. McIntyre. The diagram, a method for comparing sequences. *European Journal of Biochemistry*, 16(1):1–11, 1970.

[22] B. Gyarfas, F. Olasagasti, S. Benner, D. Garalde, K. R. Lieberman, and M. Akeson. Mapping the position of dna polymerase-bound dna templates in a nanopore at 5 angstrom resolution. *Acs Nano*, 3(6):1457–1466, 2009.

[23] O.P. Hamill, A. Marty, E. Neher, B. Sakmann, and F.J. Sigworth. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Pflugers Archiv-European Journal of Physiology*, 391:85–100, 1981.

[24] D.M. Hatters, L. Wilson, B.W. Atcliffe, T.D. Mulhern, N. Guzzo-Pernell, and G.J. Howlett. Sedimentation analysis of novel dna structures formed by homo-oligonucleotides. *Biophysical Journal*, 81:371–381, July 2001.

[25] C. Heinz, H. Engelhardt, and M. Niederweis. The core of the tetrameric mycobacterial porin mspa is an extremely stable beta-sheet domain. *J Biol Chem*, 278(10):8678–85, 2003.

[26] C. Heinz, E. Roth, and M. Niederweis. Purification of porins from mycobacterium smegmatis. *Methods of Molecular Biology*, 228(139-150), 2003.

[27] C. Hoffmann, A. Leis, M. Niederweis, J. M. Plitzko, and H. Engelhardt. Disclosure of the mycobacterial outer membrane: cryo-electron tomography and vitreous sections reveal the lipid bilayer structure. *Proc Natl Acad Sci U S A*, 105(10):3963–7, 2008.

[28] N. Hurt, H. Y. Wang, M. Akeson, and K. R. Lieberman. Specific nucleotide binding and rebinding to individual dna polymerase complexes captured on a nanopore. *Journal of the American Chemical Society*, 131(10):3772–3778, 2009.

[29] B. Ibarra, Y. R. Chemla, S. Plyasunov, S. B. Smith, J. M. Lazaro, M. Salas, and C. Bustamante. Proofreading dynamics of a processive dna polymerase. *EMBO J.*, 28:2794–2802, 2009.

[30] A. Kahvejian, J. Quackenbush, and J.F. Thompson. What would you do if you could sequence everything? *Nature Biotechnology*, 26(10):1125–1133, 2008.

[31] J. J. Kasianowicz, E. Brandin, D. Branton, and D. W. Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A*, 93(24):13770–3, 1996.

[32] C. Ke, M. Humeniuk, H. S-Gracz, and P.E. Marszalek. Direct measurements of base stacking interactions in dna by single-molecule atomic-force spectroscopy. *Physical Review Letters*, 99:018302, 2007.

[33] P. Lexow. Sequencing method using magnifying tags. Int Patent WO 2000/39333, 2004.

[34] P. Lexow. Method for producing and amplified polynucleotide sequence. Int Patent WO 2006/092582, 2006.

[35] K. R. Lieberman, G. M. Cherf, M.J. Doody, F. Olasagasti, Y. Kolodji, and M. Akeson. Processive replication of single dna molecules in a nanopore catalyzed by phi29 dna polymerase. *J Am Chem Soc*, 132(50):17961–17972, 2010.

[36] E. A. Manrao, I. M. Derrington, A. H. Laszlo, K. W. Langford, M. K. Hopper, N. Gillgren, M. Pavlenok, M. Niederweis, and J. H. Gundlach. Reading dna at single-nucleotide resolution with a mutant mspa nanopore and phi29 dna polymerase. *Nature Biotechnology*, 30:349–353, March 2012.

[37] E. A. Manrao, I. M. Derrington, M. Pavlenok, M. Niederweis, and J. H. Gundlach. Nucleotide discrimination with dna immobilized in the mspa nanopore. *PLoS ONE*, 6(10):e25723, 2011.

[38] J. Mathe, H. Visram, V. Viasnoff, Y. Rabin, and A. Meller. Nanopore unzipping of individual dna hairpin molecules. *Biophys J*, 87(5):3205–12, 2004.

[39] A. Meller, L. Nivon, E. Brandin, J. Golovchenko, and D. Branton. Rapid nanopore discrimination between single polynucleotide molecules. *Proc Natl Acad Sci U S A*, 97(3):1079–84, 2000.

[40] C. Merstorf, B. Cressiot, M. Pastoriza-Gallego, A. Oukhaled, J. M. Betton, L. Auvray, and J. Pelta. Wild type, mutant protein unfolding and phase transition detected by single-nanopore recording. *Acs Chemical Biology*, 7(4):652–658, January 2012.

[41] J.B. Mills, E. Vacano, and P.J. Hagerman. Flexibility of single-stranded dna: use of gapped duplex helices to determine the persistence lengths of poly(dt) and poly(da). *J Mol Biol.*, 285(1):245–257, 1999.

[42] N. Mitchell and S. Howorka. Chemical tags facilitate the sensing of individual dna strands with nanopores. *Angewandte Chemie-International Edition*, 47(30):5565–5568, 2008.

[43] M.C. Murphy, I. Rasnik, W. Cheng, T.M. Lohman, and T. Ha. Probing single-stranded dna conformational flexibility using fluorescence spectroscopy. *Biophysical Journal*, 86:2530–2537, 2004.

[44] M. Niederweis, S. Ehrt, C. Heinz, U. Klocker, S. Karosi, K. M. Swiderek, L. W. Riley, and R. Benz. Cloning of the mspa gene encoding a porin from mycobacterium smegmatis. *Mol Microbiol*, 33(5):933–45, 1999.

[45] J. Nivala, D.B. Marks, and M. Akeson. Unfoldase-mediated protein translocation through an -hemolysin nanopore. *Nature Biotechnology*, Advance Online Publication(doi:10.1038/nbt.2503), February 2013.

[46] G. Oukhaled, J. Mathe, A. L. Biance, L. Bacri, J. M. Betton, D. Lairez, J. Pelta, and L. Auvray. Unfolding of proteins and long transient conformations detected by single nanopore recording. *Physical Review Letters*, 98(15), 2007.

[47] R. F. Purnell, K. K. Mehta, and J. J. Schmidt. Nucleotide identification and orientation discrimination of dna homopolymers immobilized in a protein nanopore. *Nano Letters*, 8(9):3029–3034, 2008.

[48] Robert F. Purnell and Jacob J. Schmidt. Discrimination of single base substitutions in a dna strand immobilized in a biological nanopore. *ACS Nano*, 3(9):2533–2538, 2009.

[49] M. Salas, L. Blanco, J. M. Lazaro, and M. de Vega. The bacteriophage phi29 dna polymerase. *IUBMB Life*, 60(1):82–85, 2008.

[50] F. Sanger, S. Nicklen, and A.R. Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, 1977.

[51] A. F. Sauer-Budge, J. A. Nyamwanda, D. K. Lubensky, and D. Branton. Unzipping kinetics of double-stranded dna in a nanopore. *Phys Rev Lett*, 90(23):238101, 2003.

[52] S.B. Smith, Y. Cui, and C. Bustamante. Overstretching b-dna: The elastic response of individual double-stranded and single-stranded dna molecules. *Science*, 271:795–799, 1996.

[53] M. S. Soengas, J.A. Esteban, M. Salas, and C. Gutierrez. Complex formation between phage phi-29 single-stranded dna binding protein and dna. *J. Mol. Biol.*, 239:213–226, 1994.

[54] D. Stoddart, A. J. Heron, E. Mikhailova, G. Maglia, and H. Bayley. Single-nucleotide discrimination in immobilized dna oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, 106(19):7702–7707, 2009.

[55] T. C. Sutherland, M. J. Dinsmore, H. B. Kraatz, and J. S. Lee. An analysis of mismatched duplex dna unzipping through a bacterial nanopore. *Biochem Cell Biol*, 82(3):407–12, 2004.

[56] A. Tsortos, G. Papadakis, K. Mitsakakis, K.A. Melzak, and E. Gizeli. Quantitative determination of size and shape of surface-bound dna using an acoustic wave sensor. *Biophysical Journal*, 94(7):2706–2715, 2008.

[57] M. Wanunu. Nanopores: A journey towards dna sequencing. *Physics of Life Reviews*, 9:125–158, 2012.

[58] J.D. Watson and F.H.C. Crick. Molecular structre of nucleic acids. *Nature*, 171:737–738, 1953.

[59] N. A. Wilson, R. Abu-Shumays, B. Gyarfas, H. Wang, K. R. Lieberman, M. Akeson, and W. B. Dunbar. Electronic control of dna polymerase binding and unbinding to single dna molecules. *Acs Nano*, 3(4):995–1003, 2009.

[60] L. X. Zhang and J. Chen. Elastic behavior of compact polymer chain transporting through an infinite adsorption channel. *Polymer*, 47(5):1732–1740, 2006.

Appendix A

SUPPLEMENTAL INFORMATION FOR DNA IMMOBILIZATION EXPERIMENTS

This work was published on October 4, 2011 in PLOS ONE [37].

A.1 Supplemental Figures and Tables

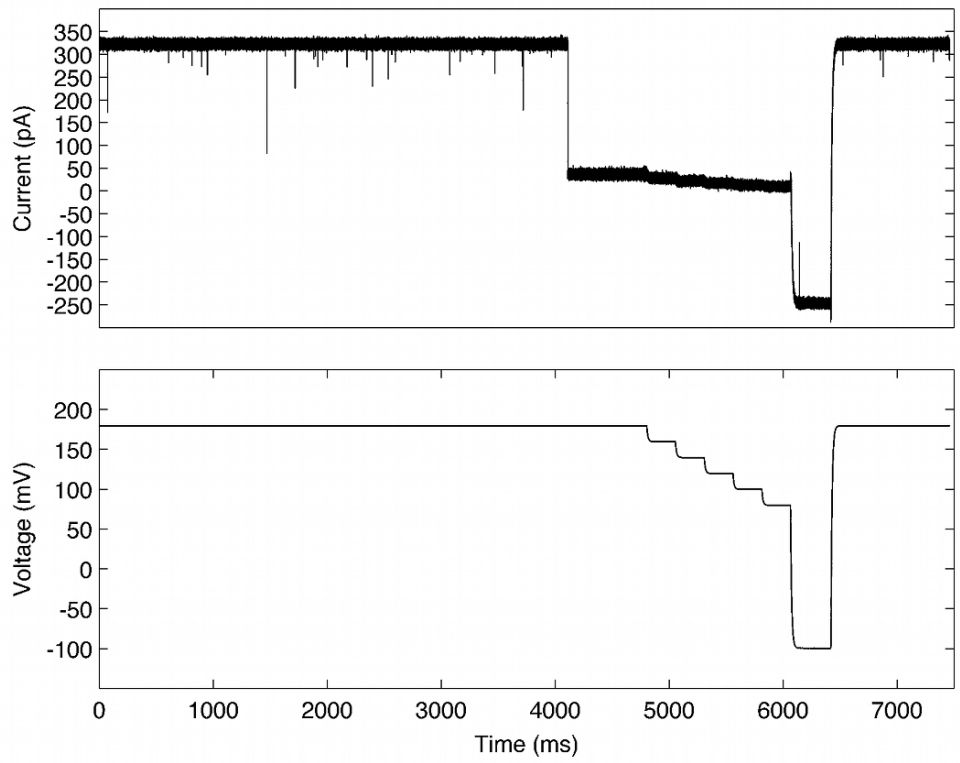


Figure A.1: **Example Voltage and Current Trace for DNA Immobilization Experiments.** A current (upper) and voltage (lower) trace is shown for an example event. The applied voltage was held at 180 mV until there was a spontaneous reduction in current corresponding to DNA entering the pore. The voltage was held at 180, 160, 140, 120, 100 and 80 mV for 250 ms per voltage level. Then, −100 mV was applied to force the DNA out of the pore and back into the *cis* compartment. The transient current spikes seen when the voltage is held at 180 mV result from DNA entering the pore vestibule but escaping before fully threading through the constriction. Residual current levels reported in Chapter 4 correspond to an applied voltage of 180 mV and the analysis in Chapter 6 uses all voltage levels.

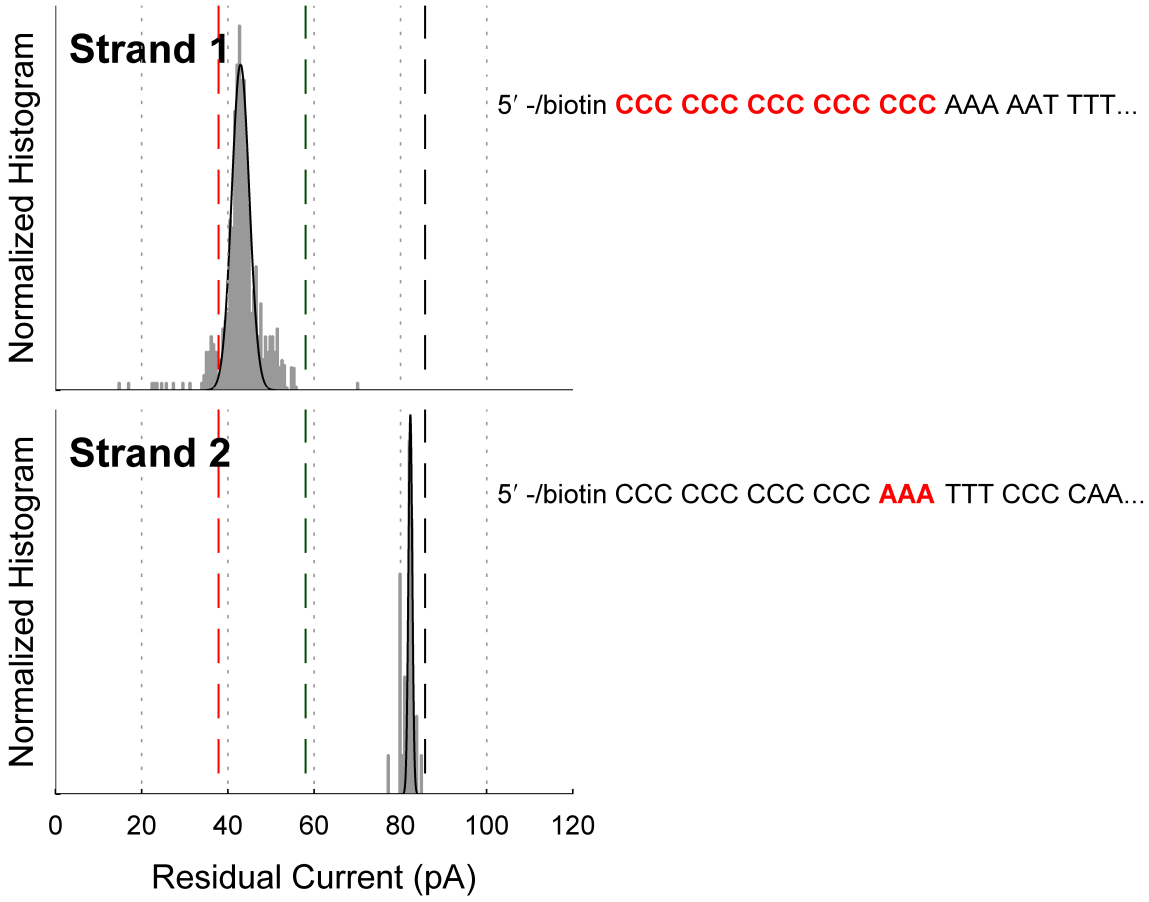


Figure A.2: **Experimental Results to Determine the Region of Sensitivity of DNA Immobilized in MspA.** To determine the region of sensitivity of M1-MspA, the residual ionic current was measured when the two strands of ssDNA, shown right, were held within the pore with a NeutrAvidin 'anchor'. The mean residual ionic current (gray) and fitted Gaussian curve (black) are shown for each strand. For reference, the Gaussian means of the mean residual ionic current are shown for poly-dC (red dashed), poly-dT (green dashed), and poly-dA (black dashed). The mean residual ionic current for 'Strand 1' is most like that of poly-dC, suggesting a recognition site within the first 15 bases (highlighted red in sequence). The mean residual ionic current for 'Strand 2' is most like that of poly-dA, corresponding to a region of sensitivity near the 13th-15th nucleotide (highlighted red in sequence).

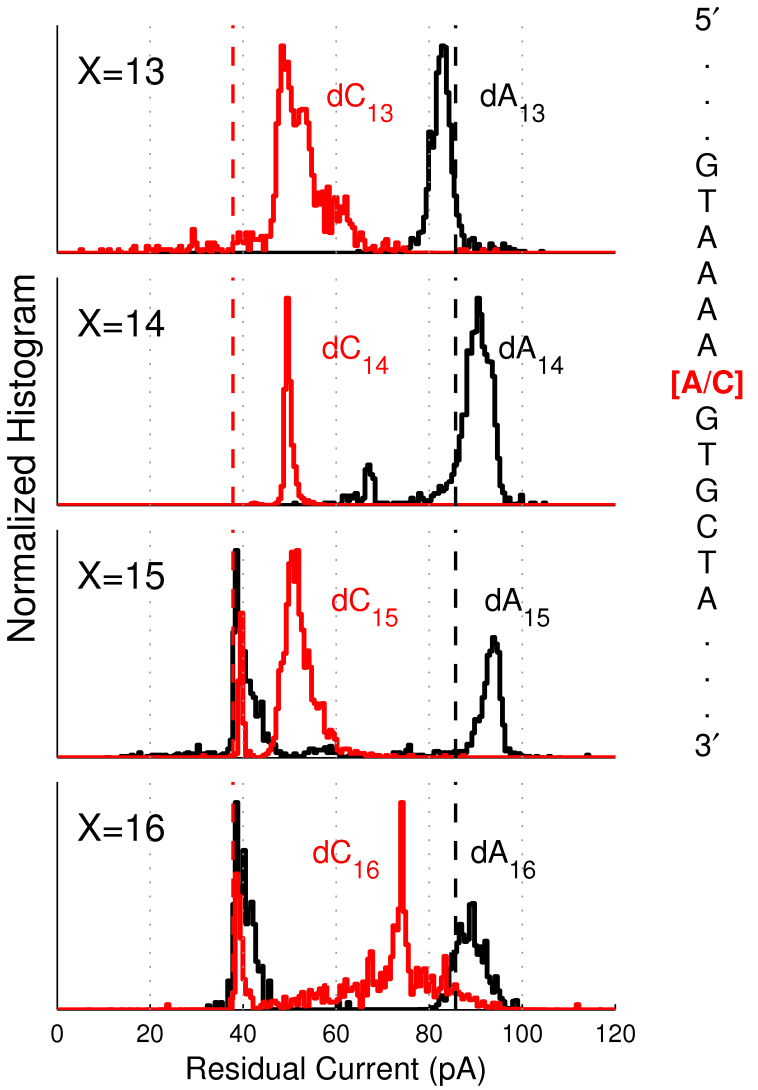


Figure A.3: **Experimental Results for Immobilized Strands Containing a Single Nucleotide Polymorphism (rs1447295).** A segment of ssDNA containing SNP rs1447295, which is associated with an increased risk of prostate cancer [1, 19], was bound to NeutrAvidin such that the polymorphism is at the 13-16th nucleotide, X, from the NeutrAvidin. Part of the surrounding sequence is shown with the nucleotide of interest, either an adenine or cytosine, highlighted in red (see Table A.1 for complete sequences). Histograms of the mean residual ionic current for both the dA_X and dC_X variants for X = 13-16 and Gaussian means of the mean residual ionic current for poly-dA (black dashed) and poly-dC (red dashed) are shown. The two variants are most resolved when X = 13 and 14. The ionic current level for dA_X is similar to that of poly-dA and the current for dC_X is similar to that found for a single dC in poly-dA. When X = 15 or 16 we observe two current levels for each variation; one level is near 40 pA for both dA_X and dC_X and the other level is unique to the SNP variation.

Name	Sequence	Vapp = 180 mV		Num. Events	Num. Pores
		Ires (pA)	HWHH (pA)		
poly-dC	5′– /biotin/ CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	37.76	0.806	2065	8
poly-dA	5′– /biotin/ AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAC CTG TCT CCC TGC CG	85.92	2.044	1637	6
poly-dT	5′– /biotin/ TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTC CTG TCT CCC TGC CG	58.01	1.211	689	3
poly$_5$-dG	5′– /biotin/ AAA AAA AAA AAA AGG GAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	89.21	5.109	1287	4
poly$_4$-mC	5′– /biotin/ CCC CCC CCC CCC (mC)(mC)(mC) (mC)CC CCC CCC CCC CCC TGT CTC CCT GCC GCT GTC TCC CTG CCT GTC TCC CTG CCG	38.86	0.284	878	3
poly-dC (5′ lead)	5′– CTG TCT CCC TGC CGC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CC /biotin/	83.19	1.43	1989	3
poly-dA (5′ lead)	5′– CTG TCT CCC TGC CGA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA /biotin/	91.40	1.81	568	3
poly-dT (5′ lead)	5′– CTG TCT CCC TGC CGT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT /biotin/	49.80	0.98	1279	3
poly$_5$-dG (5′ lead)	5′– CTG TCT CCC TGC CGA AAA AAA AAA AAA AAA AAA AGG GAA AAA AAA AAA AA /biotin/	100.70	3.54	325	3
strand1	5′– /biotin/ CCC CCC CCC CCC CCC AAA AAT TTT TCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	42.69	2.40	495	2
strand2	5′– /biotin/ CCC CCC CCC CCC AAA TTT CCC CAA AAA CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	82.09	0.59	33	3
CinA11	5′– /biotin/ AAA AAA AAA ACA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	87.02	1.10	401	3
CinA12	5′– /biotin/ AAA AAA AAA AAC AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	85.37	2.83	748	3
CinA13	5′– /biotin/ AAA AAA AAA AAA CAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	56.37	5.69	330	4
	second peak	80.45	9.54	371	4
CinA14	5′– /biotin/ AAA AAA AAA AAA ACA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	51.44	0.45	71	3
CinA15	5′– /biotin/ AAA AAA AAA AAA AAC AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	53.09	2.01	1922	5
CinA16	5′– /biotin/ AAA AAA AAA AAA AAA CAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	54.73	4.25	831	4
	second peak	76.07	3.64	589	4
CinA17	5′– /biotin/ AAA AAA AAA AAA AAA ACA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	86.47	2.16	75	3
CinA18	5′– /biotin/ AAA AAA AAA AAA AAA AAC AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	86.47	1.42	216	4
AinC11	5′– /biotin/ CCC CCC CCC CAC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	38.31	0.57	156	3
AinC12	5′– /biotin/ CCC CCC CCC CCA CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	38.31	0.71	68	3
AinC13	5′– /biotin/ CCC CCC CCC CCC ACC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	36.12	1.56	325	4
AinC14	5′– /biotin/ CCC CCC CCC CCC CAC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	35.02	0.75	584	4
AinC15	5′– /biotin/ CCC CCC CCC CCC CCA CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	40.50	1.19	498	6
AinC16	5′– /biotin/ CCC CCC CCC CCC CCC ACC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	41.59	0.79	289	4
AinC17	5′– /biotin/ CCC CCC CCC CCC CCC CAC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	38.86	1.64	616	3
AinC18	5′– /biotin/ CCC CCC CCC CCC CCC CCA CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	39.40	0.82	182	3
AinT13	5′– /biotin/ TTT TTT TTT TTT ATT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT CTG TCT CCC TGC CG	70.60	5.55	671	3
AinT14	5′– /biotin/ TTT TTT TTT TTT TAT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT CTG TCT CCC TGC CG	73.88	0.95	555	4
AinT15	5′– /biotin/ TTT TTT TTT TTT TTA TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT CTG TCT CCC TGC CG	73.88	6.83	2231	3
AinT16	5′– /biotin/ TTT TTT TTT TTT TTT ATT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT CTG TCT CCC TGC CG	61.84	3.29	860	3
SNP1-C13	5′– /biotin/ GCT GGA GAA AGG CAT GTG CAA ATT AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	51.99	4.46	353	4
	second peak	81.00	13.40	1321	4
SNP1-A13	5′– /biotin/ GCT GGA GAA AGG AAT GTG CAA ATT AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	96.87	2.50	454	3
SNP1-C14	5′– /biotin/ TGC TGG AGA AAG GCA TGT GCA AAT TAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	48.71	1.06	573	3
SNP1-A14	5′– /biotin/ TGC TGG AGA AAG GAA TGT GCA AAT TAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	89.21	5.19	599	3
SNP1-C15	5′– /biotin/ CTG CTG GAG AAA GGC ATG TGC AAA TTA AGA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	52.54	4.63	500	3
SNP1-A15	5′– /biotin/ CTG CTG GAG AAA GGA ATG TGC AAA TTA AGA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	82.09	1.99	312	3
SNP1-C16	5′– /biotin/ CCT GCT GGA GAA AGG CAT GTG CAA ATT AAG AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	81.00	13.81	1371	3
SNP1-A16	5′– /biotin/ CCT GCT GGA GAA AGG AAT GTG CAA ATT AAG AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	87.02	5.09	844	3
SNP2-C13	5′– /biotin/ AGG TAT GTA AAA CGT GCT ATG GAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	50.90	4.38	1027	3
SNP2-A13	5′– /biotin/ AGG TAT GTA AAA AGT GCT ATG GAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	82.64	2.33	1065	3
SNP2-C14	5′– /biotin/ GAG GTA TGT AAA ACG TGC TAT GGA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	49.25	0.75	803	4
SNP2-A14	5′– /biotin/ GAG GTA TGT AAA AAG TGC TAT GGA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	90.30	3.25	1152	3
SNP2-C15	5′– /biotin/ GGA GGT ATG TAA AAC GTG CTA TGG AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	39.40	0.47	125	5
	second peak	50.90	3.02	1124	5
SNP2-A15	5′– /biotin/ GGA GGT ATG TAA AAA GTG CTA TGG AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	39.40	2.28	1112	3
	second peak	93.04	2.09	874	3
SNP2-C16	5′– /biotin/ GGG AGG TAT GTA AAA CGT GCT ATG GAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	38.86	0.75	80	3
	second peak	73.34	7.36	425	3
SNP2-A16	5′– /biotin/ GGG AGG TAT GTA AAA AGT GCT ATG GAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	39.95	2.37	185	3
	second peak	88.66	3.80	178	3

Table A.1: **DNA Sequences used in DNA Immobilization Experiments.** Results for experiments conducted at an applied voltage of 180 mV are shown.

Appendix B

SUPPLEMENTAL INFORMATION FOR ENZYME CONTROLLED TRANSLOCATION EXPERIMENTS

This work was published in the April 2012 issue of Nature Biotechnology [36].

B.1 Level Detection and Comparison Methods

Events were filtered with a Savitzky-Golay filter of order 2. To identify levels during DNA synthesis events, we used a custom implemented step detection method with a gradient threshold to detect transitions between unique levels. Adjacent levels were combined if (i) the difference between the levels' mean currents was within 1-sigma of the levels' standard deviation, (ii) the levels were shorter than 800 μs, or (iii) the difference between the two current level means multiplied by the duration of the subsequent level (a charge-volume) was <20 fC. Toggles were detected as repeated transitions between two levels and were removed to facilitate comparison to other events. For the figures herein, the current traces were filtered further using a 5-point window and downsampled to 1.7 kHz.

To automatically correlate the current levels from different events, we used a Needleman-Wunsch sequence alignment algorithm with an affine-gap penalty[16]. Scoring matrix elements for the alignment were generated from the negative square of the normalized current differences multiplied by 50. Affine-gap constants (e and d[16]) were set at -0.25 and -0.21, respectively.

Small voltage offsets, due to electrode potential offsets and differences in ionic conductivity (due to evaporation), reduced comparability between events recorded on

different experimental setups. To compensate for this, we allowed for a multiplicative scale factor, ranging between 0.96 and 1.04 and an additive offset ranging from 0.03 and -0.03. These correction factors allowed for proper comparison between levels in two events.

B.2 Supplemental Figures and Tables

Figure B.1: **Experimental Results for Control Experiments of Enzyme Controlled DNA Translocation.** Events from experiments taken with 'Block Homopolymer' DNA, see Table B.1 for sequence. (a) Example events without phi29 DNAP present. Two classes of events were observed, (1) short events with a blockage current (I_b) of about 0.70 of the open pore current (I_o) that correspond to DNA entering the pore vestibule but never passing through the constriction and (2) those with I_b/I_o <0.50 that correspond to DNA passing through MspA after the blocking oligomer and primer sequentially unzip from the DNA template strand. For both classes, the translocation time is too fast to observe nucleotide specific current levels. (b) Example event with phi29 DNAP but without divalent cations and dNTPs. The phi29 DNAP binds to DNA in the *cis* volume and the bound complex is then captured in MspA (Figure 5.2). The blocking oligomer unzips from the template with the phi29 DNAP's force-activated strand displacement function. As the DNA moves through the pore, a single high peak (designated with an 'X') is observed when the two abasic residues in the template strand pass through MspA's constriction. Since phi29 DNAP is unable to synthesize in the absence of divalent cations and dNTPs, it eventually either falls off allowing the DNA primer to cooperatively unzip from the template, or continues unzipping the primer strand with phi29 DNAP's force-activated strand displacement function (as is shown in the given trace). In either case, the template strand passes through the pore to the *trans* side and the current returns to I_o. (c) Expanded view the region around the abasic peak.

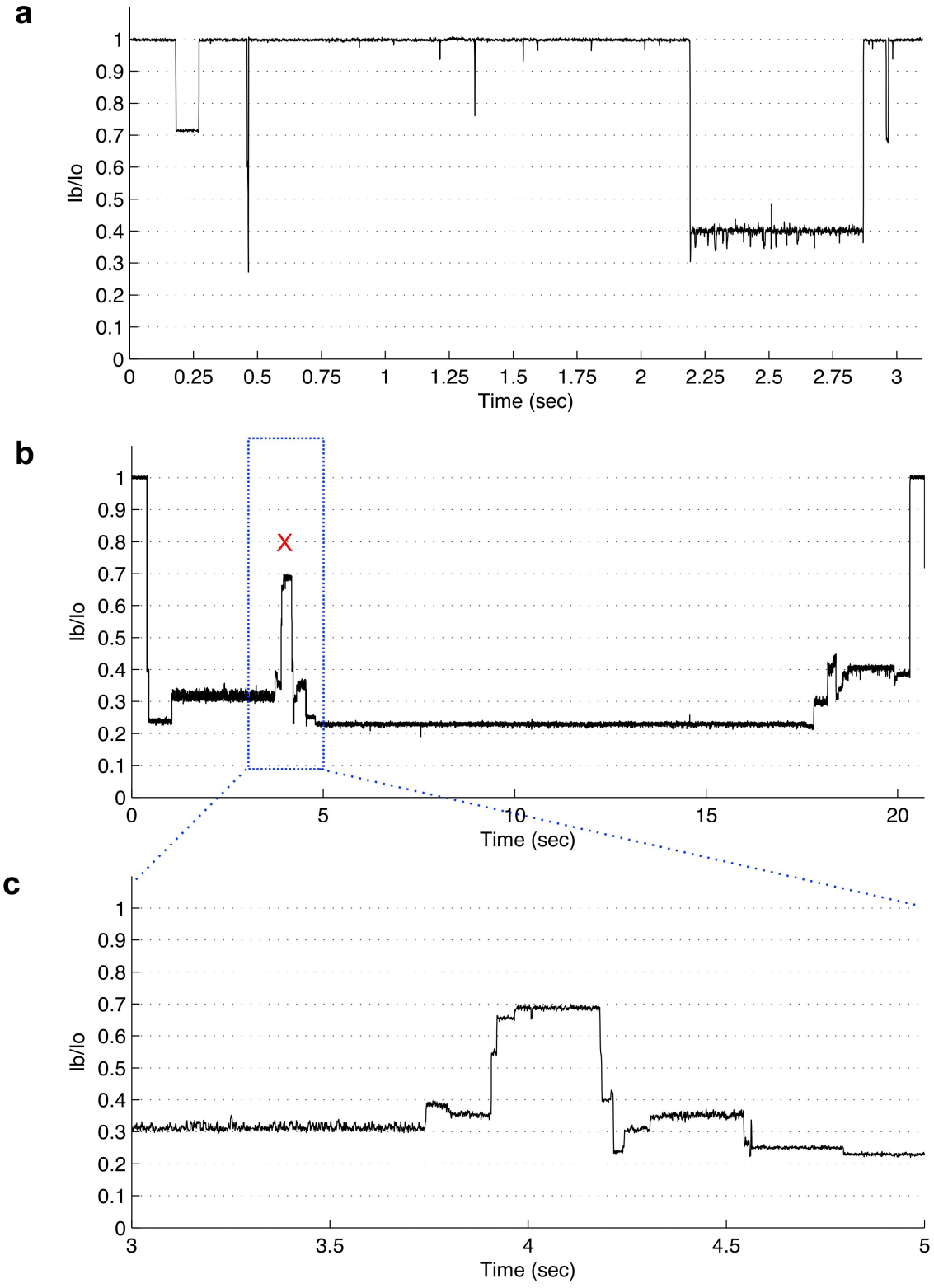
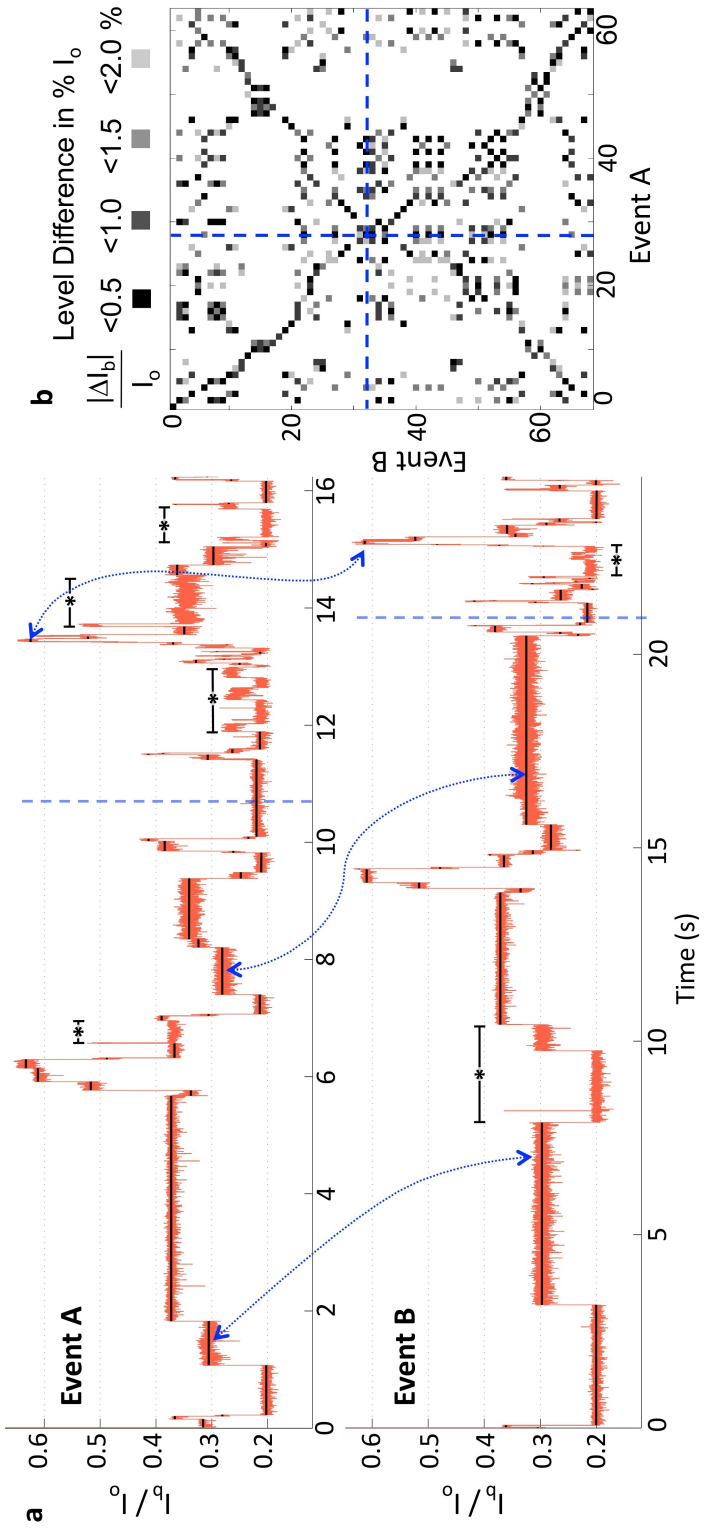

a
Ib/Io
Time (sec)
b
X
Ib/Io
Time (sec)
c
Ib/Io
Time (sec)

Figure B.2: **Similarity of 'Block Homopolymer' DNA Current Pattern for Enzyme Controlled DNA Translocation.**(a) Current traces of two example events with the 'Block Homopolymer' DNA. This DNA template first moves through MspA towards the *trans* volume as the blocking oligomer unzips, and then moves back towards the *cis* volume as phi29 DNAP extends the primer strand. The reversal of DNA motion is indicated by the vertical blue dashed line. Solid black lines indicate the average current of levels found by automated analysis (see Appendix B.1). Regions where the phi29 DNAP fails to incorporate a base result in toggles of the current and are indicated by regions with *. Blue arrows assist in identifying a few corresponding current levels in the two events. The two current traces, Event A and Event B, were recorded in two separate experiments with different pores and several days apart. Levels were found with a computer algorithm (see Appendix B.1). (b) Using a 'dot plot' [21], we compared every level found in Event A with every level found in Event B. Two levels with average current differing by less than 0.5% is indicated with a black square ('dot') and for those differing by less than 1% a grey dot was entered. For example, the dot found at (6,4) represents the similarity between the 6th observed level in Event A and the 4th observed level in Event B. The shading of this point reveals that the two levels differ by less than 1% of I_o. The diagonal line of dots that emerges (upper left to lower right) indicates that the levels of each event follow the same time-ordered sequence. A second perpendicular line of dots, demonstrates that levels occurring during blocking oligomer unzipping have the opposite time-ordering of current levels observed during synthesis. Similar results were found with each DNA sequence examined (Figures B.7, B.9, B.12 and B.15).


a
Event A
Event B
I_b / I_o
Time (s)
b
Level Difference in % I_o
$\frac{|\Delta I_b|}{I_o}$ <0.5 <1.0 <1.5 <2.0 %
Event A
Event B

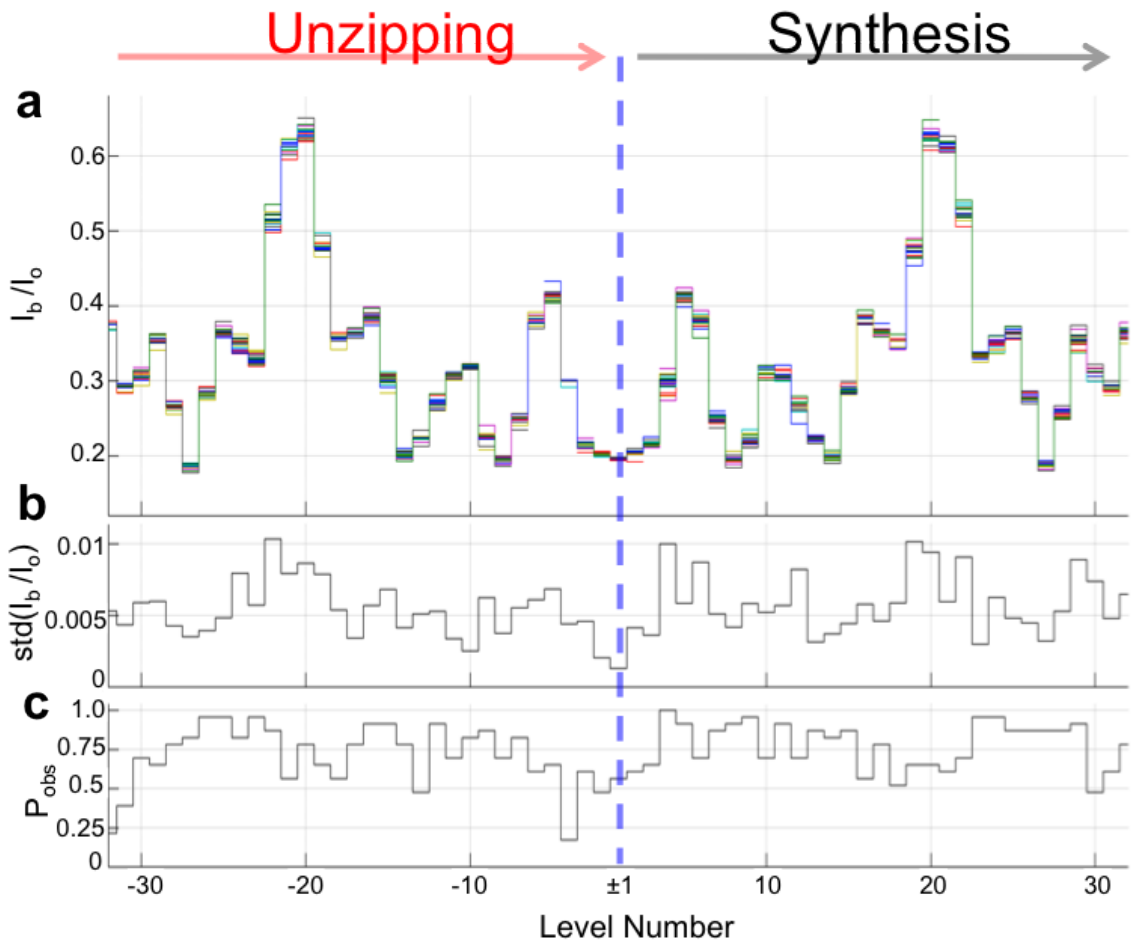


Figure B.3: **Reproducibility of 'Block Homopolymer' DNA Current Pattern for Enzyme Controlled DNA Translocation.**(a) Average current levels extracted from events with 'Block Homopolymer' DNA. Levels for n=24 events collected on N=2 pores were found using a level detection algorithm (see Appendix B.1). A consensus level sequence of currents was constructed to serve as a reference. The levels of each event were then automatically aligned to this consensus level sequence to extract average currents associated with each level (see Appendix B.1). The average current levels of each event are overlaid (one color for each event). The symmetry of the plot across the blue dashed line indicates that the DNA template changes direction allowing nucleotides to pass through the pore twice, first during unzipping of the blocking oligomer and again during synthesis. Each of these passes produces a similar current level. (b) The standard deviation of the spread in each of the overlaid levels in (a). The average 1-sigma fluctuation of all levels is 0.6% of I_o. (c) The probability, P_{obs}, of finding a given level in the proper order. Levels with P_{obs}=1, such as level 4, are found in all events. Some levels such as level -18 and level 18 are not efficiently detected due to their similarity in current to neighboring levels. Levels near the DNA template turn around point (blue dashed line) are often too fast to be detected, resulting in low P_{obs}.

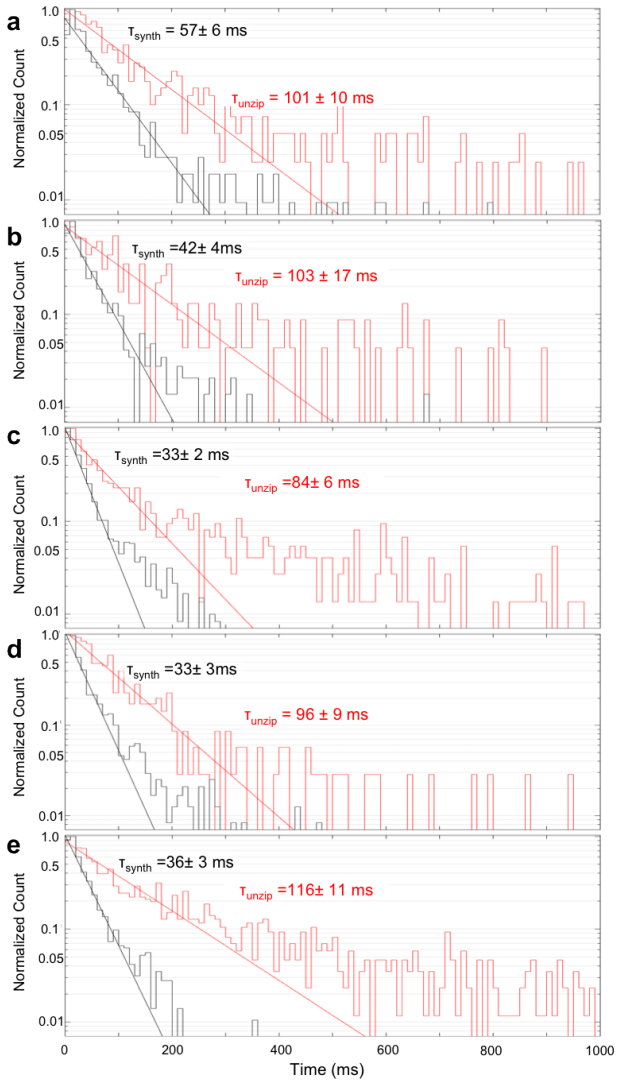
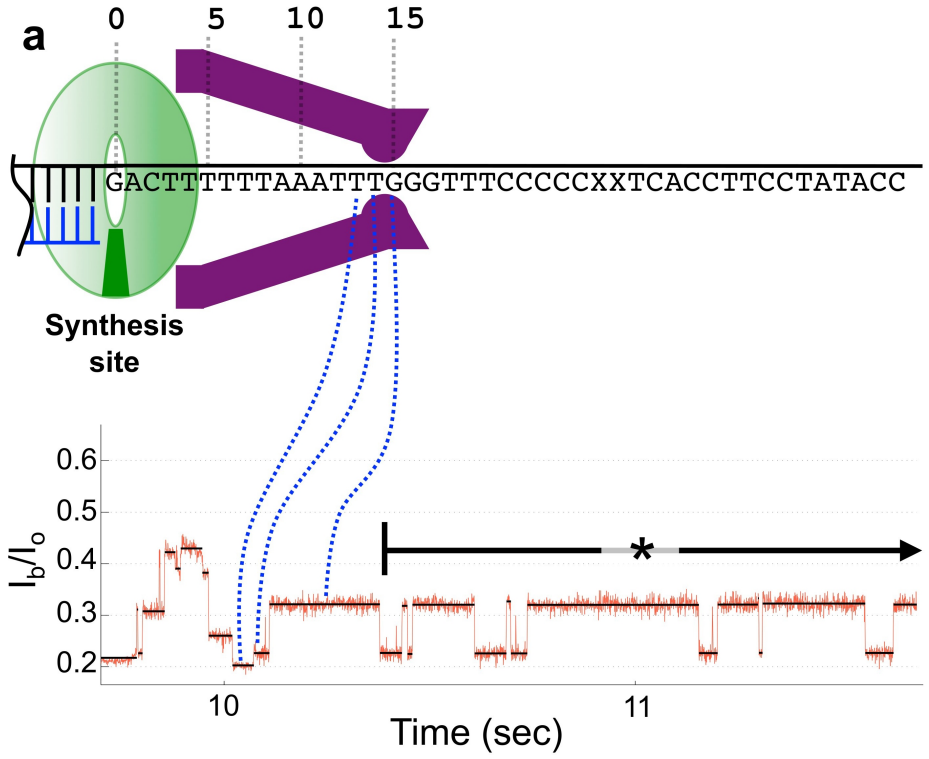


Figure B.4: **Exponential Behavior of Level Durations for Enzyme Controlled DNA Translocation.** Normalized histograms of level durations for levels found during the unzipping phase (red) and during the synthesis phase (black) for (a) 'Block Homopolymer' DNA, (b) 'Heteromer DNA 1', (c) 'Heteromer DNA 2', (d) 'Heteromer DNA 3', and (e) 'Heteromer DNA 4'. Both unzipping (red) and synthesis (black) distributions are well described by an exponential distribution suggesting that stochastic processes govern both unzipping and synthesis (see figures for exponential time constants). The level duration distribution for synthesis has a tail due to a few levels of very long duration. We note that these time constant values differ slightly, which is likely due to differences in the sequences. Several levels during the unzipping are much longer than 100 ms so that they dominate the length the unzipping process, but do not significantly skew the duration time constants.

Figure B.5: **Location of Synthesis Site Determined from Enzyme Controlled DNA Translocation Experiments.** Using 'Block Homopolymer' DNA, we determine the nucleotide-distance between the phi29 DNAP synthesis site and the constriction of MspA. We performed experiments with either no dCTP or with 100-fold reduced concentration ($\approx 1\mu$M) of dCTP but with normal concentrations (100 μM) of dATP, dGTP and dTTP. When no dCTP was present, blocking oligomer could be unzipped from the DNA template but the phi29 DNAP could not synthesize nucleotides beyond a dG within the template strand (data not shown). With the 100-fold reduced [dCTP] the phi29 DNAP was eventually able to synthesize beyond a dG of the template strand but took longer when compared to experiments run with normal dCTP concentration. This figure shows two schematic representations of the location of the DNA as the phi29 DNAP unsuccessfully attempted to synthesize a dG along the template strand. (a) After the blocking oligomer unzipped from the template (first 9.5 seconds not shown) the polymerase extended 9 nucleotides and then back-stepped to the 8th nt. The current continued to toggle between these levels for 30 seconds, as indicated by region with the *, before synthesis continued past the dG. (b) The phi29 DNAP synthesized past the first dG of the template strand allowing the abasic residues in the template strand to pass through MspA. When another dG entered the phi29 DNAP synthesis site the toggling was observed for the remainder of the event. We expect that these toggles are due to failed attempts to insert a dCTP into the growing primer stand. Comparing the ion current levels to those in Figure 5.3, the traces in (a) and (b) each individually indicate there are 14-15 nt between MspA's constriction and the phi29 DNAP synthesis site.


a
0
5
10
15
GACTTTTTTAAATTTGGGTTTCCCCCXXTCACCTTCCTATACC
Synthesis site
0.6
0.5
0.4
0.3
0.2
I_b/I_o
*
10
11
Time (sec)

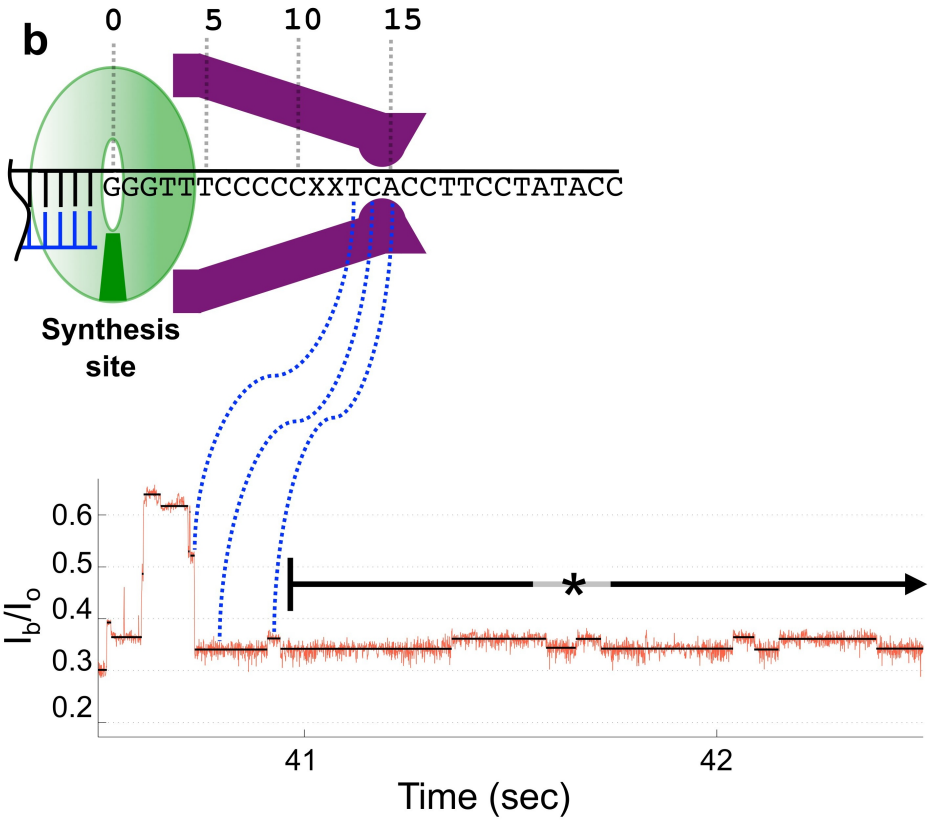
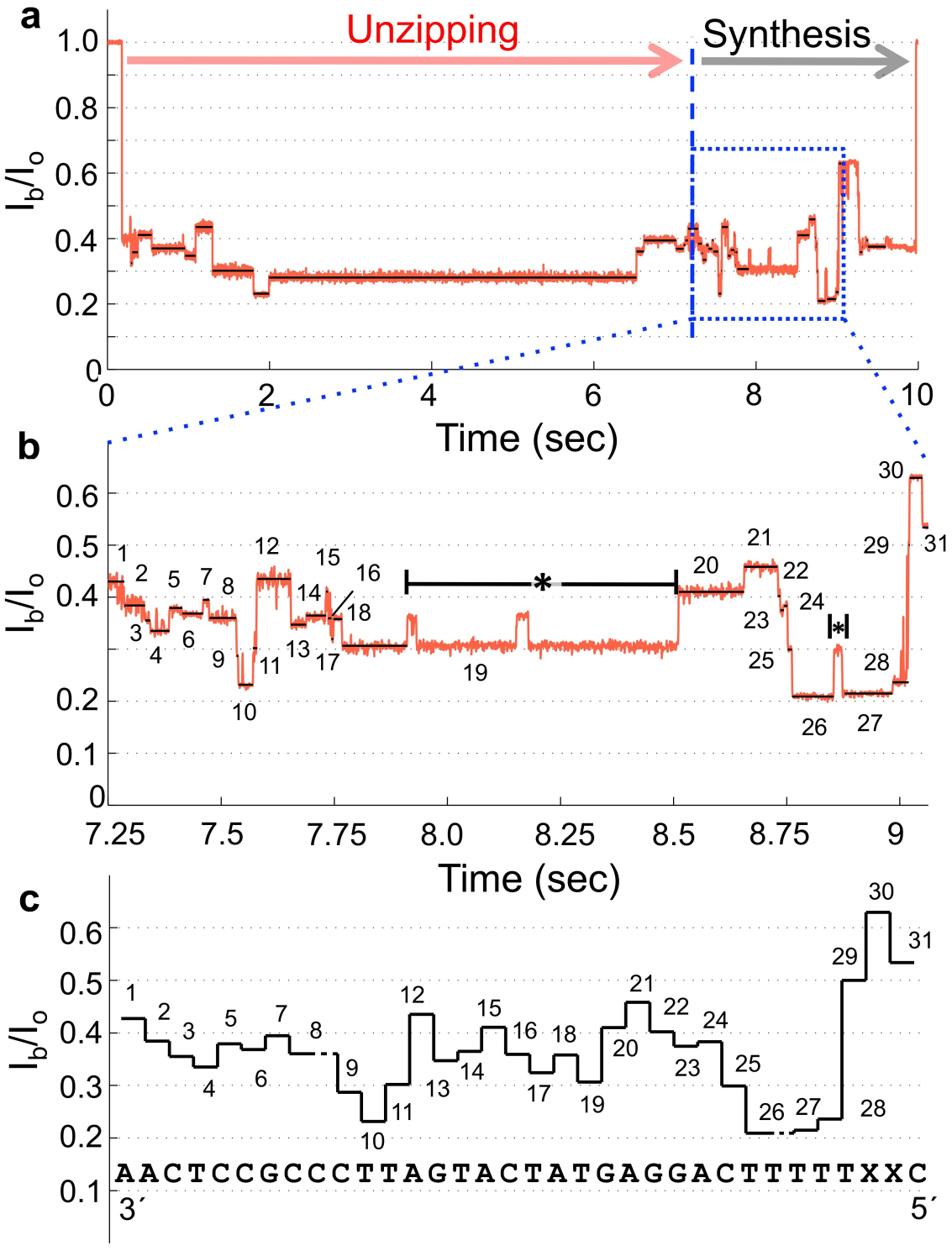

b
0
5
10
15
GGGTTTCCCCCXXTCACCTTCCTATACC
Synthesis site
0.6
0.5
0.4
0.3
0.2
I_b/I_o
*
41
42
Time (sec)

Figure B.6: **Example Event for 'Heteromer DNA 1'.** Example current trace taken with 'Heteromer DNA 1' annealed to a 23 nt primer and a 15 nt 'blocking oligomer'. (a) The current spontaneously drops from the open pore level (I_o) to <0.5 I_o when DNA enters the pore. During the unzipping of the blocking oligomer there are occasional long pauses. When the blocking oligomer is completely removed, as indicated by the vertical blue dashed line, the phi29 DNAP begins synthesis and the DNA motion changes direction. The region of synthesis (blue box) is expanded in (b). Solid black lines indicate current levels observed in many events. The levels observed during blocking oligomer removal contain a reverse time-ordered subset of the levels during the synthesis phase. (b) Expansion of the event during phi29 DNAP synthesis shows individual current levels, numbered for convenience. Toggles in current levels are marked with an *. (c) Currents from extracted levels in (b) are plotted with the matching sequence of the DNA template ($3'$ to $5'$). The nucleotides aligning to the right and left of each level most influence the current for that level. It is suspected that levels 8 and 26 each contain two levels of indistinguishable current value. This is indicated with dashes inserted to account for levels that are not observed but are expected. Two abasic residues at the end of the sequence cause a high current (level 30) and are used to indicate the end of a successful read.


a
Unzipping
Synthesis
I_b/I_o
Time (sec)
b
I_b/I_o
Time (sec)
c
I_b/I_o
AACTCCGCCCTTAGTACTATGAGGACTTTTTXXC
3′
5′

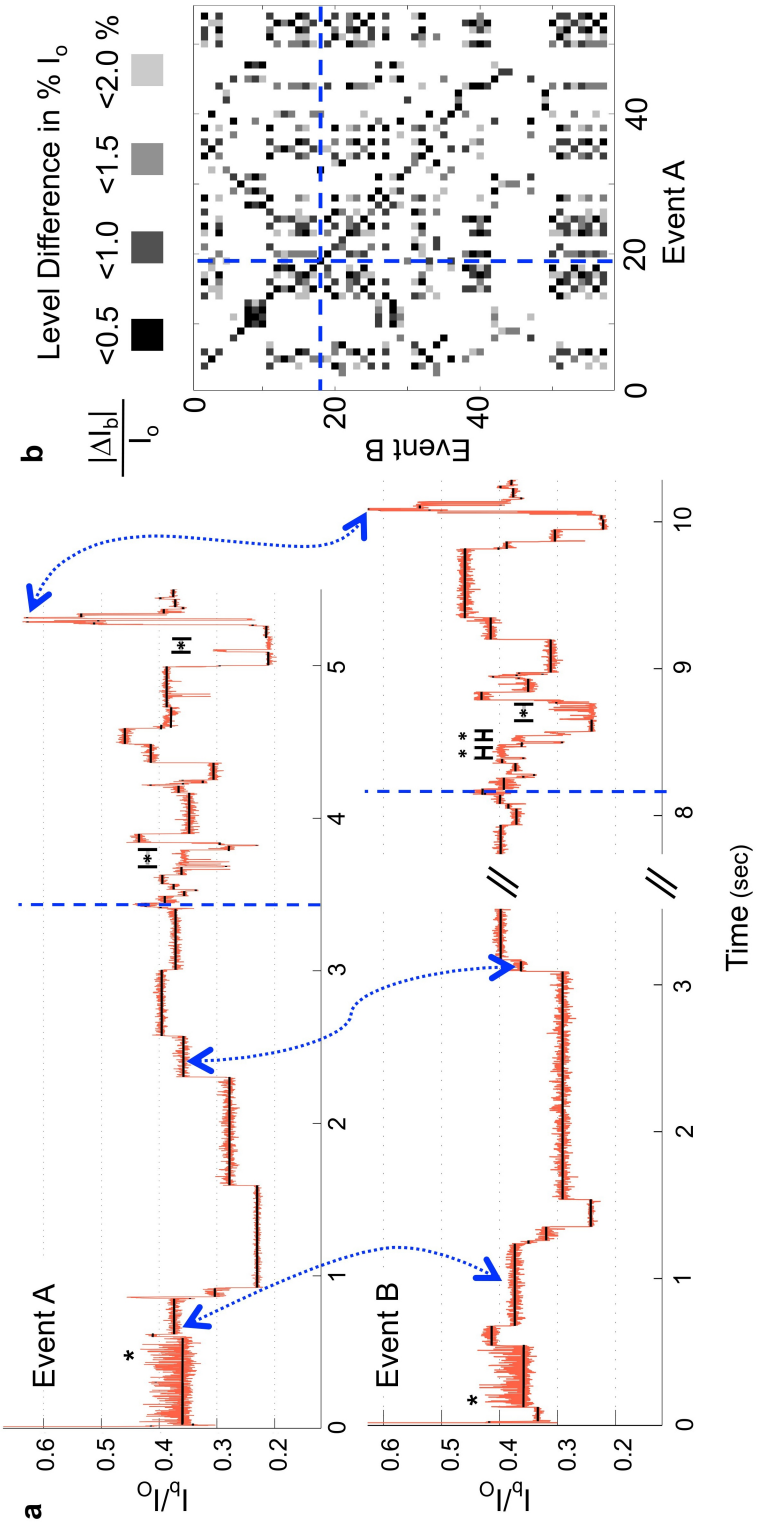


Figure B.7: **Similarity of 'Heteromer DNA 1' Current Pattern for Enzyme Controlled DNA Translocation.** Same as Figure B.2 but with the 'Heteromer DNA 1' See Table B.1 for sequences.

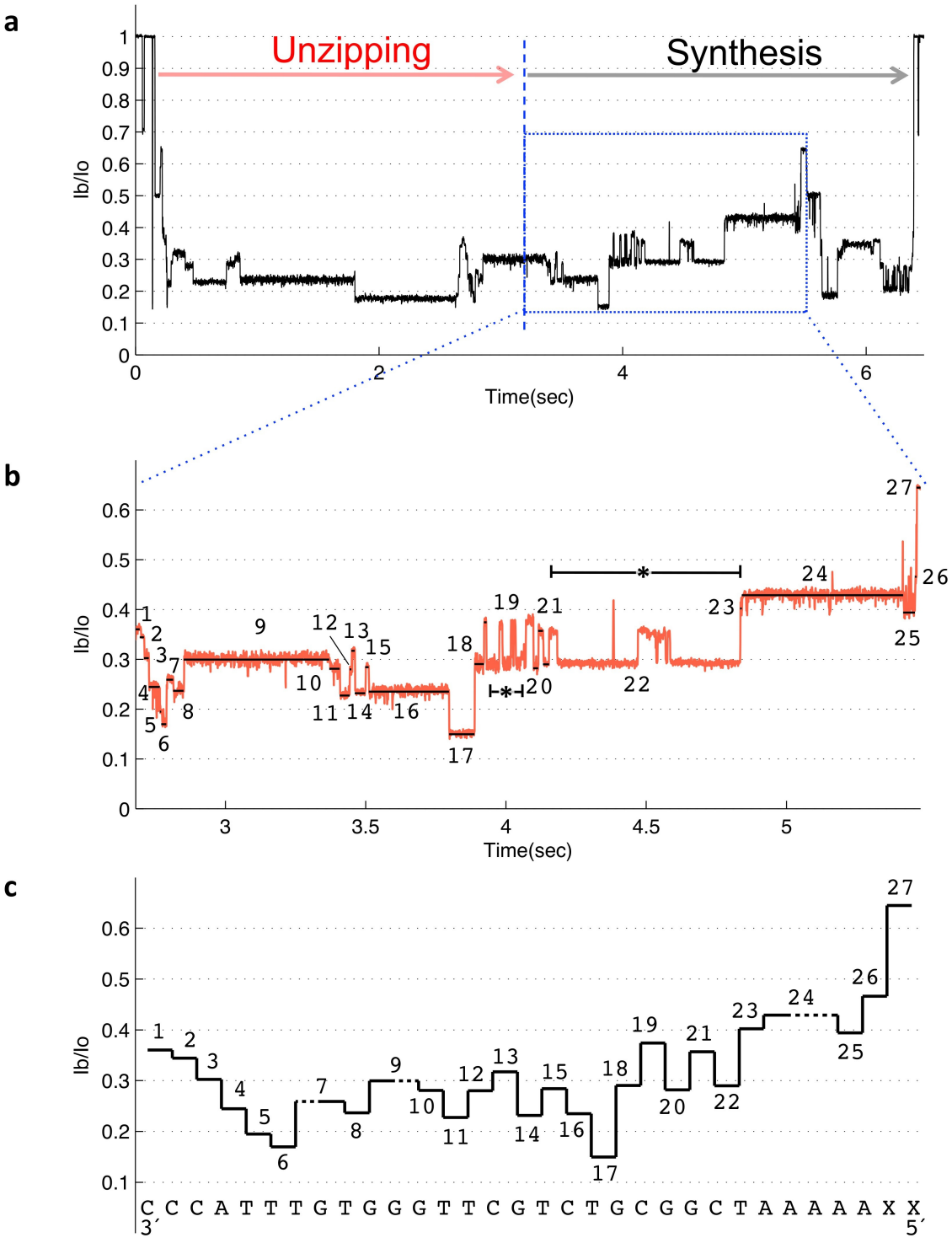


Figure B.8: **Example Event for Enzyme Controlled DNA Translocation of 'Heteromer DNA 2'.** Same as Figure B.6 but with the 'Heteromer DNA 1'. See Table B.1 for sequences.

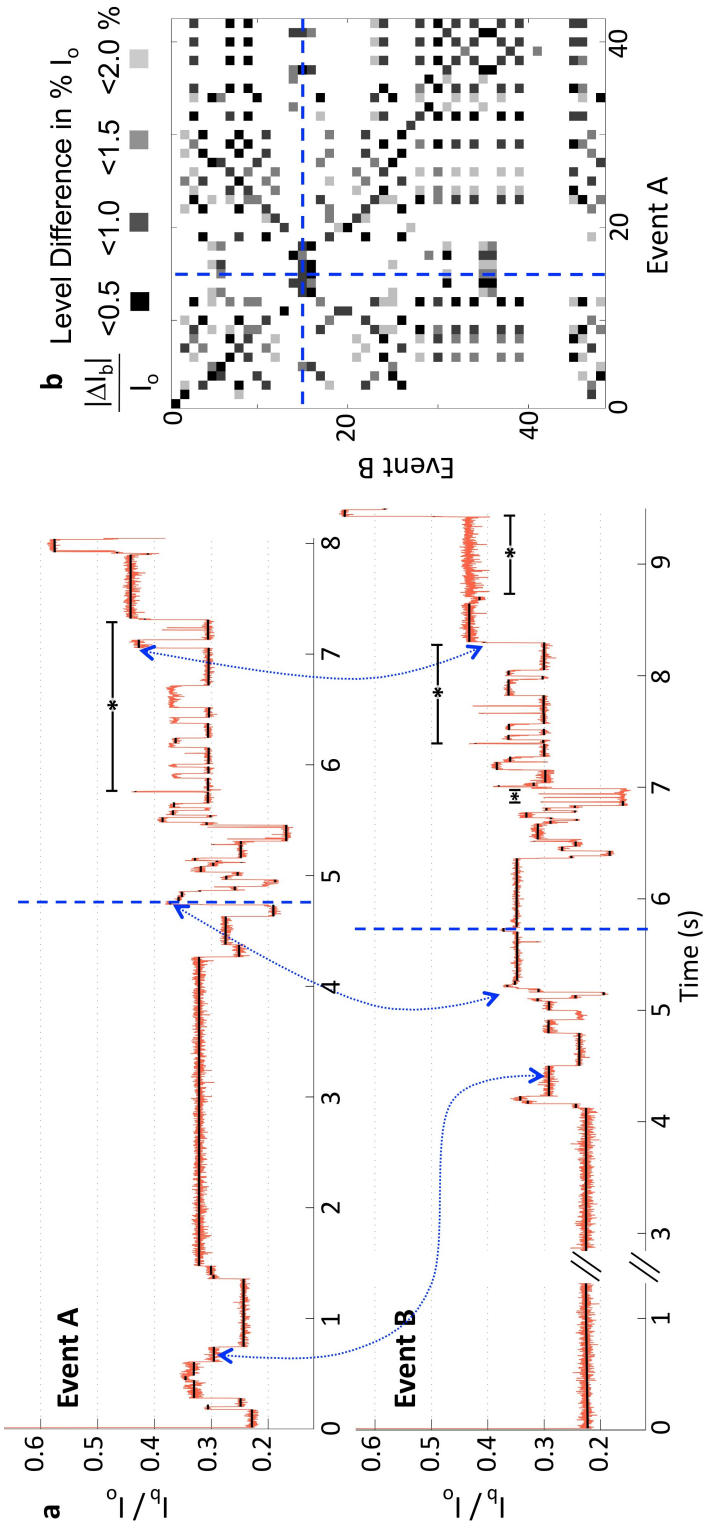


Figure B.9: **Similarity of 'Heteromer DNA 2' Current Pattern for Enzyme Controlled DNA Translocation.** Same as Figure B.2 but with the 'Heteromer DNA 2' See Table B.1 for sequences.

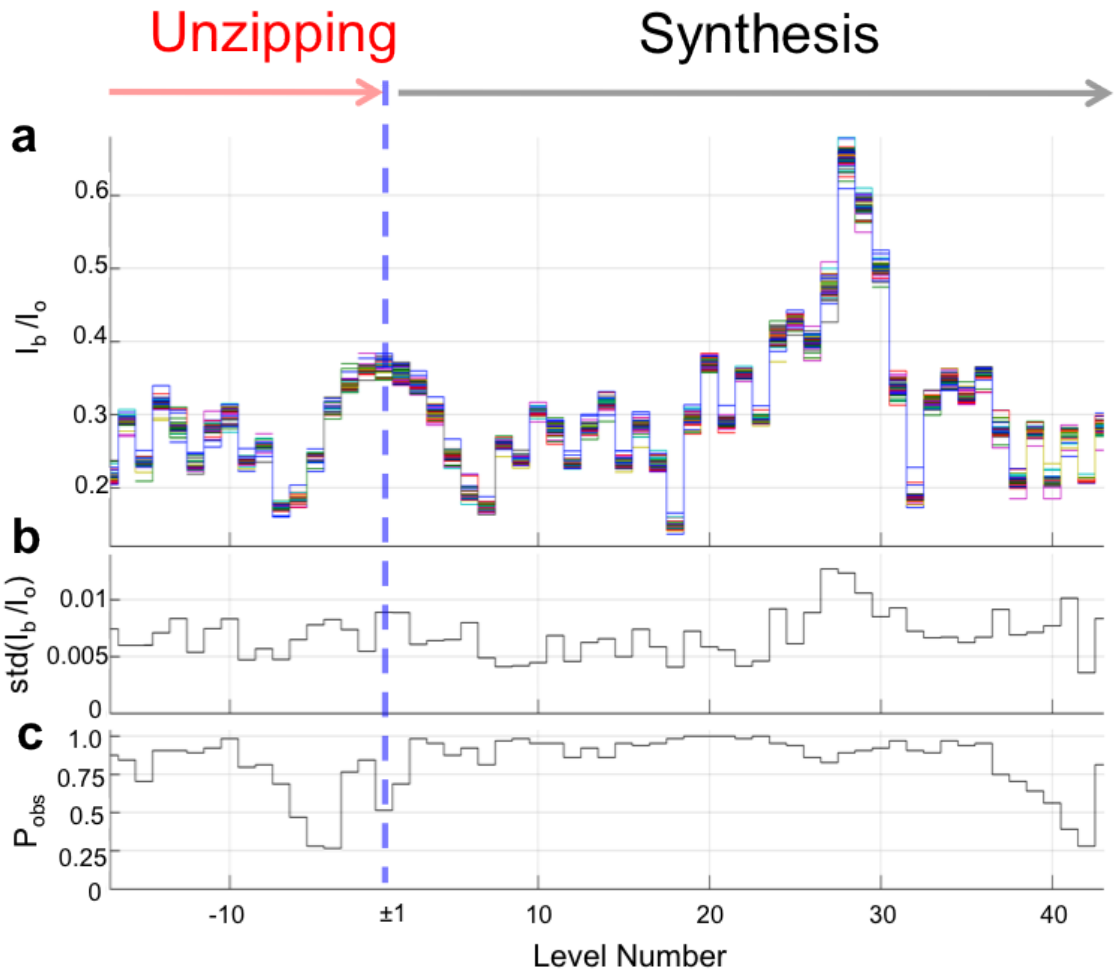


Figure B.10: **Reproducibility of 'Heteromer DNA 2' Current Pattern for Enzyme Controlled DNA Translocation.** Same as Figure B.3 but with the 'Heteromer DNA 2'. Time ordered levels for n=64 events collected on N=2 pores.

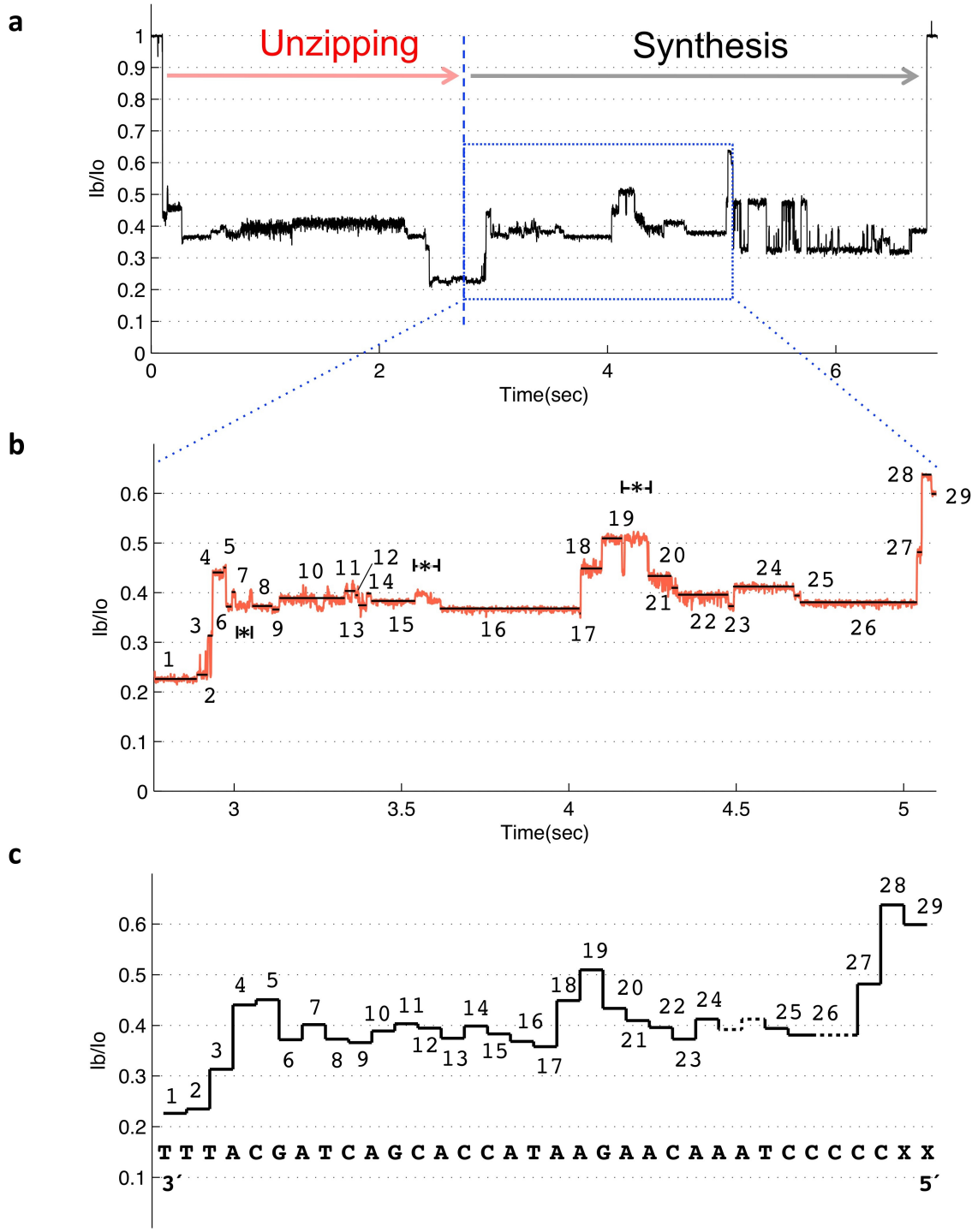


Figure B.11: **Example Event for Enzyme Controlled DNA Translocation of 'Heteromer DNA 3' Sequence.** Same as Figure B.6 but with the 'Heteromer DNA 3' template. The dashed lines between levels 24 and 25 are two levels observed in other events but are absent in this example trace. See Table B.1 for sequences.

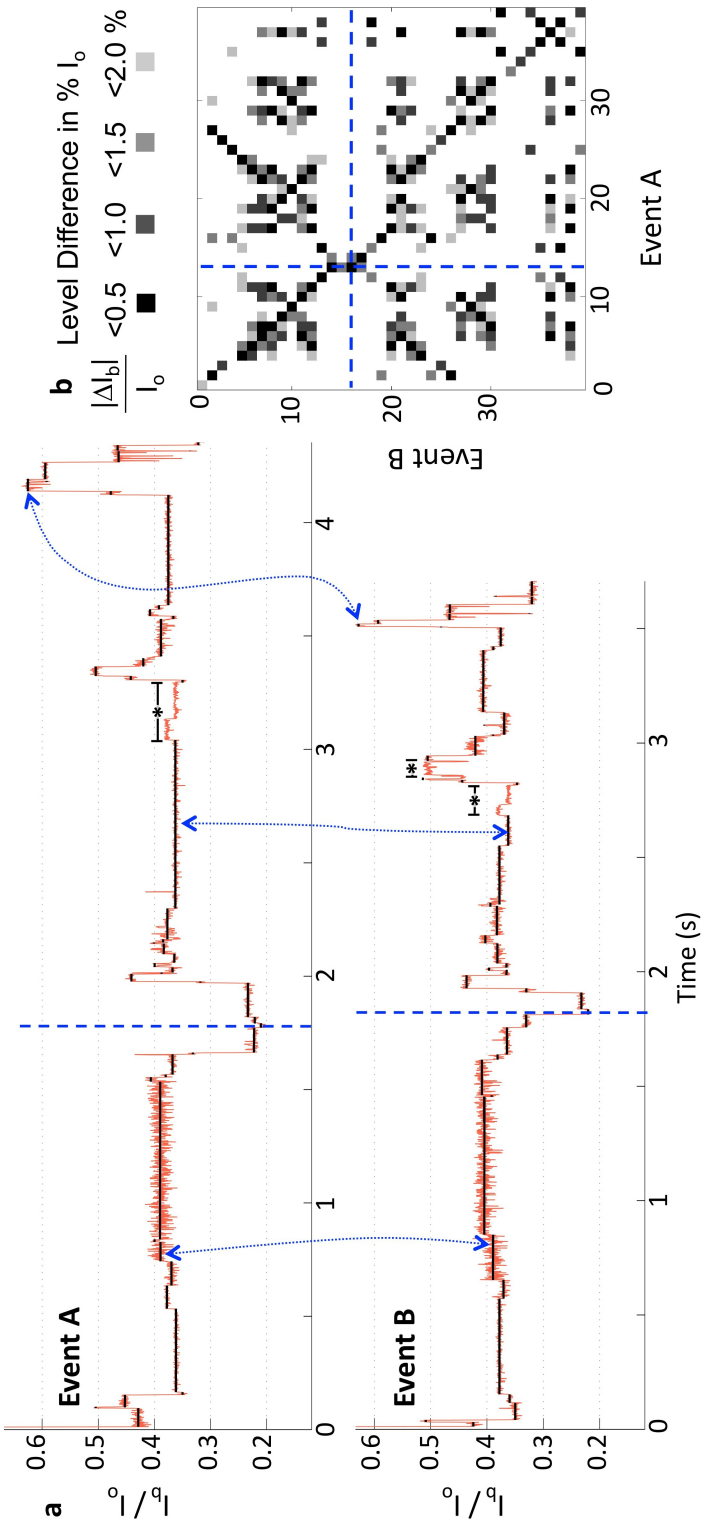


Figure B.12: **Similarity of 'Heteromer DNA 3' Current Pattern for Enzyme Controlled DNA Translocation.** Same as Figure B.2 but with the 'Heteromer DNA 3' See Table B.1 for sequences.

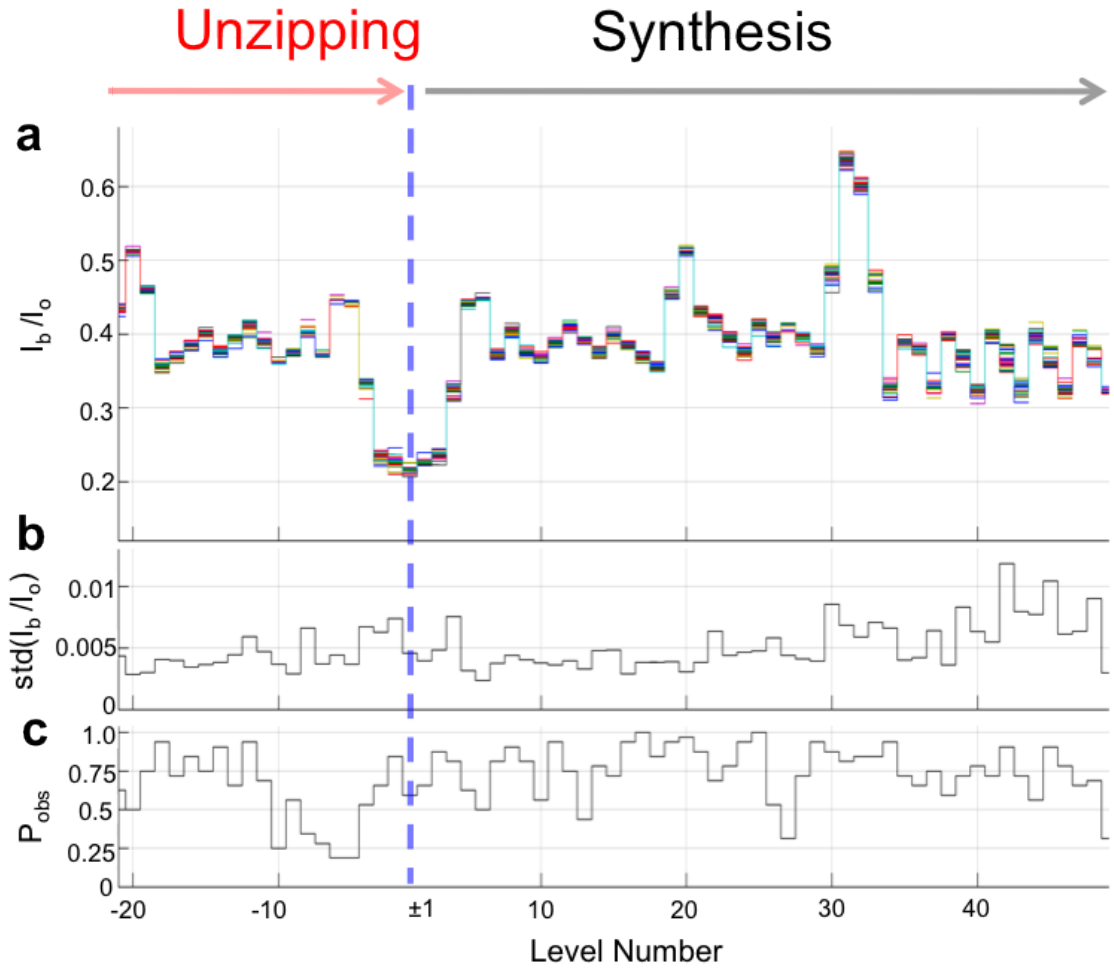


Figure B.13: **Reproducibility of 'Heteromer DNA 3' Current Pattern for Enzyme Controlled DNA Translocation.** Same as Figure B.3 but with the 'Heteromer DNA 3'. Time ordered levels for n=32 events collected on N=1 pore. Note the series of levels after the abasic peak during the synthesis phase. This series of levels reproduces the levels observed in CAT DNA (Figure 5.4), demonstrating that despite differing DNA sequence within MspA's lumen, level progressions are reproducible.

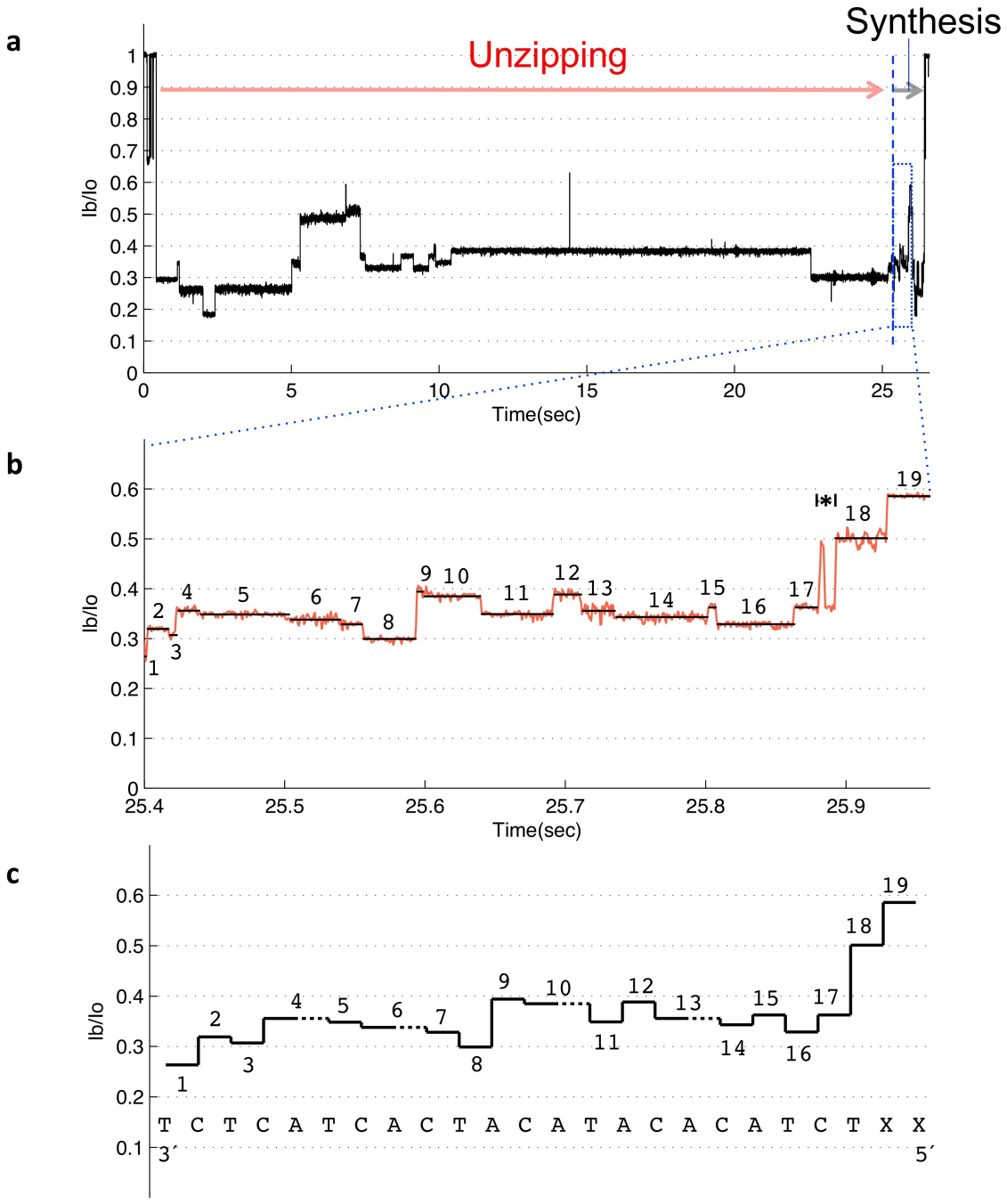


Figure B.14: **Example Event for Enzyme Controlled DNA Translocation of 'Heteromer DNA 4' Sequence.** Same as Figure B.6 but with the 'Heteromer DNA 4' template. See Table B.1 for sequences.

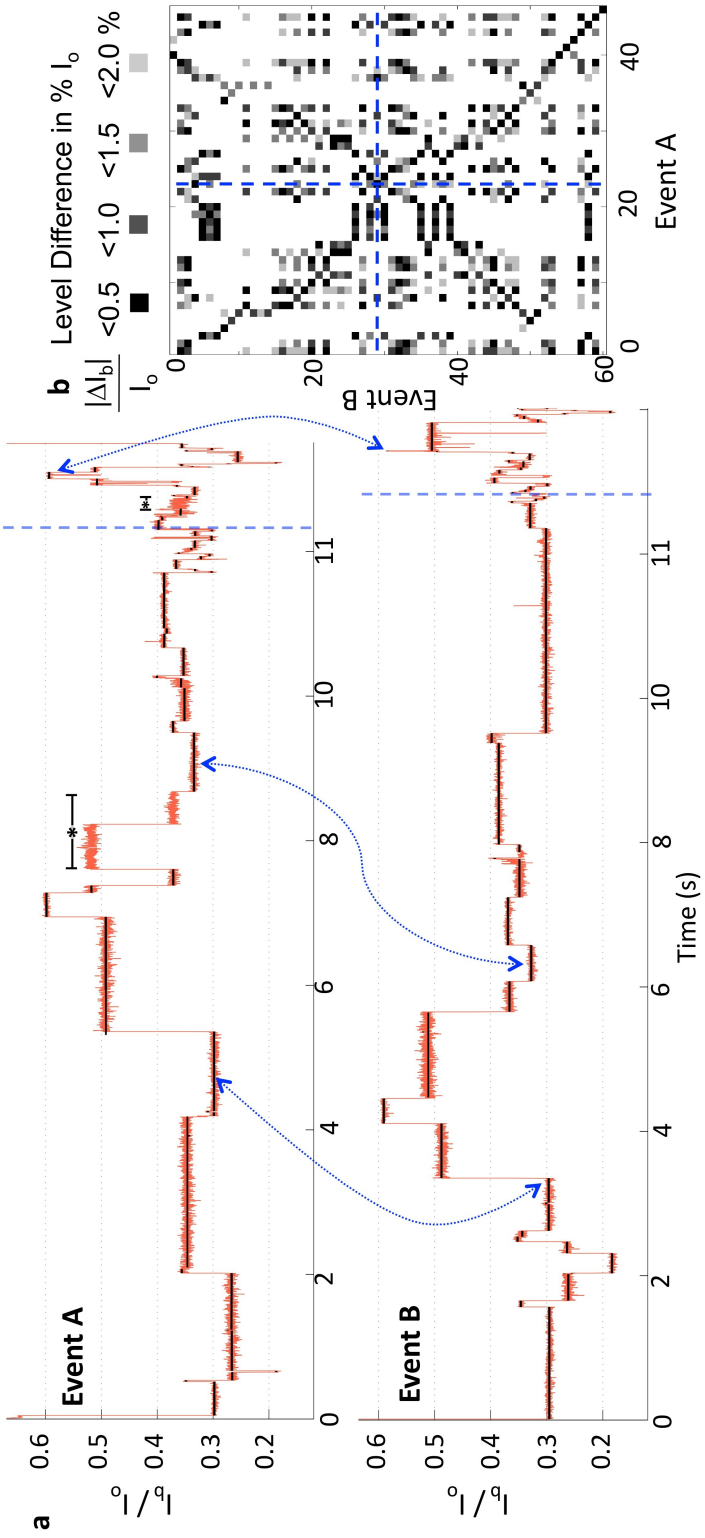


Figure B.15: **Similarity of 'Heteromer DNA 4' Current Pattern for Enzyme Controlled DNA Translocation.** Same as Figure B.2 but with the 'Heteromer DNA 4' See Table B.1 for sequences.

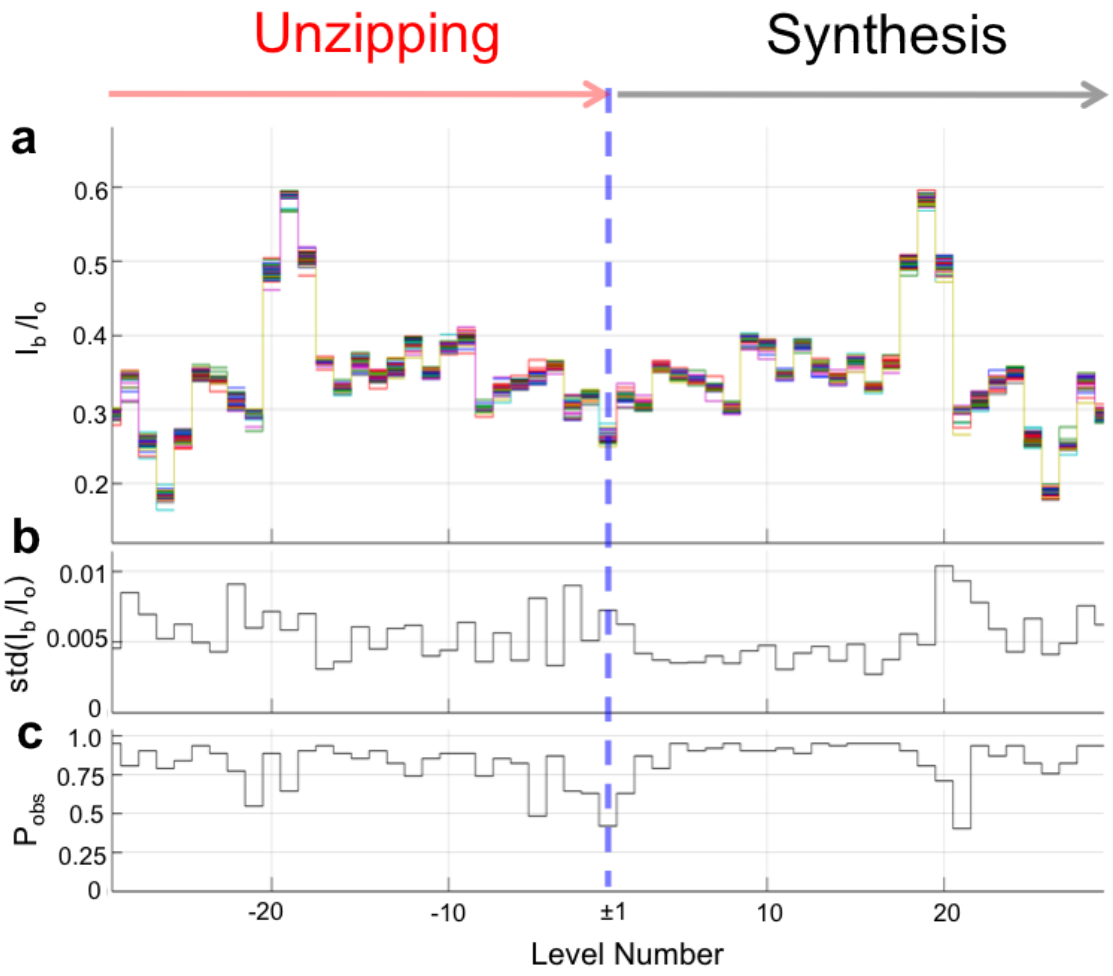


Figure B.16: **Reproducibility of 'Heteromer DNA 4' Current Pattern for Enzyme Controlled DNA Translocation.** Same as Figure B.3 but with the 'Heteromer DNA 4'. Time ordered levels for n=62 events collected on N=3 pores.

Stand	Figure Referenced	Sequence	Length (nt)	Pores (N)	Events (n)
Block Homopolymer	Template Strand 3'→5'	[CCG ATG CTG GAC GTA CTC TTA CG]C TAT CAC TCT AGA CTT TTTT AAA TTT GGG TTT CCCCC XX TCA CCT TCC TAT CCA CTC	80	5	33
	Blocking Oligomer 5'→3'	G ATA GTG AGA TCT GAA AAAA TTT AAA CCC XXXXXXXZ	29		
CAT DNA	Template Strand 3'→5'	[CCG ATG CTG GAC GTA CTC TTA CG]ATT ATT ATT ACT ACT AC TAC TAC TAC GAC TAC TAC TAC TAC TAC TAC	82	5	475
	Blocking Oligomer 5'→3'	TAA TAA TAA TGA TGA XXXXXXXZ	15		
Heteromer DNA 1	Template Strand 3'→5'	[CCG ATG CTG GAC GTA CTC TTA CG]C TAT CAC TCT AGA CTA AAA CTC CGC CCT TAG TAC TAT GAG GAC TTTTT XX CAT CAT CAT CAT CAT	91	4	47
	Blocking Oligomer 5'→3'	G ATA GTG AGA TCT GA XXXXXXXZ	15		
Heteromer DNA 2	Template Strand 3'→5'	[CCG ATG CTG GAC GTA CTC TTA CG]C TAT CAC TCT AGA CTC CCC ATT TGT GGG TTC GTC TGC GGC TAA AAA XX ATG CAC TCA TGT ATC TAT	89	3	121
	Blocking Oligomer 5'→3'	G ATA GTG AGA TCT GA XXXXXXXZ	15		
Heteromer DNA 3	Template Strand 3'→5'	[CCG ATG CTG GAC GTA CTC TTA CG]C TAT CAC TCT AGA CTT TTT ACG ATC AGC ACC ATA AGA ACA AAT CCCCC XX TAC TAC TAC TAC TAC	91	3	94
	Blocking Oligomer 5'→3'	G ATA GTG AGA TCT GA XXXXXXXZ	15		
Heteromer DNA 4	Template Strand 3'→5'	[CCG ATG CTG GAC GTA CTC TTA CG]C TAT CAC TCT AGA CTT CTC ATC ACT ACA TAC ACA TCT XX TCA CCT TCC TAT CCA CTC	80	5	142
	Blocking Oligomer 5'→3'	G ATA GTG AGA TCT GAA GAG TAG TGA TGT A XXXXXXXZ	29		
Primer for all strands	Primer 5'→3'	GCG TAC GCC TAC GG TTTT CC GTA GGC GTA CGC [GGC TAC GAC CTG CAT GAG AAT GC]	55		

Notes:
1. Primer/Template duplex region is 23 nt long and is indicated by brakets [ ]
2. Blocking Oligo / Template duplex region is indicated by underline
3. The same primer sequence is used for all DNA Hybrid Strands.
4. An abasic residue is indicated by an 'X' and a 3 Carbon Spacer is indicated by a 'Z'
5. The primer contains a hairpin structure on it's 5' end indicated by underline

Table B.1: **DNA Sequences used in Enzyme Controlled DNA Translocation Experiments.**

Appendix C

SUPPLEMENTAL INFORMATION FOR FORCE SPECTROSCOPY EXPERIMENTS

C.1 Force Calculation Details (Homopolymer Adenine Strand)

From experiment, we found the change in DNA-pore resistances due to single nucleotide substitutions at various positions along the strand. The DNA-pore resistance and error for a given strand was determined from the mean and standard deviation, respectively, of the resistances frond from multiple experiments. By fitting a Gaussian curve to the results at each applied voltage, we determined which nucleotides comprised the 'recognition site' of the DNA strand from the peak and full width at half max (FWHM) of the fit (Figures 6.3 and 6.5). The width of the recognition site was comprised of two elements; MspA's constriction size and Brownian motion of nucleotides. The constriction was the narrowest part of the MspA pore. We assumed that the applied voltage dropped entirely over the constriction so that only nucleotides within this region experienced a force. However, Brownian motion caused the nucleotides to move in and out of the pore constriction, making the total recognition site of the DNA larger than the physical size of MspA's constriction. For calculations of the force on the DNA strand, we avoided the effects of Brownian motion moving nucleotides into the constriction by excluding the region of DNA within the recognition site.

To determine the force on a DNA strand immobilized in MspA, we modeled the strand between the biotin linker and recognition site as a freely jointed chain (FJC). The DNA strand within the vestibule, then, had one end fixed by the biotin linker. The force from nucleotides in the induced electric field acted on the free end of the

strand and corresponded to the top of MspA's recognition site. See Figure C.1.

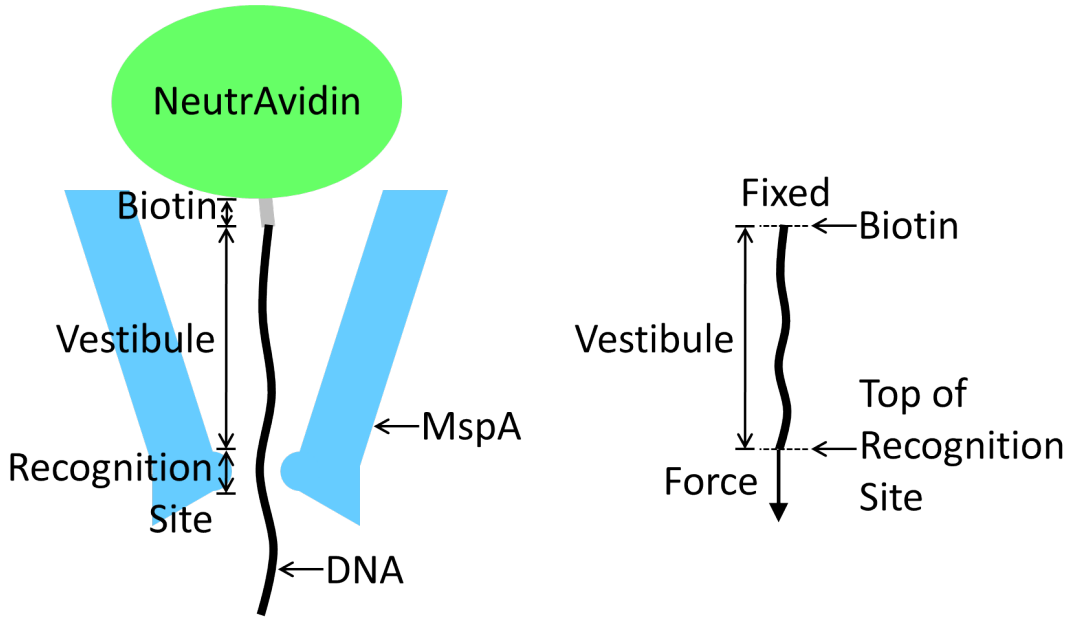


Figure C.1: **Schematic Diagram of the Freely Jointed Chain Model Applied to DNA Immobilized in MspA.** The DNA strand is immobilized within MspA by a NeutrAvidin molecule attached by a biotin linker (left). The recognition site is comprised of the nucleotides surrounding MspA's constriction which appreciably affect the DNA-pore resistance. We estimate that the induced electric field is non-negligible only within MspA's constriction and the force on the DNA stand within the vestibule was due to the nucleotides within the electric field. For our calculations, we modeled the region of the DNA in the vestibule, between the biotin linker and top of the recognition site, as a freely jointed chain (FJC) with a single applied force (right).

The force on DNA within the electric field, E, was $F = qE$ where q is the total effective charge of DNA within the electric field. The charge of DNA, q, was $q = \lambda s$ where λ is the effective linear charge density and s is the contour length of the strand. The electric field, E, can be estimated to be $E = \frac{V}{a}$, where V is the electrostatic potential drop and a is the distance over which the potential changes. The force on the FJC was then:

$$F = \frac{\lambda s V}{a} \tag{C.1}$$

To compare forces at different voltages, we estimated that the FWHM value from experiments, n, was proportional to the contour length of DNA within the electric field, s. Taking the linear charge density, λ, and width of the electric field, a, as voltage independent, the forces at two voltages were related by:

$$\frac{F_1}{F_2} = \frac{n_1}{n_2} * \frac{V_1}{V_2} \tag{C.2}$$

The extension of a FJC under a force, F, was estimated as [52]:

$$x = L * [coth(\frac{Fb}{k_b T}) - (\frac{k_b T}{Fb})](1 + \frac{F}{S}) \tag{C.3}$$

where x is the end-to-end distance of the strand, L is the contour length, k_b is the Boltzmann constant, T is the temperature, S is the stretch modulus, and b is the Kuhn length. At room temperature, the product $k_b T$ is 4.11 pN-nm. The Kuhn length, b, was estimated as 3.0 nm [43] (1M KCl) and the stretch modulus, S, was taken as 800 pN [52]. We took the contour length, L, from experiment as the length of DNA between the biotin and top of the recognition site, assuming an interphosphate distance of 0.56 nm. The extension or end-to-end distance, x, was the length of the vestibule between the biotin and top of the recognition site and was taken as unknown and voltage-independent. The applied force and end-to-end distance of the vestibule, x, were found with the FJC model by comparing the force-extension curves (Equation C.3) at two voltages using the force relationship found above (Equation C.2). Since 180 mV was our baseline voltage, we compared the extension at each lower voltage to that at 180 mV.

The force extension curves for a single dC substitution in poly-dA are shown in Figure C.2. The average force for an applied voltage of 180 mV found from the FJC comparisons described above was 15.54 $\pm$ 1.81 pN. Forces for each lower voltage

were determined using this value and the force relationship from Equation C.2. The extension of DNA, x, corresponded to the distance between the biotin linker and top of the recognition site and was found to be 6.76 ± 0.10 nm. See Figure C.1.

The linear charge density of ssDNA, determined from Equation C.1, was $\lambda = \frac{Fa}{sV}$. Assuming that the curvature of ssDNA within the constriction was similar to that within the vestibule, $\frac{a}{s}$ can be estimated to be $\frac{x}{L}$, where x and L are the extension and contour length from the FJC model, respectively. For the single dC substitution in poly-dA, we found that the linear charge density of ssDNA was $0.52 \pm 0.05 \frac{e^-}{nm}$ or $0.29 \pm 0.03 \frac{e^-}{nt}$ assuming an interphosphate distance of 0.56 nm.

C.2 Homopolymer Thymine Strand

Experiments done on a poly-dT strand which had a single dA substitution at X=13-16 produced a similar trend to that found with single nucleotide substitutions in poly-dA (Figure C.3 and Figure 6.3). Using the same procedure described above, the forces were determined to be lower than those found above for a poly-dA strand (Figure 6.5) and the effective charge was $0.39 \pm 0.02 \frac{e^-}{nm}$ or $0.22 \pm 0.01 \frac{e^-}{nt}$ (Figure C.2).

C.3 Single Nucleotide Polymorphism (SNP) Strand

Single nucleotide polymorphism (SNP) experiments were done on a heteromeric strand with a single nucleotide that could be either a dA or dC. The location of this variable nucleotide was shifted to be at X=13-16 and the difference in resistance for the dA and dC substitutions were found (Figure C.4). The peak and FWHM values were found from the fitted Gaussian curve. Using the same procedure described above, the forces were determined to be similar to those of poly-dT and the effective charge was $0.39 \pm 0.01 \frac{e^-}{nm}$ or $0.22 \pm 0.01 \frac{e^-}{nt}$. See Figure C.3.

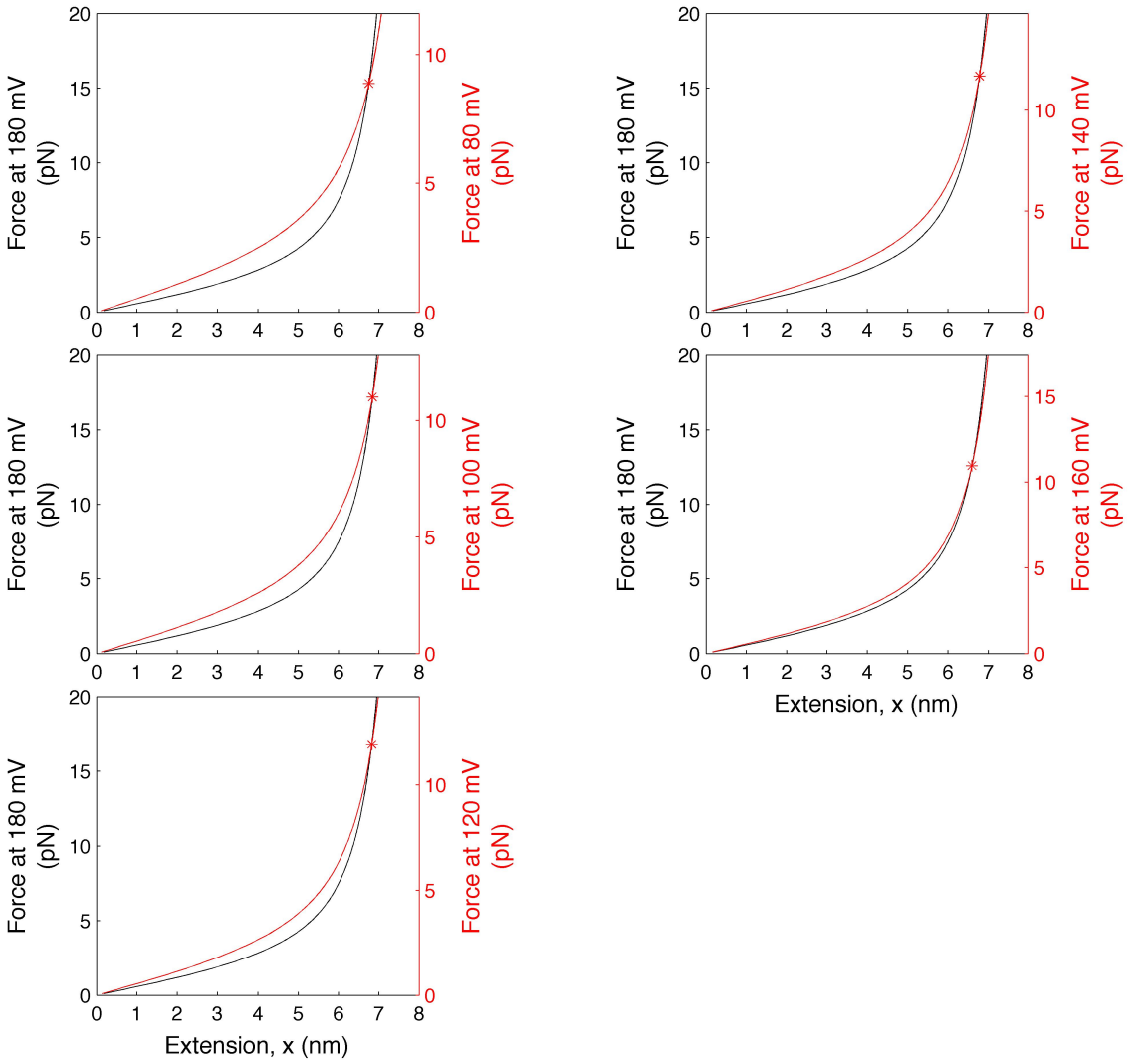


Figure C.2: **Force-Extension Curves for Immobilized DNA Modeled as a Freely Jointed Chain.** Force extension curves for a FJC (Equation C.3) are plotted for each voltage using contour lengths, L, found experimentally. The curve for an applied voltage of 180 mV is compared to the curves for each lower voltage. The red star indicates the position where the end-to-end distances of the strand, x, are equal and the forces are related by Equation C.2.

Table C.1: **FJC Calculations for Single Nucleotide Substitutions in Poly-dA.** The central nucleotide and number of nucleotides in the recognition site are taken from a Gaussian fit to experimental data (Figure 6.3). The errors are the standard errors of the Gaussian fit parameters. The contour length of DNA is measured from the biotin anchor to the top of the recognition site assuming an interphosphate distance of 0.56 nm. The force relationship $\frac{F(XmV)}{F(180mV)}$ is calculated from Equation C.2 and is used in the FJC model to relate the force extension curves for two voltages. The force at 180 mV and the extension are determined from the FJC model and correspond to the red stars in Figure C.2. The force at each lower voltage is determined using the mean value at of the force at 180 mV and the force relationship $\frac{F(XmV)}{F(180mV)}$. Finally, the charge is determined using Equation C.1 as described in the text.

Single dC Substitution in poly-dA

Voltage (mV)	Experimental Data		Contour Length L (nm)	F(X mV) F(180 mV)	FJC Solution		Force (pN)	Charge (e^-/nm)
	Central Nucleotide (X)	# nt in Recognition Site			F (180 mV) (pN)	Extension (nm)		
80	15.91 +/- 0.23	3.60 +/- 0.58	7.90 +/- 0.21	0.58	15.30	6.75	8.95 +/- 1.89	0.60 +/- 0.13
100	15.38 +/- 0.16	3.22 +/- 0.39	7.71 +/- 0.14	0.64	17.20	6.84	9.99 +/- 1.83	0.55 +/- 0.10
120	15.06 +/- 0.11	2.96 +/- 0.26	7.60 +/- 0.10	0.71	16.80	6.83	11.03 +/- 1.79	0.51 +/- 0.08
140	14.84 +/- 0.07	2.65 +/- 0.16	7.57 +/- 0.06	0.74	15.80	6.78	11.52 +/- 1.71	0.46 +/- 0.07
160	14.63 +/- 0.08	2.73 +/- 0.18	7.43 +/- 0.07	0.87	12.60	6.59	13.55 +/- 2.06	0.48 +/- 0.07
180	14.39 +/- 0.08	2.78 +/- 0.20	7.28 +/- 0.07	1.00	-	-	15.54 +/- 1.81	0.50 +/- 0.06
				Mean	***15.54***	***6.76***		***0.52***
				Std Dev	***1.81***	***0.10***		***0.05***

Single dA Substitution in poly-dT

Voltage (mV)	Experimental Data		Contour Length L (nm)	F(X mV) F(180 mV)	FJC Solution		Force (pN)	Charge (e^-/nm)
	Central Nucleotide (X)	# nt in Recognition Site			F (180 mV) (pN)	Extension (nm)		
80	15.89 +/- 0.50	3.26 +/- 0.94	7.99 +/- 0.39	0.56	13.50	6.60	6.73 +/- 2.69	0.43 +/- 0.17
100	15.37 +/- 0.27	2.81 +/- 0.72	7.82 +/- 0.25	0.60	14.70	6.68	7.24 +/- 2.73	0.37 +/- 0.14
120	15.05 +/- 0.05	2.79 +/- 0.15	7.65 +/- 0.05	0.72	11.80	6.48	8.62 +/- 2.43	0.38 +/- 0.11
140	14.72 +/- 0.06	2.78 +/- 0.18	7.47 +/- 0.06	0.84	10.00	6.32	10.05 +/- 2.86	0.39 +/- 0.11
160	14.42 +/- 0.16	2.72 +/- 0.45	7.31 +/- 0.15	0.94	9.90	6.31	11.22 +/- 3.61	0.39 +/- 0.13
180	14.21 +/- 0.20	2.58 +/- 0.55	7.23 +/- 0.19	1.00	-	-	11.98 +/- 2.12	0.37 +/- 0.07
				Mean	**11.98**	**6.48**		***0.39***
				Std Dev	**2.12**	**0.16**		***0.02***

Table C.2: **FJC Calculations for Single Nucleotide Substitutions in Poly-dT.** Same caption as Figure C.1.

SNP Strand

Voltage (mV)	Experimental Data		Contour Length L (nm)	F(X mV) F(180 mV)	FJC Solution		Force (pN)	Charge (e^-/nm)
	Central Nucleotide (X)	# nt in Recognition Site			F (180 mV) (pN)	Extension (nm)		
80	16.06 +/- 0.06	3.73 +/- 0.10	7.95 +/- 0.04	0.54	11.40	6.23	6.52 +/- 0.52	0.40 +/- 0.03
100	15.28 +/- 0.01	3.59 +/- 0.03	7.55 +/- 0.01	0.65	11.80	6.26	7.84 +/- 0.60	0.41 +/- 0.03
120	14.84 +/- 0.08	3.33 +/- 0.25	7.38 +/- 0.08	0.73	11.40	6.23	8.72 +/- 0.93	0.39 +/- 0.04
140	14.49 +/- 0.08	3.19 +/- 0.27	7.22 +/- 0.09	0.81	11.90	6.29	9.75 +/- 1.10	0.38 +/- 0.04
160	14.18 +/- 0.08	3.11 +/- 0.23	7.07 +/- 0.08	0.91	13.40	6.37	10.85 +/- 1.16	0.38 +/- 0.04
180	13.99 +/- 0.03	3.05 +/- 0.10	6.98 +/- 0.03	1.00	-	-	11.98 +/- 0.83	0.37 +/- 0.03
				Mean	**11.98**	**6.28**		***0.39***
				Std Dev	**0.83**	**0.06**		***0.01***

Table C.3: **FJC Calculations for SNP Strands.** Same caption as Figure C.1.

C.4 Brownian Motion Calculation Details

As previously discussed, a Gaussian curve fitted to experimental results (Figure 6.3) was used to determine which nucleotides most affected the DNA-pore resistance. The recognition site of the DNA strand was defined by the FWHM of the Gaussian fit. The recognition site width contained contributions from the width of MspA's constriction as well as Brownian motion of nucleotides moving in and out of the constriction. It was estimated that the applied voltage dropped entirely over the constriction so that only nucleotides within this region experienced a force. Brownian motion, however, caused nucleotides around the constriction to move in and out of the electric field making the number of nucleotides affecting the DNA-pore resistance to be larger than expected for a stationary system.

A DNA strand immobilized in MspA can be modeled as a spring with a linear restoring force. Brownian motion fluctuations resulting from changes in thermal kinetic energy caused the DNA to move around its equilibrium position. Thermal kinetic energy is characterized by the value k_bT where k_b is the Boltzmann constant and T is temperature. From the equipartition theorem, thermal kinetic energy is equal to the potential energy of a harmonic trap, $\frac{1}{2}k_bT = \frac{1}{2}\kappa < x^2 >$ where κ is the spring constant of the trap, and $< x^2 >$ is the variance in the displacement from equilibrium. The spring constant, κ, is ratio of an applied force, F, and the resulting displacement, x, of a particle; $\kappa = \frac{F}{x}$. The variance of the DNA due to Brownian motion was therefore found by:

$$< x^2 >= \frac{k_bT}{F/x} \tag{C.4}$$

where k_bT is 4.11 pN-nm at room temperature and $\frac{F}{x}$ was found from the derivative of the force-extension relation (Equation C.3). To compare these results with experiment, the variance was converted into the FWHM of a corresponding Gaussian curve by $< x^2 >= \frac{FWHM^2}{8ln2}$.

The width of the Gaussian curve from experimental data was related to the curves from MspA's constriction and Brownian motion by $FWHM^2_{experiment} = FWHM^2_{constriction} + FWHM^2_{brownian}$. The width of MspA's constriction, then, was determined from $FWHM^2_{constriction} = FWHM^2_{experiment} \;\; 8ln2 * (\frac{k_bT}{F/x})$.

Comparisons of the Brownian motion calculations and the widths of the recognition sites for experiments using poly-dA, poly-dT and SNP are shown in Figure C.5. The contributions from Brownian motion account for the majority of the width of the recognition site. For the poly-dA strand, we found that the mean width of MspA's constriction was $\approx 0.79 \pm 0.23$ nm. For poly-dT and SNP strands held with a low applied voltage, the recognition site was smaller than the expected contribution from Brownian motion. This suggests that either unstacked DNA strands cannot be modeled as a spring or the nucleotides interacted with the pore constriction constraining their motion.

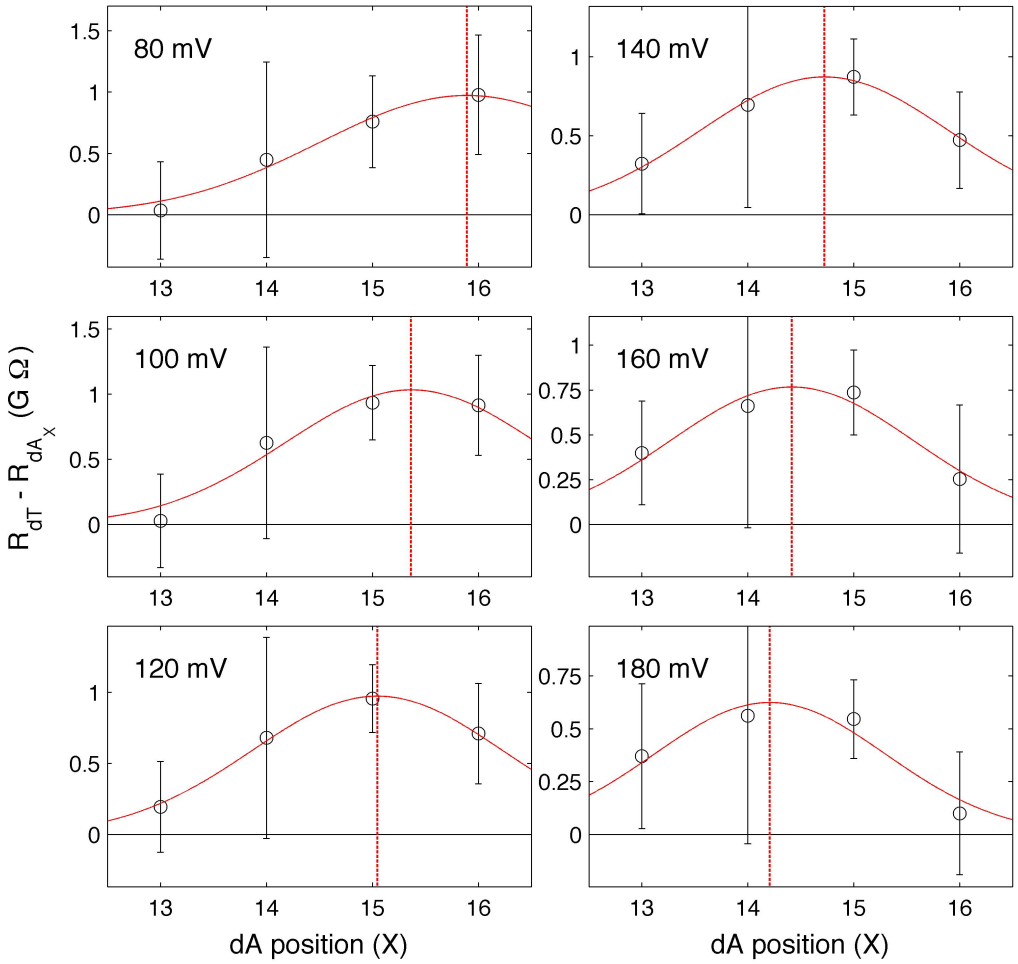


Figure C.3: **Experimental Results for Single Nucleotide Substitutions in Poly-dT strands Immobilized in MspA.** Comparison of resistance for DNA with a single adenine (dA) substitution at position X in an otherwise poly-dT strand. The deviation of the resistance from that of poly-dT is shown. The top of each plot corresponds to the resistance for a poly-dA strand. The central nucleotide in the constriction (vertical red dashed line) is taken as the peak of the fitted Gaussian curve and the width of the recognition site is taken as the full width at half max (FWHM).

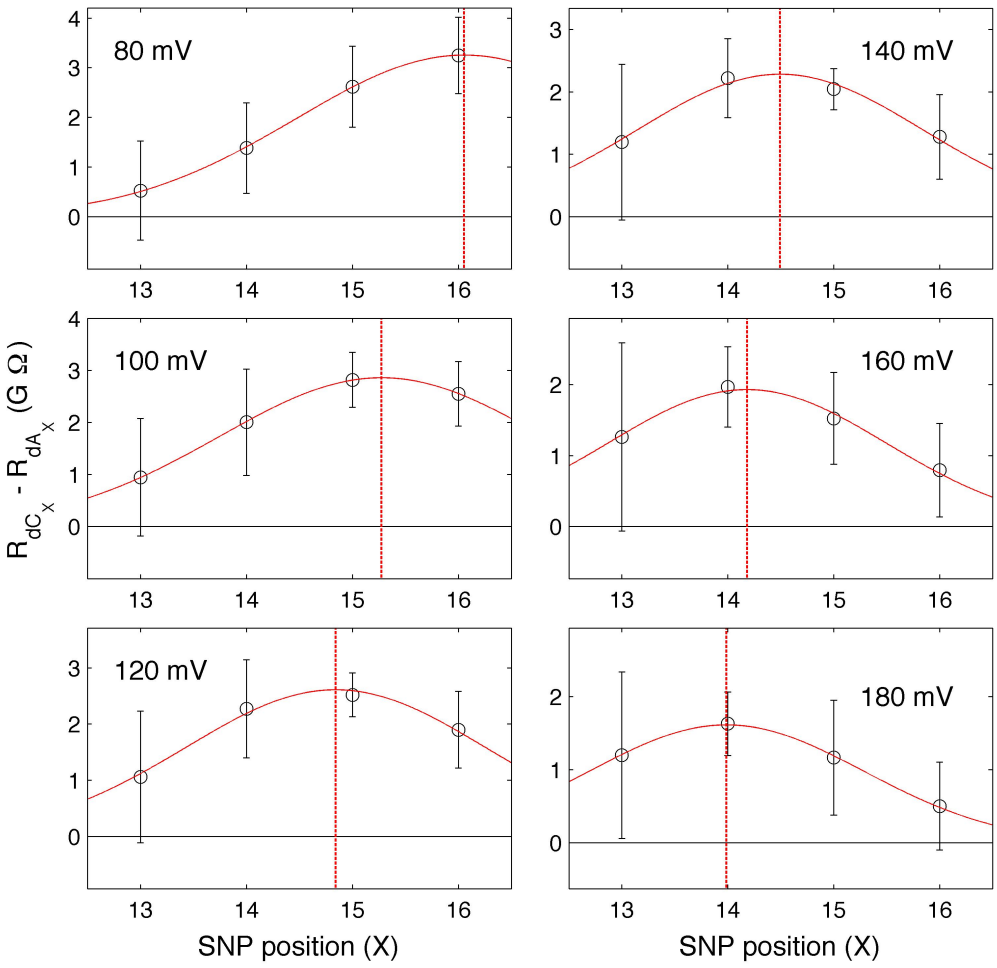


Figure C.4: **Experimental Results for Single Nucleotide Polymorphism Strands.** Comparisons of the voltage dependent resistance of the DNA-pore system is shown for experiments with a single adenine (dA) or cytosine (dC) substitution at nucleotide position X from the NeutrAvidin anchor in a heteromeric strand. The difference in resistance of for the dA and dC substitutions is shown for each nucleotide position and voltage. The top of each graph corresponds to the difference in resistance expected for poly-dC and poly-dA. The central nucleotide in the constriction (vertical red dashed line) is taken as the peak of the fitted Gaussian curve and the width of the constriction is taken as the full width at half max (FWHM).

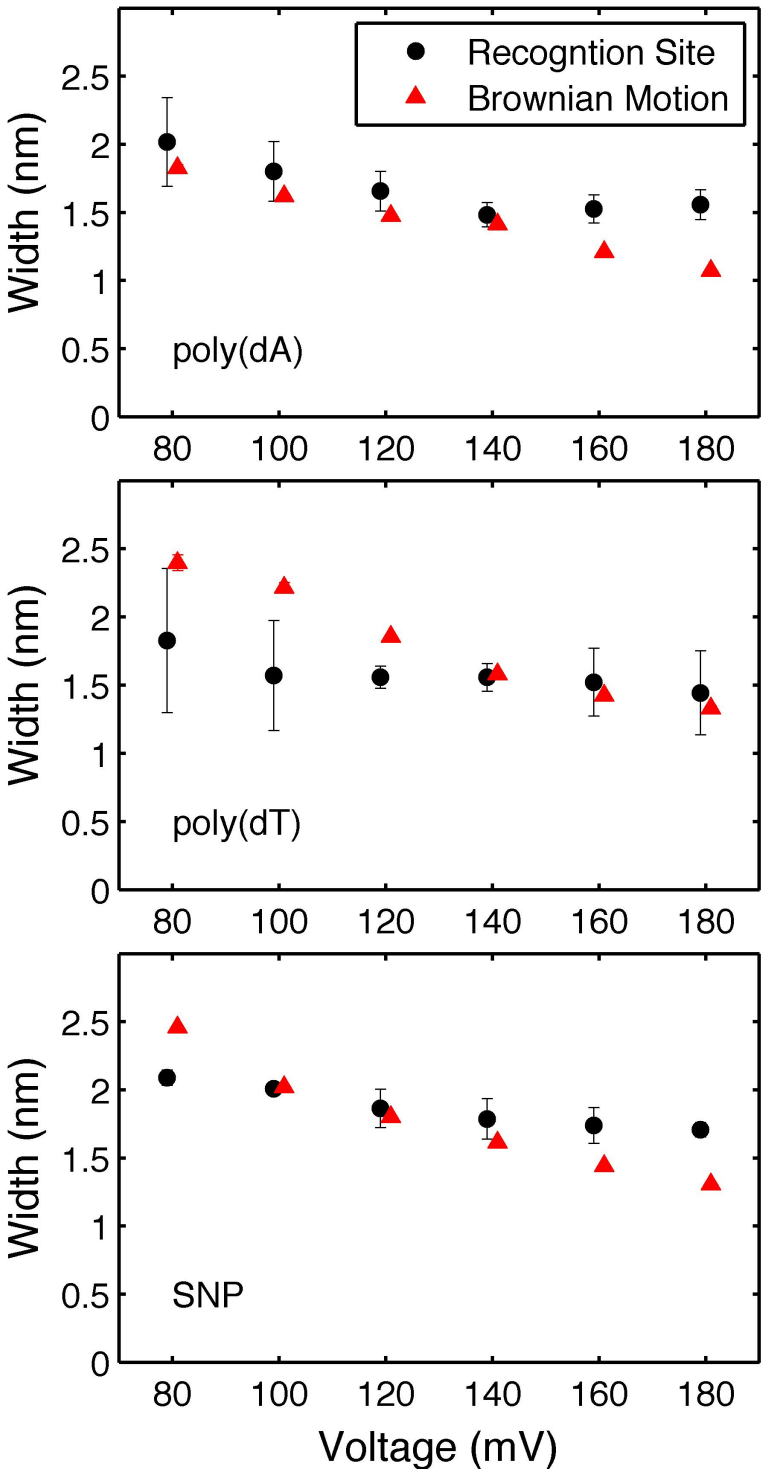


Figure C.5: **Width of MspA's Recognition Site Compared to Brownian Motion Calculations of DNA Immobilized in MspA.** The recognition sites determined from experiment (black, circles) are the FWHM from the Gaussian curve fitted to experimental results assuming an interphosphate distance of 0.56 nm (Figure 6.3). The Brownian motion values (red, triangles) are determined as described in the text. Plots are shown for Single dC substitution in poly-dA strand (top), Single dA substitution in poly-dT strand (middle), SNP Strand (bottom).

Name	Sequence	Num. Pores
Homopolymer strands: 3' leading orientation		
poly-dC	5′- /biotin/ CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	11
poly-dA	5′- /biotin/ AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAC CTG TCT CCC TGC CG	6
poly-dT	5′- /biotin/ TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTC CTG TCT CCC TGC CG	3
Homopolymer strands: 5' leading orientation		
poly-dC	5′- CTG TCT CCC TGC CGC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CC /biotin/	3
poly-dA	5′- CTG TCT CCC TGC CGA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA /biotin/	3
poly-dT	5′- CTG TCT CCC TGC CGT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT /biotin/	3
Single dC substitution in poly-dA		
dC_11	5′- /biotin/ AAA AAA AAA ACA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	3
dC_12	5′- /biotin/ AAA AAA AAA AAC AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	3
dC_13	5′- /biotin/ AAA AAA AAA AAA CAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	4
dC_14	5′- /biotin/ AAA AAA AAA AAA ACA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	3
dC_15	5′- /biotin/ AAA AAA AAA AAA AAC AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	5
dC_16	5′- /biotin/ AAA AAA AAA AAA AAA CAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	4
dC_17	5′- /biotin/ AAA AAA AAA AAA AAA ACA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	3
dC_18	5′- /biotin/ AAA AAA AAA AAA AAA AAC AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CTG TCT CCC TGC CG	4
Single dA substitution in poly-dC		
dA_11	5′- /biotin/ CCC CCC CCC CAC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	3
dA_12	5′- /biotin/ CCC CCC CCC CCA CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	3
dA_13	5′- /biotin/ CCC CCC CCC CCC ACC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	4
dA_14	5′- /biotin/ CCC CCC CCC CCC CAC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	4
dA_15	5′- /biotin/ CCC CCC CCC CCC CCA CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	6
dA_16	5′- /biotin/ CCC CCC CCC CCC CCC ACC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	4
dA_17	5′- /biotin/ CCC CCC CCC CCC CCC CAC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	3
dA_18	5′- /biotin/ CCC CCC CCC CCC CCC CCA CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CCC CTG TCT CCC TGC CG	3
Single dA substitution in poly-dT		
dA_13	5′- /biotin/ TTT TTT TTT TTT ATT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT CTG TCT CCC TGC CG	3
dA_14	5′- /biotin/ TTT TTT TTT TTT TAT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT CTG TCT CCC TGC CG	4
dA_15	5′- /biotin/ TTT TTT TTT TTT TTA TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT CTG TCT CCC TGC CG	3
dA_16	5′- /biotin/ TTT TTT TTT TTT TTT ATT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT CTG TCT CCC TGC CG	3
Single nucleotide polymorphism (SNP)		
dC_13	5′- /biotin/ GCT GGA GAA AGG CAT GTG CAA ATT AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	4
dA_13	5′- /biotin/ GCT GGA GAA AGG AAT GTG CAA ATT AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	3
dC_14	5′- /biotin/ TGC TGG AGA AAG GCA TGT GCA AAT TAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	3
dA_14	5′- /biotin/ TGC TGG AGA AAG GAA TGT GCA AAT TAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	3
dC_15	5′- /biotin/ CTG CTG GAG AAA GGC ATG TGC AAA TTA AGA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	3
dA_15	5′- /biotin/ CTG CTG GAG AAA GGA ATG TGC AAA TTA AGA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	3
dC_16	5′- /biotin/ CCT GCT GGA GAA AGG CAT GTG CAA ATT AAG AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	3
dA_16	5′- /biotin/ CCT GCT GGA GAA AGG AAT GTG CAA ATT AAG AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA	3

Table C.4: **DNA Sequences used in Force Spectroscopy Experiments.** The sequences of all strands used in these experiments are shown in the table along with the number of individual MspA pores that the strand was tested on.

VITA

Elizabeth A. Manrao

University of Washington • 3910 15th Ave NE • Seattle, WA 98195
Phone: (206) 619-2822 • E-Mail: emanrao@gmail.com

Education

Ph.D. Physics & Nanotechnology	**University of Washington, Seattle WA** Thesis: *Biological Nanopore MspA for DNA Sequencing*	Spring 2013
M.S. Physics	**University of Washington, Seattle WA**	2009
B.S. Physics	**San Francisco State University, San Francisco CA**	2005
B.S. Architectural Engineering	**California Polytechnic State University, San Luis Obispo CA** Graduated with honors	2002

Research Experience

- **University of Washington, Seattle WA** 2008-Present
  Graduate Student Researcher
  - Studied *Mycobacterium smegmatis* porin A, MspA, for use in nanopore DNA sequencing, a technique in which modulation of ionic current through a nanometer sized pore determines the sequence of DNA passing through it
  - Designed and performed experiments in which DNA was immobilized in MspA with an avidin molecule in order to determine MspA's sensitivity to base identity, methylation, and individual nucleotide differences in DNA
  - Performed force spectroscopy with MspA to study mechanical properties of DNA
  - Directed projects utilizing a molecular motor enzyme to move DNA though MspA one nucleotide at a time and provide sequence specific current traces
  - Programmed in Matlab to analyze experimental data
- **San Francisco State University, San Francisco CA** 2003-2005
  Student Research Assistant
  - Performed spectral modeling to compare abundance of volatile elements in stars with and without planetary companions for the California and Carnegie Planet Search Consortium

Publications

- **Manrao, E.A.**; Derrington, I.M.; Gundlach, J.H. (2013) *Force spectroscopy of single stranded DNA using the biological nanopore MspA*. In Preparation.
- **Manrao, E.A.**; Derrington, I.M.; Laszlo, A.H.; Langford, K.W.; Hopper, M.K.; Gillgren, N.; Pavlenok, M.; Niederweis, M.; Gundlach, J.H. (2012) *Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase*. Nature Biotechnology, 30 (4) 349-353.
- **Manrao, E.A.**; Derrington, I.M.; Pavlenok, M.; Niederweis, M.; Gundlach, J.H. (2011) *Nucleotide Discrimination with DNA Immobilized in the MspA Nanopore*. PLoS One, 6 (10) e25723.
- Derrington, I.M.; Butler, T.Z.; Collins, M.D.; **Manrao, E.**; Pavlenok, M.; Niederweis, M.; Gundlach, J.H. (2010) *Nanopore DNA Sequencing with MspA*. PNAS, 107 (37) 16060-16065.
- Fischer, D.A.; Laughlin, G.; Butler, P.; Marcy, G.; Johnson, J.; Henry, G.; Valenti, J.; Vogt, S.; Ammons, M.; Robinson, S.; Spear, G.; Strader, J.; Driscoll, P.; Fuller, A.; Johnson, T.; **Manrao, E.**; McCarthy, C.; Muñoz, M.; Tah, K. L.; Wright, J.; Ida, S.; Sato, B.; Toyota, E.; and Minniti, D. (2005) *The N2K Consortium. I. A Hot Saturn Planet Orbiting HD 88133*. Astrophysical Journal, 620, 481-486.

Work Experience

- **University of Washington, Seattle WA** 2007-2008
  Teaching Assistant, Department of Physics
  - Led tutorials and laboratory exercises for Introductory Physics courses
  - Led discussions, created sample problems, and held office hours for Science & Society, a course largely for undergraduate non-science majors
- **San Francisco State University, San Francisco CA** 2005-2007
  Assistant Director of Residential Administrative Services
  - Managed two groups, supervising 2 full time and several student employees
- **SSFM International, Honolulu HI** 2002-2003
  Entry Level Structural Engineer
  - Designed and inspected educational, military, residential, and commercial structures

Poster Presentations

- Biophysical Society 57th Annual Meeting, Philadelphia PA February 2013
  Force Spectroscopy of Single Stranded DNA with Biological Nanopore MspA
- NHGRI Advanced Sequencing Technology Development Meeting, San Diego CA April 2012
  Reading DNA with the MspA Nanopore and a Motor Enzyme
- Biophysical Society 56th Annual Meeting, San Diego CA February 2012
  Biological Nanopore MspA for DNA Sequencing
- NHGRI Advanced Sequencing Technology Development Meeting, San Diego CA April 2011
  Single-Nucleotide Resolution with DNA Immobilized in the MspA Nanopore
- Biophysical Society 55th Annual Meeting, Baltimore MD March 2011
  Determining the Resolution of Biological Porin MspA for Nanopore Sequencing
- Biosensing with Channels Summer School, Berder Island Brittany France August 2010
  Towards DNA Nanopore Sequencing with MspA
- NHGRI Advanced Sequencing Technology Development Meeting, Chapel Hill NC March 2010
  Single Nucleotide Discrimination in Single Stranded DNA Immobilized Within Biological Nanopore MspA
- Biophysical Society 54th Annual Meeting, San Francisco CA February 2010
  Single Nucleotide Discrimination in Single-Stranded DNA Immobilized within Biological Nanopore MspA
- American Astronomical Society 207th Meeting Washington, DC January 2006
  Spectroscopic Analysis of Volatile Elements

Skills

- Laboratory Skills: Form planar lipid bilayers, Perform qPCR reactions, Design custom DNA sequences, Perform data analysis, Use patch clamp amplifiers.
- Computer Skills: Matlab, C++, LabView, LATEX, PyMOL.